University of Padova

Department of Mathematics

Master Thesis in Data Science

Instance-Based Composed Image Retrieval with Vision-Language Models

Supervisor

Prof. Lamberto Ballan

University of Padova

Co-supervisor

Prof. Marco Fiorucci

University of Padova

Master Candidate

Mohammadamin Almasi Kolor

Academic Year

2025-2026

To my family,
for your love, patience, and quiet strength.
This thesis is also yours.

Abstract

Image retrieval systems increasingly need to support precise user intent rather than only broad semantic similarity. In instance-based composed image retrieval, the query combines a reference image and a textual modification. A correct target must depict the same literal instance, such as the same product, character, landmark, or object, and must also satisfy the requested visual change. This makes the task stricter than ordinary composed retrieval: a candidate is wrong if it shows the correct modification on a different instance, and it is also wrong if it preserves the instance but ignores the modification.

This thesis studies the task as a conjunction of two verification signals: identity preservation and modification satisfaction. The central idea is to keep these two forms of evidence visible during scoring and combine them multiplicatively, so that a candidate is rewarded only when both are supported. I first audit several frozen embedding backbones and training-free retrieval components under the intended query-local hard-negative protocol. The strongest first-stage system uses SigLIP2 SO400M Patch16 384 with leave-one-instance-out image centering and text contextualization, reaching 59.64 mean average precision at 100 on the full benchmark.

The thesis then introduces two second-stage reranking routes that preserve the first-stage image and text scores while adding more instance-focused evidence. Route A uses ILIAS, an image instance-level retrieval dataset, as an auxiliary source of supervision. Lightweight linear probes trained on frozen embeddings show that strong instance-level structure is already present in the representation; when used as a dual-encoder reranking signal together with a contextualized text branch, this route raises the ICIR score to 66.05 mean average precision at 100. Route B uses RZEnEmbed-7B as an instruction-aware multimodal embedder. Candidate images are embedded with an instance descriptor generated offline by open-source multimodal language models, encouraging the reranker to focus on the literal target instance rather than on generic scene content.

The best final system is the descriptor-guided Route B product, which multiplies the first-stage image score, the first-stage text score, and two RZEn reranking scores. It reaches 66.42 mean average precision at 100 and 65.65 macro mean average precision at 100 while preserving the same candidate-set recall as the first stage. The results show that modern frozen embeddings already contain useful instance information, but that high-performing instance-based composed image retrieval requires careful separation and recombination of identity and modification evidence.

Contents

Listing of figures

Listing of tables

Acronyms

AP	Average Precision.
BASIC	Training-free baseline pipeline for instance-level composed image retrieval introduced in the ICIR paper.
CLIP	Contrastive Language-Image Pre-training.
CIR	Composed Image Retrieval.
DINO	Self-Distillation with No Labels; used here to refer to self-supervised visual embedding backbones.
GPU	Graphics Processing Unit.
HN	Hard Negative.
ICIR	Instance-Level Composed Image Retrieval benchmark and task setting.
ILIAS	Auxiliary instance-level image retrieval dataset used for diagnostic and linear-probe experiments.
InternVL	Open-source multimodal vision-language model family used for descriptor-generation experiments.
InfoNCE	Information Noise-Contrastive Estimation.
mAP	Mean Average Precision.
MLLM	Multimodal Large Language Model.
OpenCLIP	Open implementation and training framework for CLIP-style vision-language models.
R@100	Recall at 100.
RZEn	RZEnEmbed, the multimodal embedding model used for second-stage reranking.
SigLIP	Sigmoid Loss for Language-Image Pre-training.
SO400M	SigLIP/SigLIP2 model variant identifier used by the model authors.
ViT	Vision Transformer.
VLM	Vision-Language Model.
VLM2Vec	Vision-language embedding model family that maps multimodal inputs to retrieval vectors.
VRAM	Video Random Access Memory.

1 Introduction

This chapter introduces the problem addressed in the thesis and explains why it is different from standard image retrieval. The discussion starts from the broad goal of retrieving images that match a user's intent, then moves to composed image retrieval, where the query contains both a reference image and a text instruction. The central focus is the stricter instance-level setting, where the retrieved image must preserve the exact object, product, character, artwork, or landmark shown in the query image while also satisfying the requested modification. The chapter also explains why the selected methods were chosen, why the task is important in practice, and how the thesis contributions address the limitations of existing approaches.

1.1 Motivation and problem setting

Image retrieval is one of the most direct ways in which people interact with large visual collections. A user has an intention, the system has a database, and the goal is to rank the database so that the most relevant images appear first. In the simplest version of the problem, the query is a text phrase such as "red shoe" or an image of an object, and the system returns visually or seman-

tically similar images. This already supports many useful applications, from web image search to product search, visual archives, and multimedia organization. However, real user intent is often more precise than a category label or a generic image similarity request.

In many practical situations, a user does not simply want an image of a shoe, a chair, a monument, a toy, or a character. The user may have a very specific visual instance in mind and want to retrieve that same instance under another visible condition. A shopper may want the same shoe model in a different color. A designer may want the same chair in a different material. A cultural-heritage researcher may want images of the same artifact photographed under another lighting condition. A tourist may want the same landmark at night, in snow, or from another viewpoint. A collector may want the same toy, poster, or object in a different state. These are not only semantic searches. They are searches in which part of the visual identity must remain fixed while another part must change.

This kind of intent is naturally expressed by combining an image and a short piece of text. The image communicates details that would be difficult to describe precisely: the exact product shape, the object proportions, the character design, the landmark structure, the artwork style, or the distinctive arrangement of parts. The text communicates the desired change: “at night”, “made of wood”, “as an emoji”, “in black”, “seen from the front”, or “covered in snow”. The query is therefore not just an image query and not just a text query. It is a composed query.

The difficulty is that the system must understand what should remain stable and what should change. If it follows the image too strongly, it may retrieve the same instance in the original state rather than the requested target state. If it follows the text too strongly, it may retrieve images that satisfy the modification but depict a different instance. A useful retrieval system must keep both constraints active at the same time. This is the central motivation for the thesis.

The applications are broad. In electronic commerce, instance-level composed retrieval can reduce friction in product discovery. Users often know a product visually but not by its exact name, brand, or catalog identifier. They may also want a variant of that product. Returning a different but visually similar product may be unacceptable because identity matters for price, availability, compatibility, or user preference. In cultural heritage, instance preservation is even more important: different buildings, artworks, or artifacts may share a category but have dis-

tinct historical meaning. In personal media collections, a user may want the same person, pet, object, or place under a changed context. In creative tools, a designer may want references that preserve a target object but change material, style, lighting, or composition. In assistive systems, instance-sensitive retrieval can help users locate or recognize specific objects across changing conditions.

These examples show why the problem should not be reduced to ordinary category retrieval. A model that returns any chair, any monument, or any character is not solving the task. It must preserve the literal target instance. At the same time, instance preservation alone is not enough: returning the same object without the requested modification is also wrong. The task is therefore inherently conjunctive. A target must satisfy two conditions simultaneously.

From image retrieval to composed retrieval. Modern image retrieval is usually formulated as embedding search. An encoder maps an image or a text query to a vector, database images are represented by vectors, and candidates are ranked by a similarity function such as cosine similarity. This formulation is attractive because it is scalable. Database embeddings can be computed once and reused. Approximate nearest-neighbor search can make retrieval efficient even for large collections. Most importantly, it provides a common interface for comparing different models and different scoring rules.

Vision-language models made this interface much more powerful. Models in the CLIP family learn image and text encoders that map both modalities into a shared embedding space [2, 3]. SigLIP and SigLIP2 improve this family with different training objectives and stronger training recipes [4, 5]. Once images and text share a space, one can retrieve images with a text query, retrieve similar images with an image query, or combine image and text scores. This is the foundation for most of the first-stage experiments in the thesis and connects directly to the training-free BASIC family introduced with the instance-level composed image retrieval benchmark [1].

However, a shared embedding space is not automatically aligned with every retrieval objective. These models are usually trained to match images with captions or related text. Captions often describe broad semantic content: object category, scene type, attributes, and visible relations. They may not force the model to preserve every detail that distinguishes one instance

from another. For example, two shoes of the same category may be close in the embedding space even if they are different products. Two landmarks may be close because they share architectural features. Two cartoon characters may be close because they share a style. This is useful for semantic retrieval but can be harmful for instance-level retrieval.

Composed image retrieval adds another layer. Instead of ranking by a single query vector, the system receives a reference image and a textual modification. A simple approach is to compute an image similarity score between the query image and each candidate, compute a text similarity score between the modification and each candidate, and combine them. More advanced approaches learn a composition module, rewrite the text, project embeddings, use query expansion, or generate an implicit target representation. These methods have been successful in several semantic composed retrieval settings, including early feature-composition approaches such as TIRG and later zero-shot or training-free methods such as Pic2Word, SEARLE, CIReVL, and FreeDom [6, 7, 8, 9, 10].

The thesis adopts the composed-retrieval view but keeps the score factors explicit. This is important because explicit factors make the system interpretable. When a candidate ranks highly, one can inspect whether it was supported by the image branch, the text branch, or both. This matters for diagnosis: many failures in instance-level composed retrieval occur because one branch dominates the other. If the image branch dominates, the system retrieves the right instance in the wrong state. If the text branch dominates, it retrieves the right state on the wrong instance. Keeping the scores visible makes these failures measurable.

Why instance-level composed retrieval changes the problem. Most composed retrieval benchmarks and methods focus primarily on semantic correctness. In such a setting, a target may be considered correct if it belongs to the correct category and roughly satisfies the modification. This is useful, but it leaves open a different and stricter question: can a model retrieve the same literal instance after a modification? The instance-level composed image retrieval benchmark addresses this question [1].

The difference is easy to state but subtle in practice. Suppose the query image shows a specific lounge chair and the modification asks for it "in a different color". A candidate image of

another lounge chair in the requested color may be semantically plausible, but it is not the same instance. Suppose the query image shows a specific landmark and the modification asks for it "at night". A night image of a different monument is not relevant. Suppose the query image shows a character and the modification asks for an emoji style. An emoji of a different character is not a correct target. These mistakes are common because semantic similarity and instance identity are related but not identical.

The instance-level constraint does not mean that every query is automatically harder than semantic composed retrieval. In some cases, the identity constraint can actually simplify rejection: if a candidate clearly depicts a different instance, the system can discard it even if the modification is visible. The setting becomes especially difficult when the same-instance evidence and modification evidence conflict, when the requested modification is subtle, or when the visual change makes the identity less obvious. Thus, instance-level composed retrieval changes the structure of the problem. It requires explicit control over what remains fixed and what is allowed to change.

Instance-level composed retrieval can therefore be viewed as a two-part verification problem. For every candidate image, the system should answer:

1. Does the candidate depict the same literal instance as the query image?
2. Does the candidate visually satisfy the requested modification?

The candidate is relevant only if both answers are positive. This is the core conceptual decomposition used throughout the thesis.

This decomposition also explains why additive fusion is often weak. If the final score is the sum of an image score and a text score, then a candidate can compensate for a failure on one branch with a high score on the other. A wrong instance that strongly matches the modification can survive. A correct instance that ignores the modification can also survive. Multiplication behaves differently. If either factor is low, the product is low. This makes multiplication a natural operator for a task whose relevance condition is close to a logical "and".

The experiments confirm this intuition. In the first-stage fusion comparison, the product of image and text scores is much stronger than additive fusion. Later, the same idea is extended to

the reranker: the final score multiplies the first-stage image score, the first-stage text score, the reranker image score, and the reranker text score. The goal is not to make the formula more complicated, but to preserve the two verification conditions across two model families.

The instance-level setting also changes how the candidate pool should be interpreted. In the benchmark used here, each query is evaluated against its own instance-hard-negative pool rather than against the whole image database. This is essential. Many images in the global database may be positives or hard negatives for other instances, and using the full database as a generic pool can distort the evaluation. The correct protocol tests each query against the candidates designed for that instance: positives and hard negatives that are actually confusable for the query. Throughout the thesis, results are reported under this query-local protocol.

1.2 Research problem

The thesis asks how far frozen embedding models can be pushed on instance-level composed image retrieval and how much can be gained by adding a second-stage reranker. This question is deliberately practical. Training a large multimodal backbone from scratch is not realistic for most research settings. Fine-tuning such models can also be expensive, unstable, and difficult to compare fairly across many variants. Frozen embeddings, by contrast, can be computed once, cached, and reused across many retrieval formulas. This makes it possible to run a broad and controlled experimental study.

The practical question is not only which backbone is best. It is also how the backbone should be used. A raw embedding space may contain useful information but organize it in a way that is not optimal for instance-level composed retrieval. Image embeddings may include background and dataset bias. Text embeddings may represent short modifications poorly. Similarity scores may be poorly calibrated across branches. Some operations may help weaker models but hurt stronger ones. The thesis therefore studies the pipeline around the embeddings, not only the embeddings themselves.

The first research direction is first-stage retrieval. The first stage must be efficient enough to search many candidates and strong enough to provide a high-quality top-100 shortlist. It

should produce high recall and meaningful scores for both identity and modification. The thesis audits several backbone families and several training-free operations, including centering, contextualization, projection, query expansion, and score normalization. The strongest first stage is the centered contextualized product using SigLIP2 SO400M Patch16 384.

The second research direction is reranking. Once the first stage has selected the top candidates, a heavier model can be used to compare them more carefully. This is where instruction-aware multimodal embeddings become useful. A large model such as RZEnEmbed-7B is too expensive to use naively over every possible candidate for every query, but it is feasible as a reranker over the top 100. The reranker can be guided with an instance descriptor so that it focuses on the literal target identity rather than generic scene content.

The third research direction is representation analysis. If instance-level information is already present in frozen embeddings, then the problem is partly about extraction and calibration rather than learning identity from scratch. The thesis investigates this through ILIAS, an auxiliary instance-retrieval dataset. Lightweight probes trained on top of frozen SigLIP2 embeddings provide evidence about whether the space already contains transferable identity information.

This framing also explains why the thesis emphasizes careful evaluation over a single end-to-end training run. A trained model may improve a benchmark, but without factor-level analysis it can be difficult to know why. Did it learn instance identity? Did it learn the modification language? Did it exploit a dataset bias? Did it improve recall, or only reorder candidates? By keeping the first-stage and second-stage scores explicit, the thesis can answer these questions more directly. Each component produces a score with an interpretable role, and each ablation tests a specific hypothesis.

The frozen-embedding setting is also practically relevant. Many real retrieval systems already contain precomputed embeddings for large image databases. Re-embedding the whole collection with every new experimental model can be expensive, and retraining a backbone can be even more difficult. Methods that improve retrieval by changing centering, text contextualization, score fusion, or reranking can often be integrated into existing systems more easily than methods requiring full retraining. This does not make training unimportant; rather, it makes

the frozen setting a strong and realistic starting point.

The two-stage design follows the same practical logic. A first-stage model must be fast, stable, and reusable. A second-stage model can be slower because it is applied only to a limited shortlist. This is common in retrieval systems, but it is especially useful for instance-level composed retrieval because the final decision often depends on subtle differences among near misses. The thesis therefore treats the first stage as the source of recall and the reranker as a verifier that improves precision within the candidate set.

Why these methods were chosen. The method choices follow from the structure of the problem. The first stage uses frozen dual-encoder embeddings because they are efficient and scalable. Among the tested models, SigLIP2 SO400M Patch16 384 provides the strongest first-stage behavior. This model is large enough to provide high-quality visual and textual embeddings, but still practical enough for full-dataset embedding, caching, and repeated scoring. It therefore provides a good balance between strength and experimental controllability.

Leave-one-instance-out image centering is used because instance-level retrieval is sensitive to global embedding bias. A dataset center can capture common visual structure that is not specific to the query instance. Subtracting this center can make the remaining similarity more focused on discriminative identity cues. The leave-one-instance-out detail matters because it avoids using images from the current instance when computing the center. This keeps the centering operation training-free and reduces leakage risk.

Text contextualization is used because many modifications are short fragments. A phrase such as “at night” or “as an emoji” may not look like the natural captions used to train vision-language text encoders. Contextualization embeds the modification inside multiple neutral sentence templates and averages the resulting text embeddings. This does not add information about the answer; it changes the linguistic form of the query so that the text encoder receives inputs closer to its training distribution.

The reranker uses RZEnEmbed-7B because it provides an embedding interface while allowing instruction-aware multimodal conditioning. This is important for the second stage. The reranker should not simply be another generic image-text similarity model. It should be asked

to focus on the literal target instance. To provide that focus, the pipeline generates an instance descriptor from the query image. The descriptor is then used as an attention guide when embedding each candidate image. This creates a candidate representation that is more explicitly oriented toward the identity question.

The final score remains multiplicative. This is not a stylistic choice. It is the mathematical reflection of the task. A high final score should require support from the first-stage image branch, the first-stage text branch, the descriptor-guided reranker image branch, and the descriptor-guided reranker text branch. This makes the final method conservative: it tries to avoid candidates that are supported by only one kind of evidence.

Several alternative directions were explored during the project and helped define the final design. Richer captions and future-state descriptions were generated with multimodal language models, but they often introduced extra variability into the text branch. Crop-based methods using grounding models were tested, but they were sensitive to localization errors and sometimes removed useful context. Linear probing on ILIAS strongly improved pure instance retrieval and later became a competitive second-stage signal when paired with a compatible text branch. These outcomes shaped the final method: keep the strongest stable first stage, add complementary verifier signals, and avoid components that make the evidence harder to interpret.

This is why the thesis is not only a search for the highest number. It is also an attempt to identify which operations are reliable for this particular kind of retrieval. The strongest result matters, but the negative results matter too. They show that instance-level composed retrieval has a specific structure, and methods that ignore that structure can fail even when they are sophisticated or computationally expensive.

1.3 Contributions

The thesis makes four main contributions, which are used repeatedly as the organizing structure of the methodology and results chapters. The strongest reported system improves the previous BASIC state of the art for the benchmark by more than six mAP@100 points in the final full-query protocol, showing that the gains are substantial even relative to an already strong training-

free baseline [1].

A broad experimental study over embedding models. The thesis audits several families of frozen embeddings for instance-level composed image retrieval. These include dual-encoder models such as CLIP and OpenCLIP, different CLIP baselines, SigLIP variants, and multimodal large language model based embeddings such as RZEn, Qwen-based embeddings, and VLM2Vec. The study is performed under the intended query-local instance-hard-negative protocol, so the numbers reflect the actual difficulty of ranking positives against confusable candidates for the same query.

A compact hard-negative version of the auxiliary instance dataset. The project constructs a smaller, more practical mini-ILIAS setting for the auxiliary experiments. Instead of repeatedly using the original multi-million distractor space, the experiments retain approximately one million useful distractors selected to preserve hard-negative value. This makes the instance-retrieval experiments cheaper to run while keeping the focus on confusable candidates rather than easy unrelated negatives. The resulting setup supports the linear-probe training and makes future instance-focused experiments more manageable.

An analysis of instance information in frozen embeddings. The thesis investigates whether instance-related information already exists in stored embeddings. Using ILIAS, lightweight linear probes are trained on top of frozen SigLIP2 embeddings and evaluated under leave-one-domain-out splits [11, 5]. The results show large improvements over the frozen baseline, suggesting that frozen embeddings already contain instance-level structure that can be extracted with a small adaptation. This evidence supports the broader claim that instance-focused retrieval is feasible without learning everything from scratch, while the ICIR transfer experiments show that the extracted instance signal must be balanced with modification evidence.

Two strong second-stage reranking routes. The strongest first-stage system uses SigLIP2 SO400M Patch16 384 with leave-one-instance-out image centering and text contextu-

alization, reaching 59.64 mAP@100 on the full benchmark. The thesis then studies two ways to move the system's focus toward the exact instance. The ILIAS route uses a lightweight linear probe trained with instance-retrieval supervision and applies it as an additional second-stage verifier. The RZEn route uses an instance descriptor and an instruction-aware multimodal embedder [12]. The final descriptor-guided RZEn route reaches 66.42 mAP@100 and 65.65 Macro-mAP@100, while the ILIAS route is close behind and demonstrates that supervised instance adaptation can also become a practical reranker.

1.4 Thesis structure

Chapter 2 reviews the literature most relevant to the thesis. It follows the progression from image retrieval to instance-based retrieval, composed retrieval, and instance-level composed retrieval. Chapter 3 introduces the benchmark, the auxiliary ILIAS dataset, the mini-ILIAS hard-negative construction, and the evaluation metrics. Chapter 4 describes the methodology, including the first-stage SigLIP2 pipeline, the ILIAS linear-probe route trained with hard negatives, descriptor generation, and RZEn reranking. Chapter 5 reports the experimental results and analyzes both second-stage routes, including their successful and unsuccessful cases. Chapter 6 concludes with limitations and future work.

2

Related Work

This chapter situates the thesis within the retrieval literature. The discussion follows the progression of the problem: general image retrieval, instance-based image retrieval, composed image retrieval, and finally instance-level composed image retrieval. This order is important because the thesis combines ideas from all four areas. It uses embedding-based retrieval for scalability, instance-retrieval concepts for identity preservation, composed-retrieval methods for image-plus-text queries, and instruction-aware multimodal embeddings for reranking difficult candidates.

2.1 Image and instance-based retrieval

Image retrieval asks a system to rank images according to their relevance to a query. The query may be another image, a text description, or a combination of modalities. Classical image retrieval relied heavily on local features, geometric matching, and carefully designed descriptors. Modern retrieval systems usually represent each query and each database item by a learned vector and perform nearest-neighbor search in that embedding space. This shift is central to the

thesis because it allows the same image database to be embedded once and then reused across many scoring rules, centering strategies, and reranking experiments.

The embedding view is powerful because it separates expensive representation extraction from relatively cheap ranking. Given an encoder f and an image x, the system stores a vector $z = f(x)$. Query-time retrieval then becomes a similarity computation between the query vector and candidate vectors. This is the same interface used by CLIP, OpenCLIP, SigLIP, SigLIP2, Qwen embedding models, VLM2Vec, and RZEn in this thesis [2, 3, 4, 5, 13, 14, 15, 12]. The common interface makes it possible to ask a controlled question: if the candidate pool and evaluation protocol are fixed, how much does the retrieval result change when the embedding model or score formula changes?

Vision-language pretraining made embedding retrieval much more flexible. CLIP trains an image encoder and a text encoder with a contrastive objective so that matched image-caption pairs are close and mismatched pairs are separated [2]. OpenCLIP showed that this recipe can be reproduced and scaled with open training data [3]. SigLIP changes the training loss from a batch-softmax contrastive loss to a pairwise sigmoid loss, while preserving the same practical retrieval interface [4]. SigLIP2 further improves the representation quality, semantic understanding, multilingual coverage, localization, and dense-feature behavior of the SigLIP family [5]. These models are natural first-stage candidates because database images and modification text can be embedded independently and compared by dot product.

The strength of such models is also their limitation. Caption-supervised image-text alignment usually emphasizes semantic compatibility: objects, attributes, actions, scenes, and relations that are likely to be written in captions. Exact instance identity is only indirectly supervised. Two visually similar products, two landmarks with related architecture, or two cartoon characters from the same visual style may be close in the embedding space even when they are not the same literal instance. This tension is not a flaw of CLIP-like models; it follows from what they are trained to do. For the present thesis, it means that the embedding geometry must be audited carefully rather than assumed to match the benchmark perfectly.

Image-only encoders provide a useful contrast. Self-supervised visual models such as DINOv2 and DINOv3 can preserve strong visual structure without relying on language super-

vision [16, 17]. They are relevant to instance retrieval because local shape, texture, and visual correspondences can matter more than caption-level semantics. However, they do not provide a text encoder in a shared space, so they cannot directly solve image-plus-modification retrieval. In this thesis they are mainly part of the broader backbone audit and the hard-negative mining process for the auxiliary dataset.

The retrieval literature therefore motivates two design decisions. First, the thesis uses frozen embeddings and cached score factors whenever possible, because this makes broad comparison feasible. Second, it avoids treating a single embedding score as fully explanatory. A high image-image similarity and a high text-image similarity mean different things. The method keeps these sources of evidence separate so that the later analysis can identify whether failures come from identity matching, modification matching, or fusion.

Instance-based image retrieval. Instance-based image retrieval is stricter than semantic retrieval. The objective is not to find any image from the same category, but images depicting the same particular object, product, artwork, landmark, character, or other recurring visual identity. This changes what should be invariant. For semantic retrieval, a representation may intentionally ignore small details so that all chairs, shoes, or monuments of a type are close together. For instance retrieval, some of those small details are the identity signal itself.

The difficulty of instance retrieval is the balance between robustness and specificity. The representation should tolerate viewpoint, illumination, scale, occlusion, background, and partial visibility, but it should not discard the details that distinguish one instance from another. For a product, these details may be shape, stitching, brand placement, part configuration, or material. For a landmark, they may be facade layout, geometry, and spatial context. For an artwork, sign, or character, they may be marks, colors, line structure, or distinctive design patterns. A representation that is too semantic will collapse these details; a representation that is too literal will fail under natural visual changes.

ILIAS is the most relevant instance-retrieval reference for this thesis [11]. It is designed to evaluate instance-level image retrieval at scale across diverse domains, with query images, positives, text queries, and a very large distractor space. The ILIAS paper is important here for

two reasons. First, it emphasizes that foundation embeddings can contain instance-level information even when they are not trained exclusively for instance retrieval. Second, it reports that lightweight adaptation layers can improve vision-language embeddings for instance-level recognition. These observations directly motivated the auxiliary linear-probe experiments in this thesis.

The connection to the present work is conceptual rather than dataset-identical. ILIAS isolates the identity subproblem: retrieve the same instance. The main benchmark in this thesis adds a modification constraint: retrieve the same instance after a requested change. Therefore, ILIAS cannot replace instance-level composed retrieval, but it can test whether the identity information needed by composed retrieval is already present in frozen embeddings. The leave-one-domain-out probe experiments in Chapter 5 use exactly this diagnostic role.

This connection also clarifies why the thesis studies both training-based and prompting-based routes for identity. There are two plausible ways to encourage a system to preserve the instance. The first is to train a small adaptation, such as a linear probe, using instance-level supervision. The second is to use a multimodal embedding model and prompt it to focus on the literal instance described by a query-specific descriptor. The final thesis contains both routes: an ILIAS-trained linear-probe reranker and a descriptor-guided RZEn reranker. They are different implementations of the same high-level goal: make the second stage more identity-aware than a generic embedding comparison.

2.2 Composed and instance-level composed image retrieval

Composed image retrieval studies queries made of a reference image and a textual modification. The reference image provides the visual anchor and the text describes the desired change. This formulation is useful because many real searches are not well expressed by text alone or image alone. A user may want “this product but in black”, “this room with a different chair”, or “this landmark at night”. The problem is to combine the preserved information from the reference image with the changed information from the text.

The task was formalized in early neural retrieval work such as TIRG, which studied how to compose image and text features so that the composed query lives in the target-image embedding space [6]. This line of work made clear that composition is not simply concatenation. The text should modify the visual representation while preserving the parts of the image that remain relevant. Later supervised methods explored attention, transformers, autoencoding, and metric-learning objectives for the same goal.

Several datasets shaped the field. FashionIQ introduced natural-language feedback for fashion retrieval and became a standard benchmark for attribute-level and product-oriented composed retrieval [18]. Its strength is realistic language feedback in a focused commercial domain, but its domain specificity limits how much it tests open-world visual reasoning. CIRR moved the task to real-life open-domain images with human-generated modification sentences and introduced subset evaluation to reduce false-negative issues among visually similar images [19]. CIRCO, introduced together with SEARLE, further addressed false negatives by using multiple valid targets per query in an open-domain setting based on common objects [8]. These datasets collectively show why composed retrieval is challenging: a single reference image can be modified in multiple plausible ways, and a single text instruction may match several targets.

The method landscape is correspondingly broad. Pic2Word maps a reference image into a pseudo-word token so that composed retrieval can be performed through the text encoder [7]. SEARLE uses textual inversion for zero-shot composed retrieval and introduced CIRCO as a benchmark with multiple ground truths [8]. CompoDiff uses latent diffusion to model composed queries and synthetic triplets at scale [20]. CIReVL uses a training-free vision-by-language strategy that recombines vision-language models and language models [9]. FreeDom studies training-free domain conversion for composed retrieval [10]. The original ICIR paper also compares methods such as WeiCom, Pic2Word, CompoDiff, CIReVL, SEARLE, MagicLens, CoVR-2, FreeDom, and BASIC under its evaluation setting [1].

These methods differ in how they represent the composed query. Some learn a direct image-text composition module. Some convert the image into language-like tokens. Some generate or rewrite text. Some use diffusion or target-image synthesis. Some remain training-free and combine separately estimated image and text scores. The thesis is closest to the last family, especially

BASIC, because it keeps score factors explicit and interpretable. This choice is deliberate: when a candidate fails, it matters whether the failure came from the identity branch, the modification branch, or the fusion rule.

Composed retrieval is not only a modeling problem but also an evaluation problem. In semantic composed retrieval, a candidate may be acceptable if it belongs to the right category and satisfies the modification. In instance-level composed retrieval, a candidate must also be the same literal instance. However, it would be too strong to say that the instance-level setting is always harder in every scenario. Sometimes the exact-instance constraint can help reject otherwise plausible semantic matches: if a candidate clearly depicts a different object, it can be discarded even if the modification is visible. The difficulty increases when the system must preserve identity under changes that make the instance less visually obvious, or when the modification is subtle and hard to verify. Thus, the instance-level setting changes the nature of the problem rather than simply making every query uniformly harder.

Instance-level composed image retrieval. Instance-level composed image retrieval combines the two constraints studied separately above. The query consists of an image and a modification, and a target is relevant only if it depicts the same literal instance and satisfies the requested change. The ICIR benchmark formalizes this setting and introduces a compact but difficult evaluation protocol with positives and semi-automatically selected hard negatives for each query [1]. Its design is especially relevant to this thesis because the negatives are not arbitrary unrelated images. They are selected to be confusable with respect to instance identity, modification satisfaction, or both.

The original ICIR paper introduces BASIC as a strong training-free baseline [1]. BASIC separately estimates query-image-to-candidate-image similarity and query-text-to-candidate-image similarity, improves each branch with simple operations, and then fuses them so that candidates satisfying both branches are upweighted. This branch-wise philosophy is the starting point of the present methodology. The thesis reimplements and audits this family of operations with stronger backbones, especially SigLIP2 SO400M Patch16 384.

The most important difference from many earlier composed retrieval benchmarks is the

query-local hard-negative protocol. Each query is evaluated against its own positives and curated hard negatives. This is not merely an implementation detail. In instance-level composed retrieval, an image can be a positive for one query group and a hard negative for another. Treating the full database as a generic candidate pool can distort the task by mixing unrelated negatives with query-specific near misses. The thesis therefore follows the intended instance-HN protocol throughout the main ICIR results.

The ICIR paper reports BASIC as the previous state of the art for this setting. The present thesis uses that result as the reference point and improves it with stronger backbones and reranking. The strongest first stage replaces the original backbone with SigLIP2 SO400M Patch16 384 and keeps only the components that remain useful in that stronger embedding space: leave-one-instance-out image centering and text contextualization. The final second stage then adds either an ILIAS-trained linear-probe verifier or a descriptor-guided RZEn verifier. Both outperform the strong first-stage baseline, and the descriptor-guided RZEn route gives the strongest result reported here.

This progression is important. The thesis does not claim that all earlier methods fail because they are weak. Rather, it shows that the instance-level setting rewards a particular structure: preserve a strong and explicit image branch, preserve a strong and explicit modification branch, then add a second verifier without collapsing the two pieces of evidence. This explains why a simple four-factor product can outperform more complex branches that add generated captions, extra difference variables, or crop-based signals.

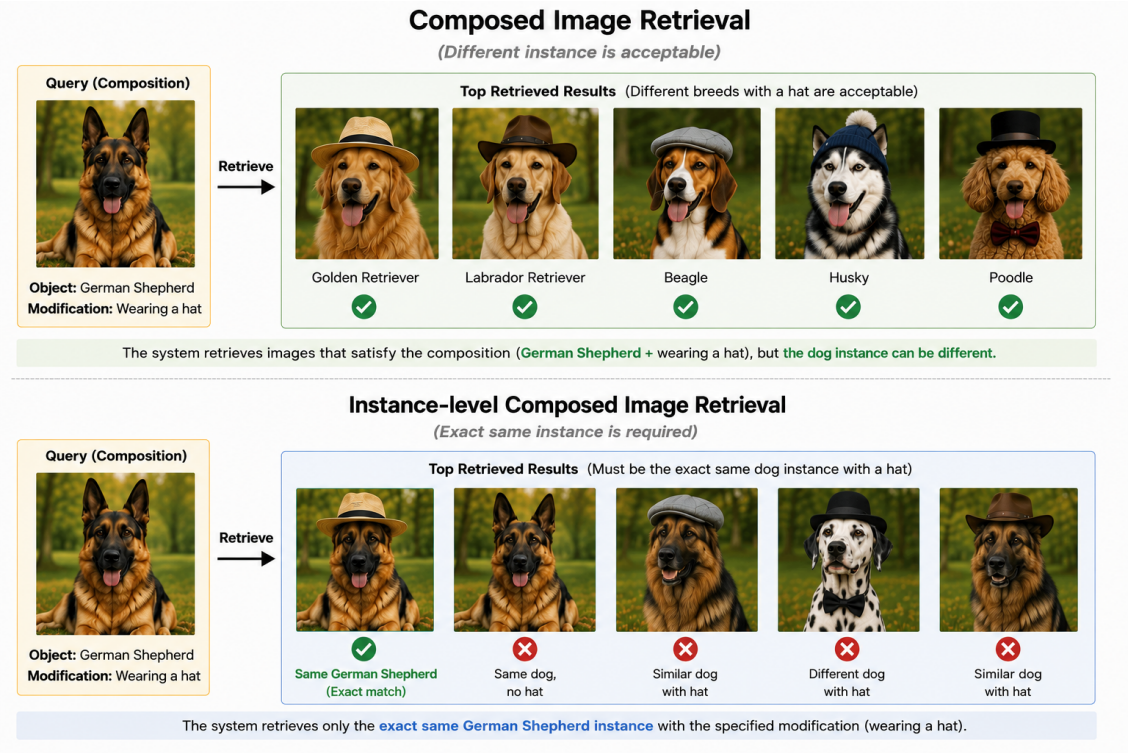

Figure 2.1: Conceptual difference between ordinary composed image retrieval and instance-level composed image retrieval. In the upper example, the query combines a German Shepherd image with the modification "wearing a hat", but any dog breed wearing a hat may be considered acceptable. In the lower example, the retrieved image must depict the exact same German Shepherd instance and must also satisfy the modification. The red negatives illustrate the two dominant failure cases in this thesis: the correct instance without the required state, and a different instance that satisfies the modification.

2.3 Multimodal large language model based embeddings

Multimodal large language model based embeddings provide a different interface from classical dual encoders. Instead of embedding an image and a caption with fixed modality-specific towers, models such as RZEn, Qwen multimodal embeddings, and VLM2Vec can embed images, text, or multimodal inputs while taking instructions into account [12, 13, 14, 15]. They are more expensive than CLIP/SigLIP-style models, but they can be useful as rerankers because the candidate set is small after the first stage.

This instruction interface is attractive for instance-level composed retrieval. The remaining errors after a strong first stage are often fine-grained: a candidate may show the right modification on the wrong instance, the right instance without the modification, or a visually similar near duplicate. A generic embedding may not know which part of the image should be treated as the stable identity. An instruction-aware embedder can be given a descriptor and asked to represent the candidate with respect to that descriptor.

The thesis uses this capability in two ways. First, open-source multimodal language models such as Qwen and InternVL are used to generate instance descriptors from query images. The modification is used only to identify the relevant instance in the query image, not to leak the desired target state. Second, RZEnEmbed-7B uses the descriptor as an attention guide when embedding each shortlisted candidate. The result is still an embedding-based reranker: the model outputs vectors, and final ranking is computed by dot products and multiplication. It is not a generative answer system.

This distinction is important for cost and reproducibility. Autoregressive generation depends on decoding settings and can be slow for long outputs. Reranking by embeddings is a bounded forward-pass computation. Once candidate embeddings are computed, the final score is deterministic. The thesis therefore uses multimodal large language model based embeddings where they are most useful: not as a full database search engine, but as a focused verifier over the top-100 candidates returned by SigLIP2.

2.4 Position of this thesis

The related work leaves a clear methodological gap. Image retrieval provides scalable embedding search, but not composed intent. Instance retrieval provides exact identity matching, but usually not text-conditioned modification. Composed retrieval provides image-plus-text search, but much of the literature evaluates semantic correctness rather than strict instance preservation. Instance-level composed image retrieval requires all three ideas at once.

The thesis addresses this gap with a two-stage, factorized design. The first stage uses a strong frozen SigLIP2 dual encoder with leave-one-instance-out image centering and text contextualization. The second stage adds a verifier: either a trained instance-aware linear probe or an instruction-aware multimodal embedder guided by an instance descriptor. The final score remains multiplicative so that a candidate must be supported by identity evidence and modification evidence from both stages.

The empirical message is that the previous BASIC state of the art can be substantially improved when the backbone and reranking design are adapted to the instance-level setting. The result is not only a higher number; it is evidence that exact-instance information already exists in modern embeddings and can be strengthened in two complementary ways: by training a lightweight instance-focused probe, or by prompting a multimodal embedding model to focus on the instance during reranking.

3

Background and Problem Setting

This chapter defines the task, datasets, and evaluation protocol used throughout the thesis. It explains why instance-level composed image retrieval must be evaluated differently from ordinary semantic retrieval: each query has its own relevant positives and carefully selected hard negatives, and the model is judged on whether it preserves the exact instance while satisfying the modification. The chapter also introduces ILIAS as an auxiliary diagnostic dataset and defines the retrieval metrics used in the experiments.

3.1 Instance-level composed image retrieval

The task studied in this thesis is instance-level composed image retrieval. A query is composed of a reference image and a textual modification. The desired target is not merely an image from the same semantic category, but an image that depicts the same literal instance after the requested change. For example, the correct target may be the same landmark at night, the same character in a different style, or the same product under a different condition.

This creates a stricter relevance definition than standard semantic composed image retrieval.

In semantic composed retrieval, a candidate can often be considered acceptable if it matches the category and roughly follows the modification. In instance-level composed retrieval, this is not enough. A candidate that correctly satisfies the modification but depicts a different instance is a false positive. Conversely, a candidate that depicts the correct instance but does not show the requested state is also a false positive. This does not mean that every individual query is always more difficult than semantic composed retrieval: when the candidate is clearly a different instance, the identity constraint can help reject it. The difficulty comes from candidates that are plausible under one condition but fail the other.

The distinction is subtle but important. A semantic retrieval system can succeed by learning that two images share a broad concept. An instance-level system must also learn which details are stable identifiers of the specific instance. These details vary by domain. For a product, they may include shape, logo placement, material, or design pattern. For a landmark, they may include architectural structure and spatial configuration. For a character, they may include face shape, costume, colors, or distinctive visual marks. The benchmark is difficult because these cues can be partially hidden, altered by viewpoint, or mixed with irrelevant background content.

The textual modification adds a second source of uncertainty. Some modifications describe a simple attribute such as color. Others describe context, action, style, time of day, viewpoint, or a state that affects only part of the image. The model must therefore avoid two opposite simplifications: treating the modification as the whole query, or ignoring it and retrieving only the same instance. Instance-level composed retrieval is precisely the regime where both simplifications fail.

This structure also motivates reporting Macro-mAP in addition to standard mAP. ICIR instances do not all contribute the same number of query images, modifications, and positives. If evaluation only averages over all queries, performance on an instance with many modifications can dominate the final number. Macro-mAP gives each instance group equal weight after first averaging within that group, which makes it easier to see whether a method improves instance-level retrieval broadly rather than only on over-represented identities.

The relevance condition can therefore be written as

$$\mathrm{rel}(q, d) = 1 \iff [\mathrm{same\_instance}(x_q, d) \wedge \mathrm{satisfies}(d, t_q)], \tag{3.1}$$

where $q = (x_q, t_q)$ is the composed query and d is a candidate image. This logical conjunction motivates the multiplicative score formulations used throughout the thesis.

The formula is intentionally written as a logical condition rather than as a neural-network architecture. It states the problem independently of any particular model. A dual encoder, an instruction-aware multimodal embedder, or a learned reranker may all produce different similarity values, but the relevance condition remains the same. This is why the thesis repeatedly returns to the same decomposition: identity preservation and modification satisfaction are not implementation details, they are the two parts of the task definition.

3.2 Instance-level composed image retrieval benchmark

ICIR is the main benchmark of this thesis [1]. It was designed specifically to evaluate whether retrieval systems can preserve literal instance identity while applying a textual modification. Figure 3.1 illustrates the core dynamic of the dataset: the query image supplies the instance, the text supplies the requested change, and the target must satisfy both. The benchmark is compact, but it is not easy: the candidate pools are built from positives and carefully crafted hard negatives, so the system is evaluated on near misses rather than on a large set of unrelated distractors.

The full ICIR benchmark used here contains 1,883 queries grouped around 202 instance identities. Each instance may have several query images, several textual modifications, multiple relevant positives, and a curated set of hard negatives. The benchmark is therefore not a flat image retrieval collection in which every unrelated image can be treated as a useful negative. Its difficulty comes from the fact that the negative pool is intentionally composed of near misses. Some negatives resemble the query image but do not show the requested modification. Others

Figure 3.1: Query structure of the instance-level composed image retrieval benchmark. The reference image identifies the literal instance, the text specifies the desired visual change, and a relevant target must satisfy both requirements. This is stricter than category-level composed retrieval because a visually plausible target is still wrong if it changes the underlying instance.

satisfy the modification but depict a different instance. Others are visually or semantically close to the target category and test whether the model can avoid broad semantic shortcuts. This design makes the benchmark a direct test of the two conditions used throughout the thesis: same-instance evidence and modification evidence.

The grouping by instance identity matters for both evaluation and interpretation. Some identities have more queries or modifications than others. Some instances are visually distinctive and therefore easier. Others are confusable because they belong to categories with many visually similar examples. If a method performs well only on a few frequent or easy identities, a global query average may hide this imbalance. For this reason, the thesis reports both query-level mean average precision and macro mean average precision over instance groups.

QUERY-LOCAL INSTANCE-HARD-NEGATIVE PROTOCOL. All ICIR experiments in this thesis use the intended query-local instance-hard-negative protocol. For a query q, the candidate set is

$$\mathcal{C}_q = \mathcal{P}_q \cup \mathcal{H}_q, \tag{3.2}$$

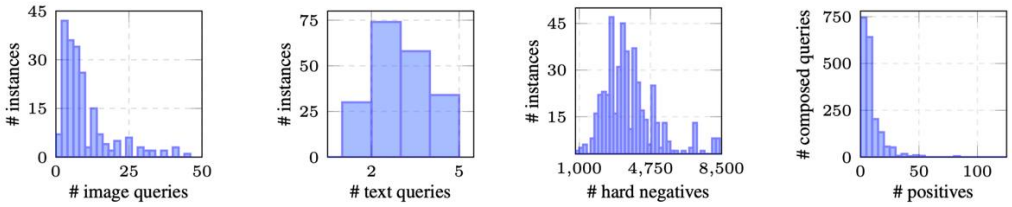

Figure 3.2: Organization of the query-local candidate pools. Each query is evaluated against its own positives and curated hard negatives. The hard negatives are intentionally close to the query instance or modification, so the evaluation tests fine-grained ranking among confusable candidates rather than rejection of unrelated images.

where $\mathcal{P}_q$ is the set of relevant positives and $\mathcal{H}_q$ is the curated hard-negative set associated with the query.

This protocol is important enough to state explicitly. Some images can be positives for one instance and hard negatives for another. Using the whole dataset as the candidate pool would mix together unrelated distractors, confusable near misses, and positives from other query contexts. That changes the benchmark. The thesis therefore does not evaluate ICIR by searching the whole image database for every query. Instead, each query is ranked inside its corresponding instance-HN pool.

This choice also makes the results more meaningful. If the whole database were used as the candidate pool, many retrieved negatives would be unrelated to the query and therefore easy to reject. A method could obtain a misleadingly strong number by separating categories rather than by solving the intended fine-grained problem. The query-local protocol keeps the focus on the hard part: deciding which candidates inside a deliberately confusable pool satisfy both the identity and modification constraints.

WHY THE BENCHMARK IS DIFFICULT. ICIR is difficult for several reasons. First, the same-instance constraint requires preserving small visual details that generic image-text embeddings may not emphasize. Second, the modification can describe a state, style, context, time, material, action, or viewpoint, so the text branch must remain sensitive to short and sometimes underspecified phrases. Third, database images are not always clean product-style shots. Many images are cluttered, partially occluded, or contain multiple potentially relevant objects. In

Figure 3.3: Example queries and positive targets from the instance-level composed retrieval benchmark. Positives may differ in viewpoint, style, or state, but they must remain tied to the same literal instance and satisfy the textual modification. The examples illustrate why the benchmark requires both visual identity preservation and modification verification.

such cases, the retrieval model must implicitly decide which object is the instance of interest and which surrounding details are incidental.

These properties explain why simple semantic matching is insufficient. A model must simultaneously recognize the instance, ignore irrelevant background variation, and verify the requested change.

The benchmark also exposes calibration failures between image and text. A model may produce a very high image similarity for the exact query instance, even when the modification is not visible. Another model may produce a high text similarity for the requested modification, even when the instance is wrong. The final ranking depends on how these two types of evidence are balanced. This is why the thesis studies fusion operators and score normalization explicitly, rather than treating the similarity scores as directly comparable by default.

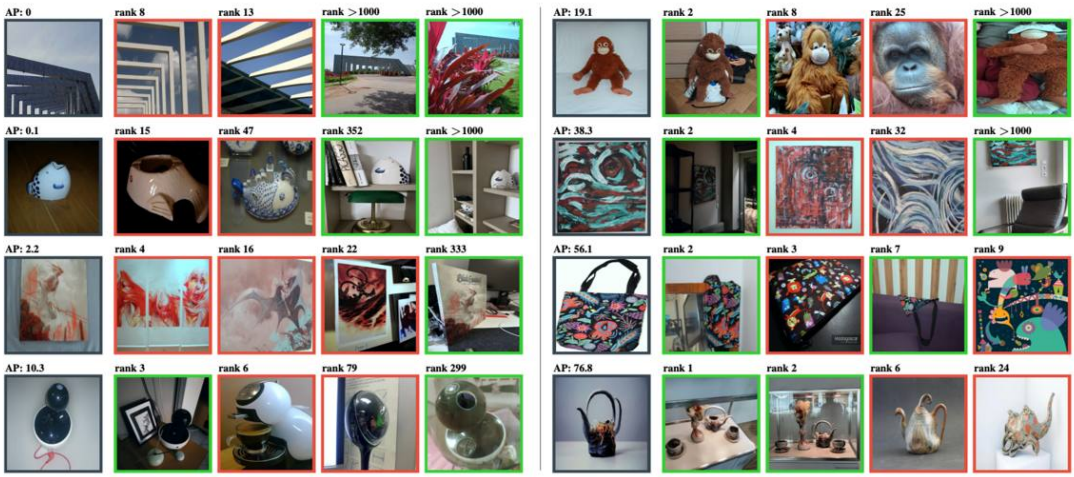

Figure 3.4: Example ILIAS auxiliary instance-retrieval sample. In this thesis, ILIAS is used both as a diagnostic dataset and as the training source for Route A, the ILIAS-trained dual-encoder reranker. The linear probe learns an instance-focused visual verification signal from ILIAS while the composed retrieval pipeline still preserves the original ICIR first-stage image and text scores.

3.3 Auxiliary instance retrieval dataset

ILIAS is used in this thesis for a different purpose [11]. It is not the final target benchmark, but an auxiliary supervised diagnostic dataset for studying whether instance-level information is already present in frozen embeddings. ILIAS contains instance-centered image retrieval supervision across high-level domains such as art, fashion, household objects, landmarks, media, products, technology, and toys. In contrast to ICIR, the auxiliary ILIAS experiments focus on the first part of the ICIR decomposition: recognizing the same instance. They do not require applying a textual modification to reach a changed target state. This makes ILIAS useful as a cleaner probe of instance identity.

The central question is:

> If a strong frozen embedding already contains instance identity information, can a small head extract that information even when tested on held-out domains?

This question is deliberately separated from the final composed retrieval objective. If the answer were negative, then the final method would likely need heavy supervised training to learn instance identity. If the answer is positive, then the problem becomes more interesting:

the information exists, but the system must expose it and combine it correctly with text. The ILIAS results later support the second interpretation.

Figure 3.5 shows the high-level categories used in ILIAS. These categories make it possible to run leave-one-domain-out experiments: train on several domains and test on a completely excluded one.

Auxiliary hard-negative construction. The original ILIAS setting contains a very large distractor space. For the experiments in this thesis, this was reduced to a more ICIR-like query-local hard-negative structure. Instead of repeatedly using a very large pool of roughly five million distractors, a smaller but more relevant mini-ILIAS setting was constructed with approximately one million related distractors. The goal was not to make the task easier, but to make the negatives more informative and computationally manageable.

The reduction was guided by strong frozen visual and multimodal similarities. CLIP/SigLIP-style embeddings and DINOv3 visual embeddings were used to identify distractors that were close to the query or positives under different notions of similarity [2, 4, 5, 17]. This produced dedicated hard-negative pools for each query group. Negatives that were retrieved near the top by both visual and multimodal models were treated as especially hard, because they were likely to share either visual appearance, semantic category, or both.

This construction mirrors the ICIR philosophy: the most useful negatives are not arbitrary unrelated images, but candidates that are plausible enough to test whether the representation really preserves instance identity. Operationally, for each query group I used the available positive images and searched the mini-ILIAS distractor pool with several frozen embedding spaces. Distractors that appeared high in the rankings of both visual and image-text models were retained as hard negatives. The hardest buckets were those retrieved very near the top by multiple models, because such images were likely to be visually or semantically close while still not being the true instance. This produces an ICIR-like local pool: each query is evaluated against positives plus carefully selected near misses, rather than against a mostly easy collection of unrelated images.

Role of the auxiliary dataset in the thesis. The ILIAS branch has two roles in the thesis. First, it is diagnostic: it tests whether the first subproblem, same-instance recognition, is already encoded in frozen SigLIP2 SO400M Patch16 384 embeddings. A lightweight linear probe is trained on top of frozen embeddings. In the strongest evaluation, one high-level ILIAS category is left out entirely during training and used only for testing. This creates an out-of-domain setting: the probe must transfer the notion of instance identity to categories it did not see during training.

If such a probe succeeds, it indicates that the frozen embedding already contains extractable instance-level structure. The important distinction is that ILIAS tests exact-instance recognition directly, while ICIR asks for the same instance under a requested change. Therefore ILIAS is not a replacement benchmark for ICIR, but it is a supporting source of supervision for the identity-preservation half of the ICIR task. The methodological details and results of this branch are discussed later, where they are used to support the claim that strong frozen embeddings already contain usable instance information.

The transfer experiment back to the composed retrieval benchmark is also important. A probe that improves pure instance retrieval can still hurt composed retrieval if it changes the image space without a matching text-side adaptation. Direct replacement is therefore not the final use of ILIAS. Instead, the thesis also uses the trained probe as a second-stage verifier: the original composed SigLIP2 image and text scores remain active, and the probe contributes an additional instance-focused image score paired with a compatible contextualized text score from a second backbone. In this role, ILIAS is not only diagnostic. It provides a practical reranking route that competes with the descriptor-guided multimodal reranker while remaining much cheaper computationally.

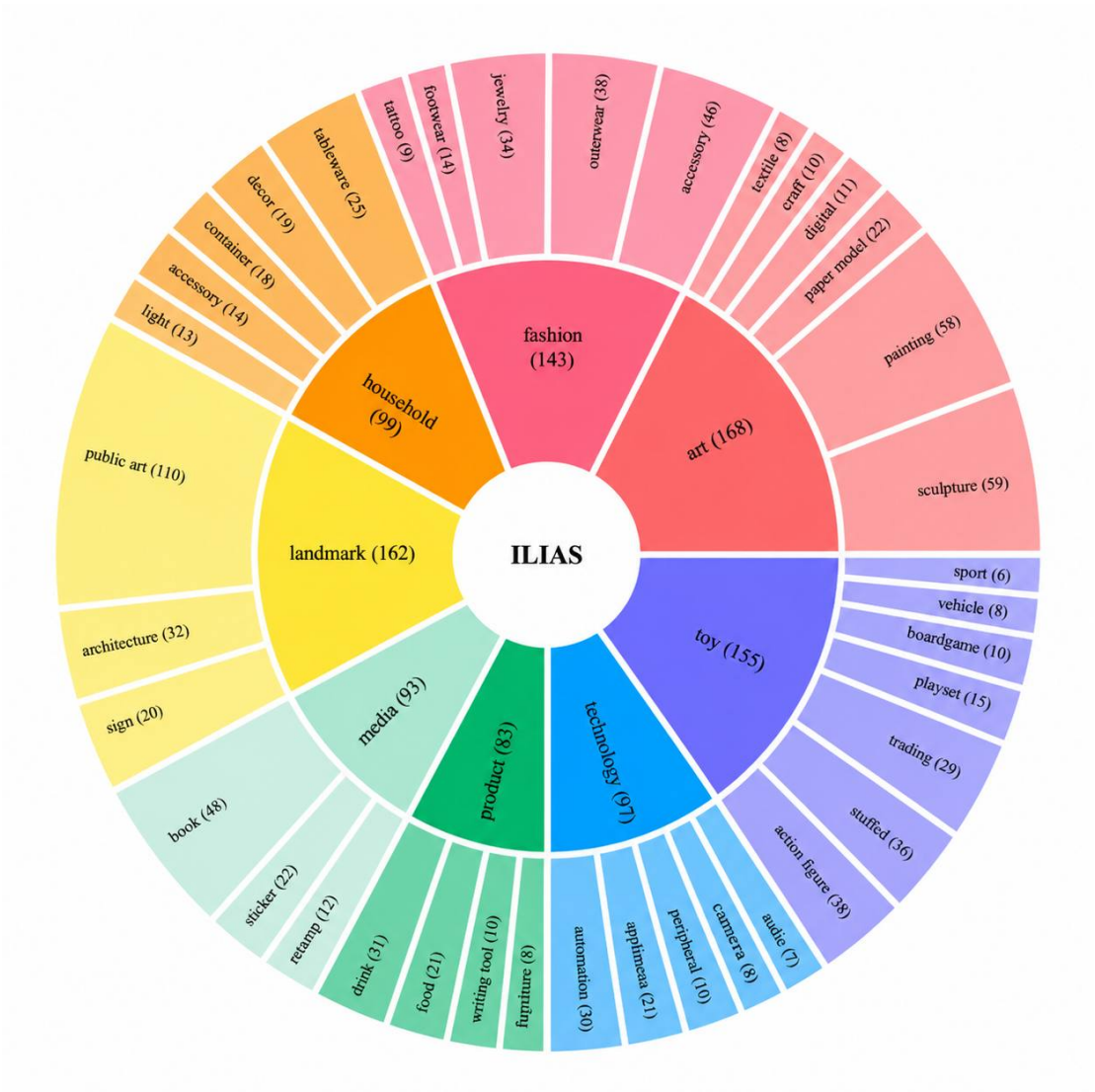

Figure 3.5: High-level categories used for leave-one-domain-out diagnostic experiments. In each run, one category is excluded from training and used only for testing, which evaluates whether a learned instance-retrieval head transfers beyond the domains seen during training.

3.4 Embedding retrieval and cosine similarity

The thesis operates primarily in an embedding-retrieval regime. An encoder maps an image or text input to a vector:

$$z = f(x) \in \mathbb{R}^d. \tag{3.3}$$

Embeddings are L_2-normalized:

$$\hat{z} = \frac{z}{\|z\|_2}, \tag{3.4}$$

so cosine similarity reduces to a dot product:

$$\mathrm{sim}(x_i, x_j) = \hat{z}_i^\top \hat{z}_j. \tag{3.5}$$

This geometry is efficient and matches the retrieval interface of CLIP, SigLIP, Qwen embedding models, VLM2Vec, and RZEn [2, 4, 13, 14, 15, 12]. It also makes it possible to cache database embeddings and run large sweeps over score formulas without recomputing the image encoder each time.

Caching is a major practical reason for using embedding retrieval in this thesis. Once the database side has been embedded, many experiments become score-level experiments rather than model-inference experiments. This makes it possible to compare backbones, fusion rules, centering variants, contextualization variants, and reranking formulas in a controlled way. It also reduces the risk of accidentally changing the image encoder while evaluating a scoring idea.

3.5 Contrastive multimodal representation learning

Most first-stage models in this thesis are dual encoders. They contain an image encoder f_I and a text encoder f_T:

$$z_i^I = f_I(x_i), \qquad z_i^T = f_T(t_i). \tag{3.6}$$

After normalization, image-text similarity is

$$s_{ij} = (\hat{z}_i^I)^\top (\hat{z}_j^T). \tag{3.7}$$

CLIP learns this shared space with a symmetric contrastive loss over matched image-text pairs [2]. SigLIP replaces the batch-softmax objective with a pairwise sigmoid loss [4]. SigLIP2 keeps the same general dual-encoder retrieval interface while improving the training recipe and supervision [5]. These models are attractive for first-stage retrieval because candidate images can be embedded once and queried efficiently.

For a batch of N matched image-text pairs, a CLIP-style loss forms an $N \times N$ similarity matrix and applies a cross-entropy objective in both directions. If s_{ij} is the similarity between image i and text j, the image-to-text loss is

$$\mathcal{L}_{I \to T} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}/\tau)}, \tag{3.8}$$

with an analogous text-to-image term. The temperature τ controls how sharply the model separates positives from in-batch negatives. SigLIP keeps the same retrieval interface but replaces the softmax normalization over the batch with independent sigmoid pair classification. This makes the training objective less dependent on the exact global batch normalization while still learning an image-text similarity space.

The auxiliary linear-probe experiments use related retrieval losses on top of frozen embeddings. In the pairwise sigmoid objective, an anchor-positive pair is assigned label 1 and anchor-negative pairs are assigned label 0. In local InfoNCE, each anchor sees one positive and several sampled hard negatives, and the positive is the correct class within that local set:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(s(a,p)/\tau)}{\exp(s(a,p)/\tau) + \sum_{k=1}^{K} \exp(s(a,n_k)/\tau)}. \tag{3.9}$$

In the multi-negative margin objective, each sampled negative is penalized if it is too close to

the anchor:

$$\mathcal{L}_{\text{margin}} = \frac{1}{K}\sum_{k=1}^{K} \max(0, m - s(a,p) + s(a,n_k)). \tag{3.10}$$

These losses are useful for ILIAS because the dataset is naturally local and instance-based: a query should be close to its positives and separated from hard negatives that may share category or visual appearance.

For instance-level composed retrieval, the advantage of dual encoders is also their limitation. They provide a single vector for an image and a single vector for a text. This is efficient, but it compresses many kinds of evidence into one representation. Fine-grained identity, scene context, object category, style, and modification state may all contribute to the same similarity score. The methodology chapter therefore introduces operations that try to separate or rebalance these factors without retraining the backbone.

The thesis also considers image-only encoders as diagnostic tools. DINOv2 and DINOv3 are self-supervised visual representation models that do not rely on paired image-text captions [16, 17]. They can be strong for visual similarity and hard-negative mining because their features often preserve local visual structure. However, without a matched text encoder they cannot directly evaluate whether an ICIR candidate satisfies a textual modification. In this thesis, DINO-style embeddings are therefore useful for constructing and analyzing hard negatives, while the final composed retrieval pipeline relies on models that can score both image and text evidence.

3.6 Evaluation metrics

Average precision and mean average precision. Let $\pi_q(k)$ be the candidate ranked at position k for query q, and let $r_q(k) = 1$ if the candidate is relevant and 0 otherwise. Precision at rank k is

$$P_q(k) = \frac{1}{k}\sum_{i=1}^{k} r_q(i). \tag{3.11}$$

Average Precision at 100 is

$$AP_q@100 = \frac{1}{|\mathcal{P}_q|} \sum_{k=1}^{100} P_q(k) r_q(k). \tag{3.12}$$

The denominator is the number of all positives for the query, not only the positives retrieved in the top 100. Mean Average Precision is

$$mAP@100 = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} AP_q@100. \tag{3.13}$$

Average precision is particularly useful because it rewards both retrieving positives and ranking them early. In this benchmark, many queries have multiple positives. A method that retrieves one positive at rank 1 but misses the rest is different from a method that retrieves several positives throughout the top 100. Average precision captures this ranked structure better than a single binary success measure.

Recall at one hundred. Recall at 100 is

$$Recall_q@100 = \frac{1}{|\mathcal{P}_q|} \sum_{k=1}^{100} r_q(k). \tag{3.14}$$

Dataset-level Recall@100 is the mean over queries.

Recall@100 has a different interpretation from mean average precision. In the two-stage pipeline, the first stage determines which candidates are available to the reranker. A second-stage reranker that only reorders the top 100 cannot improve Recall@100 unless the candidate union is enlarged. It can, however, improve mAP@100 by moving positives higher within the same candidate set. This distinction is important when interpreting the RZEn reranking results.

Macro mean average precision. Macro-mAP is important for ICIR because instances have different numbers of query images and modifications. If one instance contributes many

easy queries, global mAP can be skewed toward that instance. Macro-mAP first averages within each instance group and then averages across instances:

$$\mathit{Macro\text{-}mAP}@100 = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \frac{1}{|g|} \sum_{q \in g} AP_q@100. \tag{3.15}$$

This metric checks whether improvements are broadly distributed across instance identities rather than concentrated in a few easier groups.

The thesis reports Macro-mAP because the benchmark contains multiple modifications per instance and the number of queries per instance is not perfectly uniform. If a method improves mostly on identities with many queries, the global mean can overstate how broadly useful the method is. Macro-mAP is a stricter companion metric: it asks whether each instance group benefits on average. The final method improves both the global and macro means, which strengthens the interpretation of the result.

4

Methodology

This thesis treats instance-level composed image retrieval as the conjunction of two subproblems:

1. **instance verification**: does the candidate image depict the same literal instance as the query image?
2. **modification verification**: does the candidate image visually satisfy the textual modification?

The full pipeline is designed around this decomposition. The first stage uses a frozen dual encoder to retrieve and score candidates efficiently, following the scalable embedding-retrieval interface established by CLIP-style and SigLIP-style models [2, 3, 4, 5]. The second stage then verifies the shortlisted candidates with an additional source of evidence. In the final experiments this second source takes two forms: an ILIAS-trained linear-probe dual-encoder route and a descriptor-guided RZEn multimodal route [11, 12]. Both routes reuse the same first-stage shortlist and the same first-stage image and text scores. In both cases, the final score is multiplicative because the task behaves like a logical "and": a candidate should score highly only when the identity evidence and the modification evidence are both strong.

Figure 4.1 summarizes the main two-stage design in the visual style of the BASIC pipeline. The query image and modification first pass through a SigLIP2 stage that computes a centered image score and a contextualized text score inside the query-local instance-HN pool. The top-100 candidates and the two stored first-stage scores are then passed to a second-stage verifier. In the dual-encoder route, the verifier contributes an additional image score from an ILIAS-trained probe and a contextualized text score from the corresponding frozen backbone. In the RZEn route, an instance descriptor is generated for the query, RZEn uses this descriptor to produce descriptor-guided candidate embeddings, and the final ranking multiplies the two first-stage scores with the two RZEn scores.

Figure 4.2 shows the BASIC-style retrieval pipeline that motivates the first-stage experiments. In this thesis, the same family of components is implemented and audited, but the strongest final first-stage recipe is simpler than the full BASIC stack.

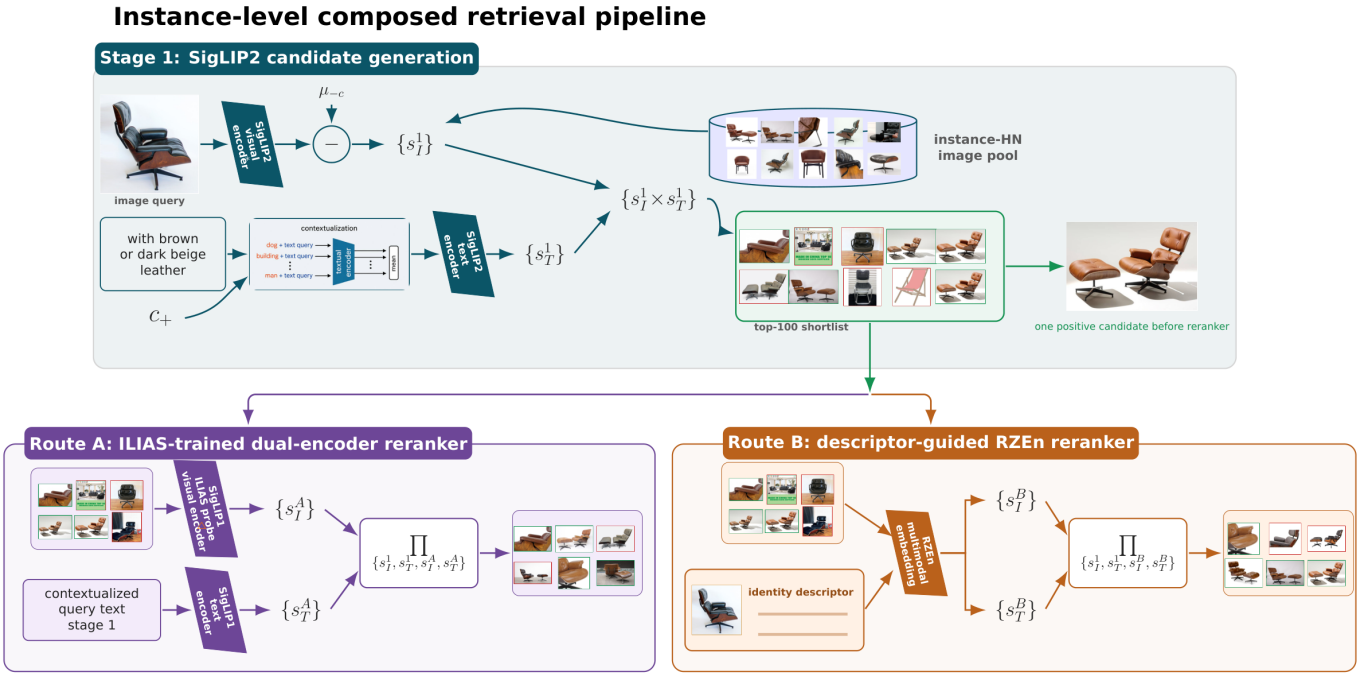

Figure 4.1: Final two-route retrieval pipeline used in this thesis. Stage 1 is the shared SigLIP2 candidate-generation baseline: the query image is encoded and leave-one-instance-out centered to produce the image score s_I^1, while the modification text is contextualized and encoded to produce the text score s_T^1. Their product retrieves the top-100 shortlist from the query-local instance-hard-negative pool, and both Stage 1 scores are stored for reranking. Route A is the ILIAS-trained dual-encoder reranker: it adds an ILIAS linear-probe visual score s_I^A and a SigLIP1 contextualized text score s_T^A. Route B is the descriptor-guided RZEn reranker: an MLLM-generated identity descriptor guides the candidate embedding, producing RZEn image and text scores s_I^B and s_T^B. The final score in each route is a four-factor product that preserves the Stage 1 baseline evidence while adding one route-specific verification signal. The visual example is the real ICIR query `q_0180`, an Eames lounge chair with the modification "with brown or dark beige leather"; green borders mark positives in the local shortlist.

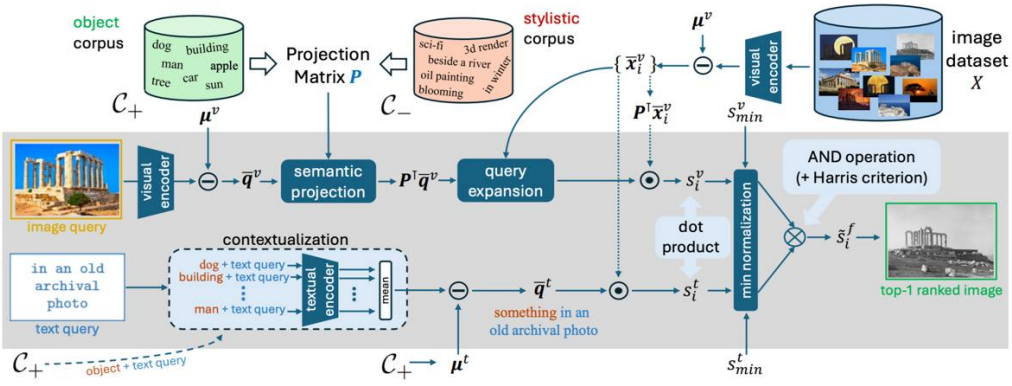

Figure 4.2: Training-free BASIC-style pipeline used as the starting point for the first-stage experiments [1]. The benchmark pipeline improves the image and text branches through operations such as centering, contextualization, projection, query expansion, and score calibration before fusing them. This thesis reimplements these components and evaluates which ones remain useful when the backbone is replaced by stronger modern embedding models.

4.1 First-stage retrieval

Base score. Let x_q be the query image, t_q the textual modification, and d a candidate image. A dual encoder provides an image encoder f_I and a text encoder f_T, as in the CLIP, OpenCLIP, SigLIP, and SigLIP2 family of retrieval models [2, 3, 4, 5]. The normalized embeddings are

$$z_q^I = f_I(x_q), \qquad z_q^T = f_T(t_q), \qquad z_d^I = f_I(d). \tag{4.1}$$

The image and text scores are

$$s_I(q,d) = \langle z_q^I, z_d^I \rangle, \qquad s_T(q,d) = \langle z_q^T, z_d^I \rangle. \tag{4.2}$$

The simplest composed score is

$$S(q,d) = s_I(q,d)s_T(q,d). \tag{4.3}$$

This product is the starting point for all first-stage experiments. Addition was also tested, but it performed much worse because it allows a high score in one branch to compensate for a low

score in the other.

The base score is useful because it exposes the simplest possible hypothesis: a good target should be similar to the query image and similar to the modification text. It also provides a clean reference for later operations. If a component improves retrieval, it should improve over this simple product rather than only over image-only or text-only retrieval. In the experiments, the raw product is already much stronger than either branch alone, which supports the conjunctive interpretation of the task.

However, the raw product has two weaknesses. First, the image score is computed in the original embedding space, where common dataset-level directions may dominate. Second, the text score is computed from the raw modification, which may be a short fragment rather than a caption-like sentence. The next two operations, image centering and text contextualization, address these weaknesses directly.

Leave-one-instance-out image centering. The most important first-stage operation is image centering. For a query belonging to instance c, I compute a mean image embedding μ_{-c} using database images outside that instance and subtract it from both query and candidate image embeddings:

$$\tilde{z}_q^I = z_q^I - \mu_{-c}, \qquad \tilde{z}_d^I = z_d^I - \mu_{-c}. \tag{4.4}$$

The centered image score becomes

$$s_I^c(q, d) = \left\langle \frac{\tilde{z}_q^I}{\|\tilde{z}_q^I\|_2}, \frac{\tilde{z}_d^I}{\|\tilde{z}_d^I\|_2} \right\rangle. \tag{4.5}$$

The leave-one-instance-out design prevents leakage from the same instance into the centering vector. In practice, centering removes dataset-level visual biases and reveals more discriminative identity directions. I also tested text centering, including external caption corpora such as Flickr-style natural captions, but text centering did not improve the final SigLIP2 setting. This is plausible because the text branch is much less visually redundant than the image branch: the modification is a short phrase, and contextualization adds only a small set of generic linguistic wrappers. There are fewer repeated visual-background-like directions to remove. Database

images, however, are centered consistently on the image side.

The leave-one-instance-out detail is important. If the center included images from the same instance, the operation could accidentally use information from the evaluation identity itself. That would make the centered representation less clean as an evaluation protocol. By excluding the current instance, the center represents broad dataset structure rather than the target identity. This keeps the operation training-free while reducing the risk of leakage.

Intuitively, centering subtracts a common visual bias. Many image collections have dominant directions corresponding to dataset style, object frequency, background distribution, or image quality. These directions can make two candidates appear similar for reasons unrelated to exact instance identity. After centering, the score is more sensitive to residual directions that distinguish the query from the general dataset. This is why centering produces one of the largest first-stage gains in the results.

TEXT CONTEXTUALIZATION. ICIR modifications are often short fragments, while CLIP/SigLIP-style text encoders are trained mostly on caption-like sentences. Text contextualization addresses this mismatch by embedding multiple harmless sentence-like variants of the same modification and averaging them. The additional words come from the generic subject/context corpus distributed with the BASIC implementation. They are not instance labels and they do not describe the target candidate. They only make the modification phrase look more like a natural caption. For each modification, the implementation embeds both orders, generic context before the modification and generic context after the modification, and averages the resulting vectors. Appendix A.1 gives the full construction.

More explicitly, let $\mathcal{C} = \{c_1, \ldots, c_M\}$ denote the neutral contextualization corpus. Each c_i is a generic caption-like fragment rather than an answer-specific description. For a modification such as "at night", the contextualized set contains strings that place the phrase into a more natural linguistic environment, for example a generic context followed by "at night" and the reverse order "at night" followed by the same generic context. The implementation embeds all valid realizations with the same text encoder, normalizes the resulting vectors, and averages them. The operation is therefore a distributional correction for short text, not a source of extra

supervision.

Let $g_k(t_q)$ be a contextualized realization of the modification. The contextualized text embedding is

$$z_q^{T,\mathrm{ctx}} = \frac{1}{K} \sum_{k=1}^{K} f_T(g_k(t_q)). \tag{4.6}$$

The contextualization procedure does not add relevant new information. It randomly wraps the modification with generic linguistic context before or after the phrase, embeds the variants, and averages the result. This makes the text input more sentence-like while preserving the intended modification. This operation is particularly suitable for CLIP/SigLIP-style dual encoders because their text towers are trained on natural image captions. It was less consistently useful for MLLM-based embedding models such as Qwen, VLM2Vec, and RZEn, whose training already includes more flexible instruction and sentence lengths.

This distinction matters because contextualization can easily be misunderstood as adding semantic knowledge. It does not use the instance name, the target image, or labels from the candidate pool. It only changes the linguistic form of the query text. The average over many neutral contexts makes the resulting vector less dependent on one accidental phrasing. In practice this helps short modifications because the embedding becomes closer to the text distribution expected by the encoder.

The thesis also tested caption-like variants generated from query images, including short and long instance descriptions. Those variants were useful for analysis, but they did not replace the simpler contextualized modification in the final first stage. The reason is that richer text can introduce extra information that is not always aligned with the requested modification. A longer caption may emphasize background, pose, or current state, and these details can distract from the modification branch.

Projection, query expansion, normalization, and Harris-style fusion. The thesis also implements the main BASIC-style enhancements beyond centering and contextualization [1]. These operations are important because they test whether the gain comes from a complete retrieval recipe or from a smaller number of robust components:

1. **projection**, which applies an image-side transformation before scoring;
2. **local query expansion**, which updates the query representation using nearby retrieved images;
3. **score normalization**, which attempts to make image and text scores more comparable before multiplication;
4. **Harris-style fusion**, which penalizes strong disagreement between the image and text branches.

Projection is intended to remove or reorient directions in the embedding space that are not useful for the retrieval task. In a weak or biased embedding space, this can improve the alignment between the query and the desired candidate. In an instance-level setting, however, projection has a risk: the directions that look like nuisance variation globally may also contain fine-grained identity cues locally. For this reason, projection is evaluated rather than assumed to help.

Local query expansion starts from an initial ranking and updates the query representation using nearby retrieved images. The intuition is that the neighborhood can provide a more stable estimate of the query intent. This is useful in many retrieval settings, but it is delicate in instance-level composed retrieval. The top neighbors may include hard negatives that are visually close to the instance but do not satisfy the modification, or images that match the requested state while depicting the wrong instance. Averaging such candidates into the query can blur the very distinction the benchmark is designed to test.

Score normalization attempts to make the image and text branches comparable before fusion. This matters because the two branches can have different score ranges, and a product can be dominated by the branch with the larger dynamic range. I evaluated both the official min-score scaling idea and local variants. These normalization variants were useful diagnostically, but they did not improve the strongest SigLIP2 configuration.

Harris-style fusion introduces an explicit disagreement penalty:

$$F(a, b) = ab - \lambda(a + b)^2, \tag{4.7}$$

where a and b are the image and text scores after scaling, and λ controls the strength of the quadratic penalty. This is the BASIC-style analogue of the Harris corner response: the product term rewards joint support from both branches, while the penalty term discourages rankings dominated by broad one-sided score effects. The rule is conceptually attractive for a conjunctive task, but in the final SigLIP2 setting it did not outperform the simpler centered contextualized product. The empirical lesson is that additional calibration can help in some regimes, but it can also reduce useful score contrast when the backbone is already strong.

These components are therefore retained as important ablations rather than as part of the final first-stage recipe. They helped or remained competitive for weaker backbones, but projection, query expansion, min-score normalization, and Harris-style fusion did not improve the final SigLIP2 SO400M Patch16 384 result. This distinction is useful for the thesis: stronger backbones do not always benefit from heavier hand-designed retrieval machinery.

Projection is the clearest example. It is attractive because it can suppress broad nuisance directions, but in a strong instance-sensitive embedding space those same directions may carry small identity cues. Removing them can make the representation more semantically regular while making it worse at distinguishing positives from hard negatives. This is also why centering the ILIAS-trained linear-probe image space did not give the same kind of gain as centering the original frozen SigLIP2 space: the probe has already shifted the representation toward instance evidence, so an additional global centering step can remove part of the very signal the probe was trained to emphasize.

This negative result is methodologically important. It prevents the thesis from presenting the final pipeline as a blind accumulation of every available refinement. The experiments show that each component must be revalidated when the backbone changes. In particular, instance-level retrieval is sensitive to operations that smooth, project, or normalize too aggressively. A component that improves global semantic alignment can still remove small visual cues needed for exact identity.

Final first-stage formula. The final first-stage score used for the main thesis pipeline is

$$S_{\text{stage1}}(q, d) = s_I^c(q, d)\, s_T^{\text{ctx}}(q, d). \tag{4.8}$$

The score is computed only inside the query's own ICIR candidate pool. The top-100 candidates become the input to the second-stage verifier. Importantly, Stage 1 is not only a high-recall candidate generator. Its two scores are reused in the final four-variable reranking products.

Reusing the first-stage scores is a deliberate design choice. The first stage contains useful information even after reranking begins. If the second stage replaced the first stage completely, the final method would depend entirely on the additional verifier. Instead, the final systems treat the first-stage scores as two reliable evidence terms and ask the second stage to provide two additional verification terms. This makes the final score easier to interpret and also reduces the risk that a second-stage signal overrides a strong first-stage signal without support.

4.2 Second-stage verification routes

The first-stage dual encoder is fast and strong, but many of its remaining mistakes are fine-grained. A candidate may contain a visually similar but different product, a different version of a character, a landmark viewed under a confusable condition, or the correct object without the requested change. These cases require an additional verification signal. This thesis evaluates two complementary ways to obtain that signal.

The first route remains in the dual-encoder family. It asks whether the shortlist can be improved by adding a second frozen backbone, and then whether an ILIAS-trained linear probe makes the image branch more instance-sensitive. This route is attractive because it is relatively cheap: the embeddings are compact, the scoring remains a dot-product computation, and the learned component is only a small projection head on top of frozen features.

The second route uses an MLLM-based embedding reranker. Such models are trained not only to align captions and images but also to follow multimodal instructions. RZEnEmbed-7B is used because it preserves an embedding-retrieval interface while allowing the candidate image representation to be guided by text. The architecture shown later in Figure 4.4 is the

same RZEn embedding architecture used by the reranker; the figure is discussed there together with the final instance-focused instruction.

Both routes use the same role split. The first stage supplies recall and coarse ranking; the second stage supplies more precise comparison among plausible candidates. The verifier is not used as a general search engine over the full database. It only reranks the top-100 candidates returned by the first stage. This design is important because it makes stronger or more specialized models usable without paying their cost over the full candidate pool.

4.3 Auxiliary linear-probe reranking

The ILIAS experiments show that frozen embeddings contain instance-level information that can be exposed by a small supervised head [11]. This is the first of the two identity-preservation routes studied in the thesis: use supervised instance retrieval to train a small projection that makes same-instance evidence more explicit. The linear probe is a learned projection

$$h(z) = \frac{Wz}{\|Wz\|_2}, \tag{4.9}$$

where the backbone embedding z is frozen and only the projection W is trained. The probe is trained on ILIAS as an instance-retrieval task, not by fine-tuning the original image encoder. This keeps the adaptation lightweight and makes the experiment a test of whether instance information is already present in the representation.

In the ICIR reranking pipeline, the probe is not used to replace the first-stage product. Direct replacement can make the image side too dominant and can reduce compatibility with the original text branch. Instead, the probe is used as an additional second-stage image-verification factor. For the strongest SigLIP2 first-stage setting, the competitive linear-probe route uses

$$S_{\text{probe}}(q,d) = s_{S,I}(q,d)\, s_{S,T}(q,d)\, s_{P,I}(q,d)\, s_{B,T}(q,d), \tag{4.10}$$

where $s_{P,I}$ is the image score from the ILIAS-trained probe and $s_{B,T}$ is the contextualized text score from the corresponding frozen second backbone. For the SigLIP2 first-stage experiments,

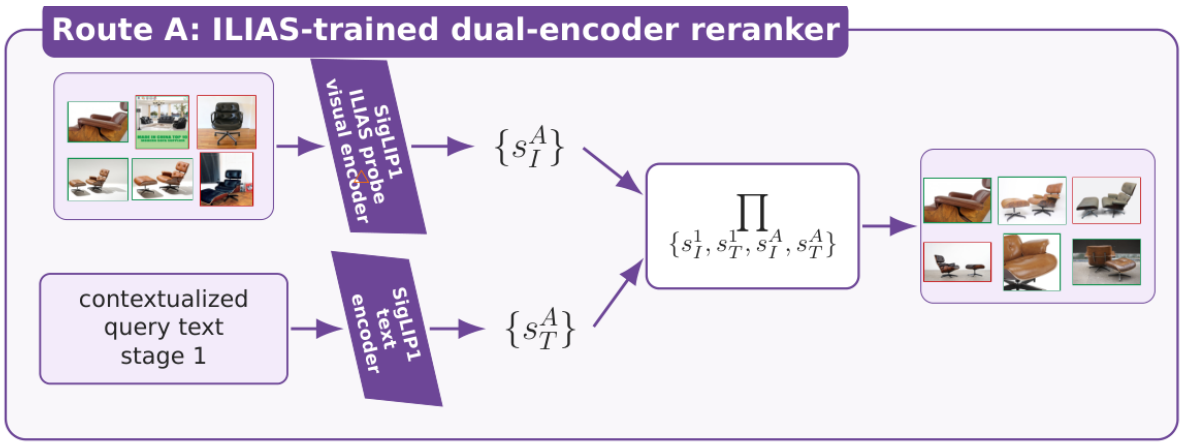

Figure 4.3: Route A of the final pipeline. This branch reranks the Stage 1 top-100 shortlist with an ILIAS-trained dual-encoder verifier. The candidate images are scored with a SigLIP1 linear probe trained on ILIAS instance-retrieval supervision, producing s_I^A, while the modification side uses the SigLIP1 contextualized text score s_T^A. These two scores are multiplied with the stored Stage 1 scores s_I^1 and s_T^1, so Route A adds supervised instance-focused evidence without discarding the original composed-retrieval baseline.

this second backbone is SigLIP1. For the CLIP ViT-L/14 first-stage experiments, the second backbone is SigLIP2.

This design is deliberately balanced. The probe contributes a stronger instance-oriented image signal, but the additional contextualized text factor keeps the composed retrieval objective from collapsing into pure instance retrieval. This is why the probe route is treated as a second full pipeline rather than as a small diagnostic: in the final results it becomes competitive with the RZEn route while remaining much cheaper computationally. It also explains why directly replacing the original image branch with the probe is not sufficient: the probe changes the image neighborhood substantially, but without a matched text-side adaptation the modification branch can become underweighted.

4.4 Instance descriptor generation

The reranker needs to know which instance in the query image should be preserved. This is the second identity-preservation route studied in the thesis: rather than training a probe, use a multimodal language model to produce an identity descriptor and then prompt the reranker

to focus on that descriptor. The descriptor is not manually written. It is generated offline with open-source multimodal language models, including Qwen-family models used as auxiliary description generators [13]. The prompt asks the model to describe the literal instance using discriminative identity cues while avoiding background details and avoiding leakage of the requested modification.

Several descriptor variants were tested, including short five-line captions, longer five-line captions, future-state captions, and modification-aware rewrites. The descriptor instruction was selected on a validation subset: several prompt families were generated, inspected, and evaluated before the most reliable identity-oriented version was used for the full dataset. The most useful version for the final reranker was an exact identity-oriented descriptor line. This is intentionally different from a generic caption. A generic caption may describe the scene, style, or even the current state of the image, which can entangle the identity branch with the modification branch. The descriptor used here is instead meant to say what the instance is and which visual cues identify it. Appendix A.3 and Appendix A.4 provide the full descriptor-generation details and the RZEn prompt text.

The descriptor-generation prompt was designed with two constraints. First, the descriptor should identify the instance, not the target state. The modification can help the generator identify which object in the image is relevant, but the descriptor should not leak the desired modified condition. Second, the descriptor should avoid unnecessary background information. Background can be useful in some landmark or scene cases, but for many objects it can distract the reranker from the stable identity cues.

The validation process was partly automatic and partly manual. Several prompt families were run on sample queries, and their outputs were inspected for common failure modes: generic captions, background-dominated descriptions, mention of irrelevant text in the image, leakage of the modification, or descriptions that were too vague to guide retrieval. The selected descriptor format was the one that best balanced specificity and stability. This matters because the reranker can only be as useful as the signal used to guide it.

4.5 Descriptor-guided multimodal reranking

Candidate embedding. For each shortlisted candidate d, RZEn receives the image together with a fixed instance-retrieval instruction and the query's instance descriptor u_q [12]. The instruction used in the final reranker is summarized in Figure 4.4. Its purpose is to make the candidate representation focus on the literal object, product, character, or landmark described by u_q, rather than on generic scene captioning.

The candidate embedding is

$$e_d^{\text{inst}} = f_{\text{RZEn}}(d, u_q, p_{\text{inst}}), \tag{4.11}$$

where p_{inst} is the fixed instance-focused instruction.

The same descriptor is used for all top-100 candidates of a query. This makes the reranking comparison consistent: each candidate is judged with respect to the same identity guide. The descriptor does not tell the model which candidate is positive; it only defines the visual identity that should be checked. The candidate image still determines the resulting embedding.

Reranker image and text scores. The descriptor-guided RZEn route embeds the candidate once under the instance-focused instruction. That same candidate embedding is compared with two query-side RZEn embeddings:

$$s_{R,I}(q, d) = \langle e_d^{\text{inst}}, e_q^I \rangle, \tag{4.12}$$

and

$$s_{R,T}(q, d) = \langle e_d^{\text{inst}}, e_q^T \rangle. \tag{4.13}$$

This is a key clarification. The RZEn text score in the reported descriptor-guided pipeline does not come from a separate modification-aware candidate prompt. The candidate is embedded with the instance-aware instruction, and the resulting representation is compared to the RZEn text embedding of the modification. Modification-aware candidate prompts were tested, but they did not improve the reported four-variable result and are not used in the main

thesis pipeline.

This design keeps the RZEn route efficient. Embedding each candidate once is cheaper than embedding it separately under an identity prompt and under a modification prompt. It also keeps the method conceptually cleaner: the candidate representation is made instance-aware, and then the same representation is compared against both query-side signals. The results suggest that this simpler design is sufficient for the main gain.

FOUR-VARIABLE MULTIMODAL PRODUCT. Let $s_{S,I}$ and $s_{S,T}$ be the first-stage SigLIP2 centered-image and contextualized-text scores. Let $s_{R,I}$ and $s_{R,T}$ be the RZEn reranking scores. The descriptor-guided RZEn score is

$$S_{\text{final}}(q,d) = s_{S,I}(q,d)\, s_{S,T}(q,d)\, s_{R,I}(q,d)\, s_{R,T}(q,d). \tag{4.14}$$

This formula has two roles. First, it adds a second opinion from a different model family. A candidate verified by both SigLIP2 and RZEn receives amplified support, while a candidate supported by only one model is suppressed. This is an ensemble effect, but it is stricter than averaging two rankings. Because the four signals are multiplied, agreement is amplified and disagreement is penalized. If both backbones assign high identity and modification evidence to the same candidate, the product grows sharply; if either backbone rejects the candidate, the product drops. In this sense the score behaves like a cross-model verification rule. Second, descriptor guidance makes the RZEn opinion more instance-specific, which is especially useful when the shortlist contains visually or semantically similar distractors.

The multiplication also makes the method conservative. It does not reward a candidate merely because one model is confident. This is appropriate for the benchmark because false positives often arise from partial evidence. A near duplicate may have high image similarity but fail the modification. A modified image may have high text similarity but depict the wrong instance. The four-variable product is designed to suppress these partial matches.

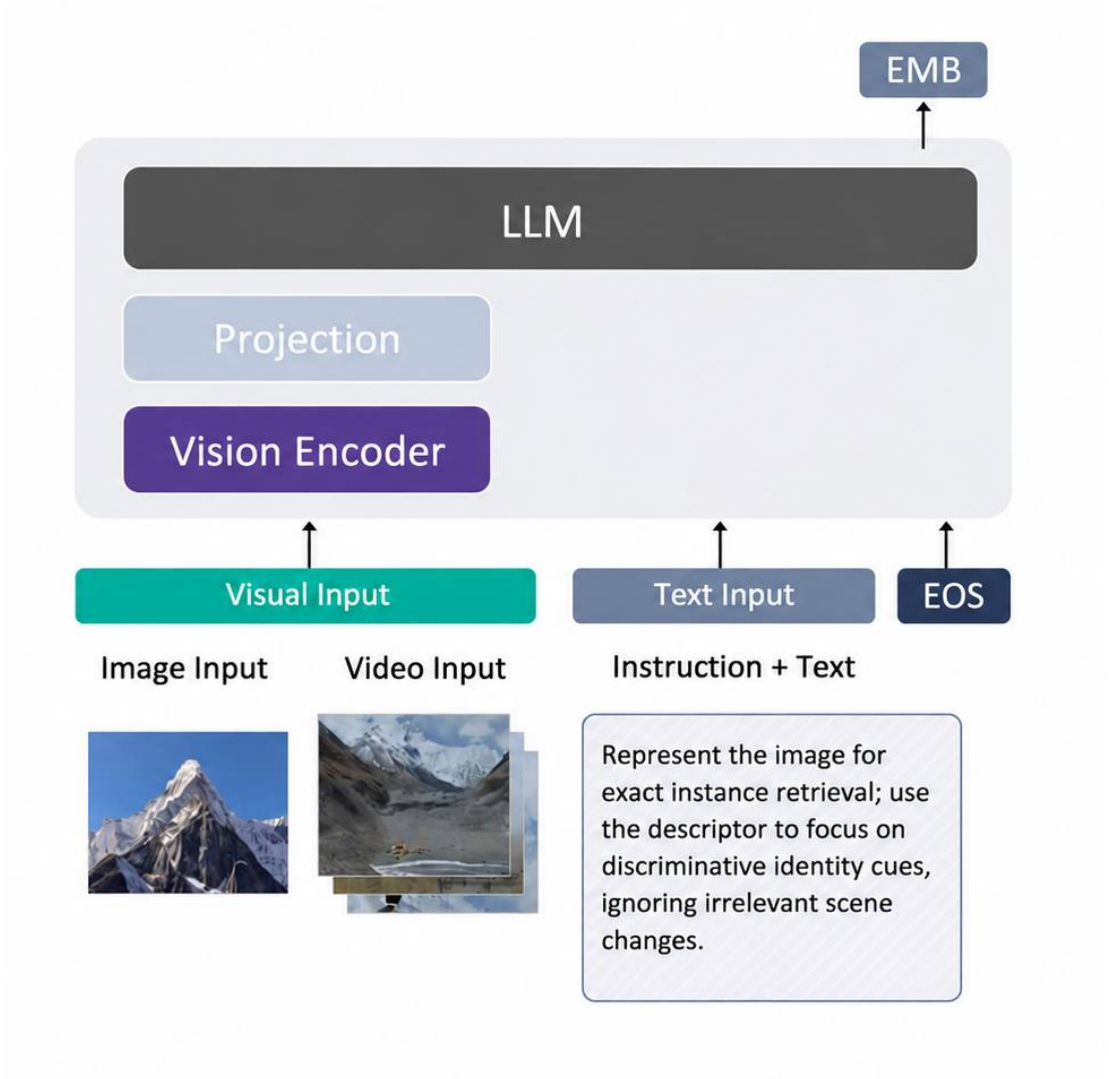

Figure 4.4: Instance-focused instruction used for the RZEn second-stage route. The instruction tells the embedding model to represent each candidate image according to the generated instance descriptor, so the candidate vector emphasizes discriminative identity cues rather than generic scene content. The descriptor-guided RZEn pipeline uses this instance-aware candidate representation for both reranker comparisons; a separate modification-aware candidate prompt was tested only as an ablation.

4.6 Plain and descriptor-guided reranking variants

Two reranking variants are used to separate the source of the gain.

Plain four-variable product. In this variant, RZEn is used without the instance descriptor instruction. It tests whether the improvement comes merely from using a second backbone. This no-instruction branch behaves like a heterogeneous ensemble between a dual encoder and an MLLM-based embedder.

Descriptor-guided four-variable product. In this variant, the candidate embedding is guided by the generated instance descriptor. It tests whether explicit instance guidance improves beyond the ensemble effect. The experiments show that it does.

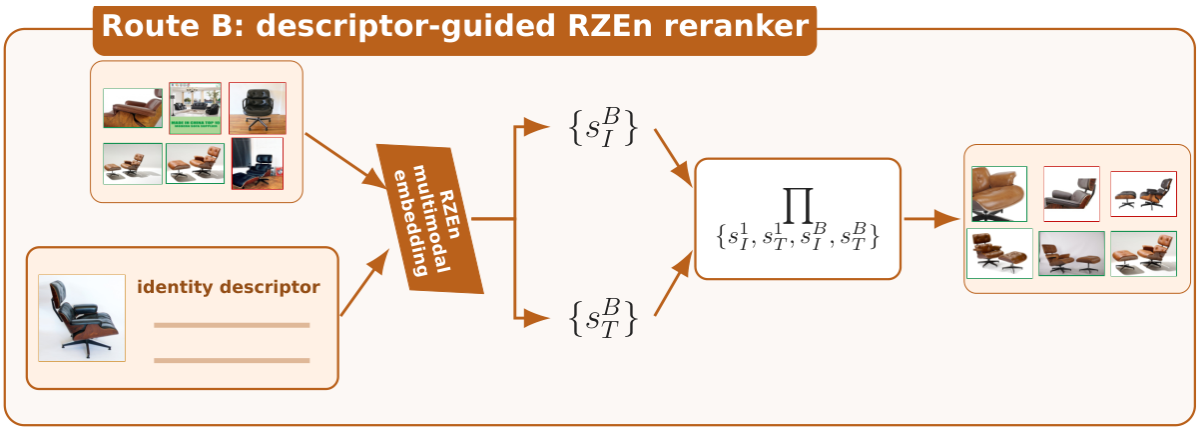

Figure 4.5: Route B of the final pipeline. This branch reranks the Stage 1 top-100 shortlist with descriptor-guided RZEn embeddings. An MLLM-generated identity descriptor tells RZEn which literal instance should be emphasized when representing each candidate image. The resulting candidate embedding is compared with the RZEn image and text query embeddings to obtain s_I^B and s_T^B, which are then multiplied with the stored Stage 1 scores s_I^1 and s_T^1. This route tests whether explicit identity guidance improves over using a second backbone alone.

4.7 Computational design

RZEn reranking is much heavier than SigLIP2 first-stage retrieval. However, after Stage 1 only 100 candidates per query must be processed. Since the same RZEn weights and the same instance instruction are reused, the mathematical computation is fixed; the practical bottleneck is how many of the 100 candidate images fit into GPU memory at once. Larger VRAM mostly increases usable batch size and reduces overhead. The implementation therefore uses cached query embeddings, cached descriptor texts, adaptive batch sizes, OOM backoff, and per-shard partial outputs.

The experiments were run on a shared GPU cluster with heterogeneous nodes. The main hardware used during the thesis included NVIDIA A100 PCIe GPUs with 40GB of memory, RTX PRO 6000 Blackwell-class nodes with approximately 96GB of memory, RTX A6000 and RTX 6000 Ada nodes with 48GB of memory, and RTX A5000 nodes with 24GB of memory. The smaller GPUs were useful for frozen dual-encoder scoring, lightweight sweeps, and shard-level jobs. The larger A100, A6000/Ada, and Blackwell nodes were used for RZEn, Qwen, InternVL, and other multimodal embedding or captioning jobs where image-token memory

dominated. The code was written to distribute shards across available idle nodes, resume partial outputs, and reduce batch size automatically after out-of-memory errors.

This reranking use case is different from next-token generation with an MLLM. The model is not asked to decode a long answer for every candidate. Instead, it performs one forward embedding pass and returns a fixed-dimensional vector. Once that vector is produced, all scoring is done with dot products. This makes the second stage more predictable than autoregressive generation: there is no sampling, no decoding length, and no output parsing for the reranker. The cost is still high because the model is large and the image tokens must be processed, but the computation is a bounded embedding computation rather than an open-ended generation process.

This distinction also matters for reproducibility. Generation experiments can vary with decoding parameters, sampling settings, and output parsing. The reranker used in the final system is deterministic once the descriptors and embeddings are fixed. The implementation stores intermediate scores and rankings so that later analysis can separate first-stage behavior, reranker behavior, and final multiplicative fusion. This was important during development because many ideas improved a subset of queries while hurting others. Keeping the score factors explicit made those tradeoffs measurable.

5

Experimental Results

This chapter presents the empirical results in the same two-stage structure introduced in Chapter 4. I first analyze the first stage, where frozen backbones are used to retrieve and score candidates inside each query's local ICIR pool. I then analyze the second stage, where those first-stage scores are combined with two different verifier families: an ILIAS-trained linear-probe dual-encoder route and a descriptor-guided RZEn route. Finally, I discuss the ILIAS experiments that explain why a lightweight probe can become competitive as a second-stage signal.

Throughout the chapter, all ICIR numbers follow the intended query-local instance-hard-negative protocol. Each query is evaluated only against its associated positives and curated hard negatives. This point is essential for interpretation. The goal is not global retrieval over unrelated distractors, but fine-grained ranking among the confusable candidates associated with the same instance family.

The main result is strong: the best first-stage system reaches 59.64 mAP@100, the ILIAS linear-probe route raises this to 66.05 mAP@100, and the descriptor-guided RZEn route reaches 66.42 mAP@100. The chapter therefore emphasizes not only which components work, but also why the improvement is meaningful: both second-stage routes improve the already strong

SigLIP2 baseline while keeping the same candidate-set recall, which means the gain comes from better ranking of difficult candidates rather than from changing the evaluation pool.

5.1 Experimental protocol and reporting conventions

All numbers in this chapter should be read under the same evaluation protocol unless explicitly stated otherwise. A query is evaluated against its own candidate pool, composed of the positives and the hard negatives associated with that query. This protocol is the one used throughout the final thesis because it matches the construction of the benchmark. It also avoids a misleading evaluation in which unrelated images from the global database dominate the candidate set. The task is not to reject arbitrary unrelated images; it is to rank confusable candidates for the same instance-level query.

The main metric is mAP@100. This metric is sensitive to the position of all positives in the ranked list, so it is more informative than a top-one or top-ten success rate for this benchmark. Recall@100 is also reported because it indicates whether the candidate set contains the relevant positives at all. In the two-stage experiments, Recall@100 has a special interpretation: if the reranker only reorders the top-100 candidates returned by the first stage, then it cannot improve Recall@100. It can only improve mAP@100 by moving already retrieved positives upward. For this reason, an unchanged Recall@100 in the reranking tables is not a weakness of the reranker; it is a consequence of the fixed-candidate reranking design.

Macro-mAP@100 is reported whenever the matching artifact was available. This metric is important because the benchmark contains multiple instances, and each instance can have a different number of queries and modifications. A global query average can be influenced by instances with more queries. Macro averaging first aggregates within each instance and then averages over instances, giving each instance group equal weight. When both mAP@100 and Macro-mAP@100 improve, the result is stronger because the gain is not limited to a small number of over-represented identities.

The tables also distinguish between first-stage results and second-stage reranking results. First-stage results compare backbones and training-free retrieval operations. Second-stage results

keep the strongest first-stage system as the base and then study which additional verifier improves the ranking of its top candidates. This separation is important because the two stages solve different computational problems. The first stage must be fast enough to search the candidate pool; the second stage can be more specialized because it only processes a shortlist.

Finally, I avoid mixing artifacts that were produced under different protocols. Some exploratory runs were performed on validation subsets, full-database candidate pools, or alternative candidate unions. These runs were useful during development, but the main tables report only results that match the final query-local protocol. When a value is unavailable in a fully comparable artifact, the table leaves it blank rather than reconstructing it from a mismatched run.

5.2 Why multiplicative fusion is the right first-stage operator

Before comparing backbones, I first tested whether multiplication is really the best way to combine the image branch and the modification branch. Table 5.1 reports this comparison for the centered SigLIP2 image branch fused with the *raw* modification branch. The purpose of the table is not to tune transformations, but to compare the fusion operators themselves under the same basic setup. Appendix A.2 gives the mathematical form and interpretation of the additional fusion operators considered during development.

Fusion operator	mAP@100	Recall@100	Macro-mAP@100
Product	50.09	88.42	50.13
Soft minimum	49.90	88.08	49.96
Harmonic mean	39.62	82.64	35.33
Minimum	18.22	64.02	15.07
Additive	15.84	54.98	22.65

Table 5.1: Comparison of first-stage score-fusion operators on the centered SigLIP2 image branch and the raw modification branch. The product is the strongest operator in this controlled comparison, reaching 50.09 mAP@100. Additive fusion performs much worse because it allows one branch to compensate for the other, which is undesirable when a candidate must satisfy both identity preservation and modification satisfaction.

The result is clear. Multiplication is not only conceptually aligned with the conjunctive structure of the task, but also empirically the strongest operator in this comparison. Soft minimum comes very close, which is not surprising because it also behaves like a strict agreement operator. Additive fusion performs much worse because it allows one branch to compensate for the other too easily, and pure minimum is too harsh in practice. This table therefore justifies the use of multiplication as the default operator in the rest of the thesis.

The gap between multiplication and addition is especially important because additive fusion is a common default in retrieval systems. If two scores are useful, summing them seems natural. In this benchmark, however, summation gives the wrong inductive bias. It rewards candidates that are very strong on only one branch. This is exactly the failure mode the benchmark is designed to expose. A candidate that is visually close to the query image but does not satisfy the modification can receive a high image score. A candidate that clearly satisfies the modification but depicts another instance can receive a high text score. Addition can keep both candidates near the top. Multiplication suppresses them unless both branches agree.

The harmonic mean and soft minimum are more conservative than addition, but they behave differently from the product. The harmonic mean is strongly controlled by the smaller factor, which can be helpful when scores are calibrated but unstable when one branch has a compressed dynamic range. The minimum operator is even more severe: it discards useful information from the stronger branch and depends entirely on the weaker score. The soft minimum is the closest alternative to the product because it behaves like a fuzzy logical conjunction. Its near-competitive result supports the same conclusion: the task benefits from an "and-like" operator. The product is selected because it is both strongest in this comparison and simplest to extend to the four-variable reranker.

This experiment also explains why later reranking uses multiplication rather than a learned or additive combination. The purpose of the final formula is not only to maximize a validation score, but to keep the evidence interpretable. If the final product improves a query, one can inspect whether the improvement comes from the original image score, the original text score, the RZEn image score, or the RZEn text score. A learned fusion layer might hide that structure. Since the thesis is partly an analysis of how identity and modification evidence interact, explicit

multiplication is preferable.

5.3 First-stage retrieval results

Backbone comparison. The next question is which frozen backbone provides the strongest first-stage retrieval quality. Table 5.2 summarizes the strongest full-query stage-one results available for the main backbone families audited in this project. For each backbone, I report the raw simple product, the strongest full BASIC-style variant, and the best overall stage-one variant. Missing entries are left blank when a fully comparable full-query artifact was not available, rather than mixing incompatible protocols. The models are ordered first by classical dual encoders and then by MLLM-derived embedders. In each numeric column, the best result is marked with a solid underline and the second-best result is marked with a dashed underline when a comparable value is available.

Backbone	Simple product mAP / R@100	BASIC-style mAP / R@100	Best stage-one mAP / Macro	Best variant
Dual-encoder backbones				
OpenAI CLIP ViT-L/14 ~428M parameters, 768 dimensions	17.48 / –	31.64 / –	41.00 / 35.85	Centered contextualized product
OpenCLIP ViT-H/14 ~1B parameters, 1024 dimensions	31.82 / 70.18	41.69 / 84.60	47.07 / 50.95	Centered contextualized product
SigLIP1 SO400M Patch14 384 ~0.9B parameters, 1152 dimensions	10.64 / 29.38	32.27 / 71.70	48.86 / 46.93	Centered contextualized product
SigLIP2 SO400M Patch16 384 ~0.9B parameters, 1152 dimensions	35.57 / 75.57	51.90 / 91.43	59.64 / 60.24	Centered contextualized product
Multimodal large language model based embedding backbones				
Qwen3-VL-Embedding-2B 2B parameters, 2048 dimensions	24.56 / 63.36	25.80 / 66.11	37.61 / 35.51	Centered contextualized product
VLM2Vec-V2.0-Qwen2VL-2B 2B parameters, 1536 dimensions	23.39 / 61.41	27.74 / 66.97	39.64 / 40.88	Centered contextualized product
RZEnEmbed-7B 7B parameters, 3584 dimensions	49.44 / 86.95	46.22 / 87.65	51.97 / –	Centered contextualized product

Table 5.2: Portrait first-stage comparison under the query-local instance-hard-negative protocol. Each row reports one backbone and groups its metrics into the raw simple product, the strongest BASIC-style result, and the best overall stage-one result. Parameter count and embedding dimension are shown under each backbone name. The previous BASIC-style state of the art is represented by the CLIP ViT-L/14 reference row from the ICIR paper [1], while the strongest first-stage result in this thesis is SigLIP2 SO400M Patch16 384 with the centered contextualized product. This reaches 59.64 mAP@100 and 60.24 Macro-mAP@100. Solid underlines mark best values and dashed underlines mark second-best values within comparable numeric columns.

The most important pattern in the table is that the strongest practical stage-one variant is not the heaviest BASIC-style stack, but the simpler *centered contextualized product*. This holds for the dual-encoder family and also for the MLLM-derived embedders when they are used in retrieval mode. The main first-stage gains therefore come from two operations:

1. cleaning the image geometry with leave-one-instance-out centering, and
2. stabilizing the short modification text with contextualization.

Among all fully audited backbones, SigLIP2 SO400M Patch16 384 is the strongest overall first-stage model. This is why it becomes the reference first-stage backbone for the final thesis pipeline.

The difference between backbones is not only a matter of parameter count. The OpenCLIP ViT-H/14 row is stronger than the original CLIP ViT-L/14 reference, but it still remains below the SigLIP2 family in the strongest full-query setting. SigLIP1 improves substantially once centering and contextualization are applied, but its raw simple product is weak. This suggests that its embedding space benefits strongly from retrieval-specific correction, yet its uncorrected image-text geometry is not as well matched to the ICIR query structure. SigLIP2, in contrast, starts from a stronger raw product and also benefits strongly from the same two key operations. This combination makes it the best first-stage backbone.

The MLLM-based embedding rows show a different pattern. RZEnEmbed-7B has a strong raw simple product relative to some dual encoders, but the full BASIC-style recipe does not improve it in the same way. This is not surprising. A model trained to support instruction-aware or multimodal embedding may not need the same text contextualization strategy as a CLIP-style text tower. The linguistic mismatch that contextualization fixes for short modification fragments is less severe when the embedding model already accepts richer instructions or more varied text forms. Therefore, the BASIC-style operations should not be expected to transfer uniformly across model families.

This is a useful warning for method design. A pipeline component is not intrinsically good or bad; it is good or bad relative to an embedding geometry. Image centering helps strongly for the final SigLIP2 first stage because it removes common visual directions and improves identity-sensitive comparison. Contextualization helps because the SigLIP2 text tower benefits from

more caption-like modification inputs. The same operations may be neutral or harmful for a different model. The thesis therefore treats the final recipe as an empirically validated combination, not as a universal checklist.

The table also motivates the two-stage design. If RZEn were clearly the strongest first-stage model, one might simply use it for all retrieval. Instead, the strongest first-stage system is SigLIP2, while RZEn becomes most useful later as a reranker. This division is computationally and empirically sensible: SigLIP2 provides fast and strong first-stage retrieval, and RZEn provides a complementary second opinion on a much smaller candidate set.

First-stage component ablation on the full benchmark. After selecting SigLIP2 SO400M Patch16 3
as the main first-stage backbone, I ran a full ablation over the main BASIC-style components on all 1,883 ICIR queries. Figure 5.1 reports the main results in the same component-style format used by the ICIR paper.

This figure defines the first-stage story of the thesis. The first rows show weak single-branch references: image-only retrieval reaches 4.71 mAP@100 and text-only retrieval reaches 6.82 mAP@100. Additive fusion is also weak, reaching only 15.84 mAP@100 even with image centering. The largest useful gain comes from multiplicative fusion and centering: the raw product starts at 35.57 mAP@100 and rises to 50.09 once image centering is applied. Text contextualization produces a second major jump, reaching 59.64 mAP@100, 93.08 Recall@100, and 60.24 Macro-mAP@100. These two components together form the strongest stage-one recipe.

The rows should be read as a controlled buildup of the first-stage recipe. The key comparison is not only between final scores, but between which components are active: the centered contextualized product is the turning point, while later BASIC-style additions mostly reduce the score for this stronger SigLIP2 embedding space.

SigLIP2-SO400M Patch16 384 - full ICIR BASIC-style components

Variant	Img	Txt	+=sum	x=×	C	CTX	Min	P	QE	mAP	mmAP	R@100
Image only	●	✗	✗	✗	✗	✗	✗	✗	✗	4.7	8.8	23.9
Text only	✗	●	✗	✗	✗	✗	✗	✗	✗	6.8	7.1	40.5
Additive + C	●	●	●	✗	●	✗	✗	✗	✗	15.8	22.6	55.0
Raw product (no C/CTX)	●	●	✗	●	✗	✗	✗	✗	✗	35.6	38.8	75.6
Product + C	●	●	✗	●	●	✗	✗	✗	✗	50.1	50.1	88.4
Product + C + CTX	●	●	✗	●	●	●	✗	✗	✗	**59.6**	**60.2**	**93.1**
+ official min scaling	●	●	✗	●	●	●	●	✗	✗	55.8	55.5	90.7
+ P	●	●	✗	●	●	●	✗	●	✗	56.4	54.6	92.5
+ P + QE	●	●	✗	●	●	●	✗	●	●	53.6	49.8	91.8
BASIC without QE	●	●	✗	●	●	●	✗	●	✗	53.8	52.1	91.6
BASIC without CTX + QE	●	●	✗	●	●	✗	✗	●	●	39.7	35.5	85.1
BASIC without P + QE	●	●	✗	●	●	●	✗	✗	●	57.7	56.6	92.7
Full LQE BASIC	●	●	✗	●	●	●	✗	●	●	51.9	48.4	91.4

Figure 5.1: Main full-dataset first-stage ablation for the strongest SigLIP2 backbone. The strongest first-stage recipe is the centered contextualized product, which reaches 59.64 mAP@100, 60.24 Macro-mAP@100, and 93.08 Recall@100. The figure also shows that image-only, text-only, and additive baselines are far weaker, while projection, local query expansion, and min-score scaling do not improve the strongest setting. Green circles indicate active components and red crosses indicate omitted components; $+$ denotes additive fusion, $\times$ denotes multiplicative score fusion, C denotes leave-one-instance-out centering, CTX denotes text contextualization, Min denotes official min-score scaling, P denotes projection, QE denotes local query expansion, and mmAP denotes Macro-mAP.

The remaining BASIC-style components did not improve the strongest SigLIP2 setting. Official min-score scaling was tested on the same full-query protocol and reduced the centered contextualized product from 59.64 to 55.75 mAP@100. Projection reduces both the global mean and the macro mean. Local query expansion pushes the score down further. Tuned Harris-style fusion was also explored, but the available tuned Harris artifacts correspond to a 400-query validation split rather than the full 1,883-query protocol used in Figure 5.1; therefore I do not mix that number into the full-dataset figure. This is an important empirical point of the thesis: once the backbone becomes sufficiently strong, additional retrieval machinery does not automatically help. In the final SigLIP2 regime, it mostly hurts.

The projection result deserves attention because projection is intuitively appealing. A projection can remove nuisance directions and may improve alignment between query and target distributions. The negative result here suggests that the nuisance directions are not easy to separate from useful instance cues. In an instance-level benchmark, small visual details can be decisive. A transformation that smooths the space or removes directions that look globally unimportant can also remove the local identity evidence needed to distinguish positives from hard negatives. This is one reason the thesis does not simply stack every BASIC component into the final system.

The query-expansion result has a similar interpretation. Query expansion often helps in classical retrieval because top neighbors provide a better estimate of the query intent. In ICIR, however, the top neighbors can include near misses: images that share the instance but not the modification, or images that share the modification but not the instance. Incorporating such neighbors into the query representation can blur the two-part relevance condition. The method then becomes less strict about which evidence should remain fixed and which evidence should change.

The min-score scaling result also illustrates the difficulty of calibration. Score normalization is attractive because image and text scores can have different ranges. However, the lower tail of a score distribution is not always a stable calibration anchor. In a query-local hard-negative pool, the weakest candidate under one branch may still contain useful information under the other branch. Scaling by such values can compress or exaggerate score differences in ways that

hurt the product. This does not mean that score normalization is unimportant in general; it means that the particular normalization tested here is not beneficial for the strongest SigLIP2 setting.

The full first-stage ablation therefore leads to a compact but important conclusion. The strongest training-free first-stage system consists of exactly two useful corrections: leave-one-instance-out image centering and modification text contextualization. These two operations directly address the two halves of the relevance condition. Centering improves identity comparison, and contextualization stabilizes modification comparison. More global operations that smooth, project, or normalize the space are less reliable for this fine-grained benchmark.

5.4 Auxiliary evidence of extractable instance structure

Why the auxiliary evidence matters. The ILIAS experiments have two roles in the thesis. First, they are diagnostic: they ask whether frozen SigLIP2 SO400M Patch16 384 embeddings already contain enough structure for a small learned head to recover instance-level retrieval under stronger supervision. Second, after the second-stage experiments in Table 5.6, they become methodologically important because an ILIAS-trained probe is competitive as an additional verifier in the ICIR pipeline.

The most informative setup is leave-one-domain-out evaluation. In each run, one domain is excluded from training and used only for testing. If a lightweight probe trained on the remaining domains still performs well, then the frozen representation already contains transferable instance-level information rather than merely memorizing within-domain category cues.

Macro leave-one-domain-out results. Table 5.3 summarizes the leave-one-domain-out ILIAS results aggregated over the eight held-out domains.

Setting	Macro-mAP@100	Macro-Recall@100
Frozen SigLIP2 baseline	31.39	52.92
Best pairwise-sigmoid probe	57.71	80.32
Best local-InfoNCE probe	71.31	89.98
Best multi-negative-margin probe	73.36	92.31
Best overall probe per domain	73.26	91.90

Table 5.3: Leave-one-domain-out auxiliary instance-retrieval results on frozen SigLIP2 embeddings. Only a lightweight projection head is trained, while the backbone remains frozen. The large improvement from 31.39 to 73.26 Macro-mAP@100 shows that frozen embeddings already contain substantial instance-level information that can be extracted with a small supervised head.

The gain is large and consistent. The frozen baseline reaches only 31.39 Macro-mAP@100, while the best overall probe configuration raises this to 73.26. Macro-Recall@100 rises from 52.92 to 91.90. This shows that the frozen SigLIP2 space already contains substantial instance-level information that can be extracted by a lightweight head, even under strong out-of-domain separation.

The size of this gain is important for interpreting the rest of the thesis. If the frozen representation did not contain instance-level information, then a linear probe would have little chance of recovering it under leave-one-domain-out evaluation. The probe cannot invent new visual evidence that is absent from the frozen embedding. It can only reweight and reorient the information already encoded by the backbone. The fact that a small head produces such a large improvement therefore indicates that the frozen space contains a latent same-instance structure that is not fully exposed by raw cosine similarity.

The comparison between loss functions also supports this interpretation. All three supervised objectives improve substantially over the frozen baseline, but the local ranking objectives perform best. Pairwise sigmoid improves the space, but InfoNCE and multi-negative margin training are more directly aligned with retrieval because they compare positives and negatives within the same local decision context. The best margin-based probe slightly outperforms the other objectives, suggesting that explicitly enforcing separation between positives and hard negatives is useful for instance retrieval.

This does not mean that the probe learns a general semantic classifier. The split is by high-level domain, so the held-out categories are not used during training. If the improvement came

only from memorizing category-specific cues, performance would be expected to collapse on the excluded domain. Instead, every held-out domain improves strongly, as shown in Table 5.4. The result is therefore better interpreted as a learned shift toward instance-sensitive geometry.

PER-DOMAIN LEAVE-ONE-OUT EVIDENCE. Table 5.4 shows the same result broken down by held-out domain.

Held-out domain	Frozen baseline mAP@100	Best probe mAP@100	Δ mAP@100
art	22.29	64.21	+41.92
fashion	36.97	79.09	+42.12
household	26.77	71.89	+45.12
landmark	20.75	53.24	+32.49
media	36.99	82.11	+45.12
product	40.34	89.16	+48.82
technology	37.29	75.32	+38.03
toy	29.71	71.09	+41.38

Table 5.4: Per-domain leave-one-domain-out results for the auxiliary instance-retrieval experiment. Each row uses one domain only for testing and trains on the remaining domains. The consistent positive deltas show that the probe does not merely memorize a single domain; it transfers the instance-retrieval signal across object types, products, landmarks, media, and other categories.

Every held-out domain improves strongly. This is exactly the behavior that supports the thesis interpretation: frozen SigLIP2 embeddings already encode substantial instance-level structure, and that structure is not confined only to the domains seen during training. The ILIAS branch therefore supports the claim that instance-level information is *extractable* from the frozen space. This matters directly for ICIR because the later second-stage results show that an ILIAS-trained probe can provide a competitive image-verification signal when it is used as an additional reranking factor rather than as a replacement for the whole composed score.

The domain-level table also shows that the improvement is not confined to one type of visual identity. Product, media, household, toy, and fashion instances all improve, even though the visual cues defining identity differ across them. A product may be identified by shape and design pattern. A media instance may be identified by character or artwork structure. A landmark may require global geometry and facade layout. A toy may require color, part configuration, and small marks. The fact that all domains improve suggests that the probe is not learning one

narrow cue, such as logo matching or color matching. It is shifting the representation toward a broader notion of same-instance retrieval.

The landmark domain has a smaller absolute score than some other domains even after probing. This is also informative. Landmarks often have large viewpoint variation, lighting variation, occlusion, and background dependence. They may also share architectural features with nearby or stylistically similar structures. A single global embedding can struggle to preserve all the geometric evidence needed for exact landmark identity. This limitation reappears in the ICIR qualitative failure cases, where scene-level and landmark-like queries can be more sensitive to background and viewpoint.

Transferring the auxiliary probe back to composed retrieval. As an additional diagnostic, I first tested whether the ILIAS-trained linear probe could be transferred back into the ICIR pipeline by directly replacing or augmenting the SigLIP2 image branch. This experiment is important because the ILIAS results show that the probe learns useful instance-level structure, but ICIR requires a different composed objective: the representation must preserve the query instance while still allowing the modification branch to control the final ranking.

I therefore compared the frozen centered SigLIP2 image score with the centered image score obtained after applying the ILIAS-trained probe. Both scores used their own leave-one-instance-out centering vectors. The probe was then tested in two ways. First, I replaced the frozen image branch in the first-stage product while keeping the same contextualized SigLIP2 text score. Second, I used the probe image score only as an additional reranking factor on top of the frozen first-stage score. Table 5.5 summarizes the result.

ICIR setting	mAP@100	Macro-mAP@100	Recall@100
Frozen centered image × contextualized text	59.64	60.24	93.08
Probe centered image × contextualized text	55.95	58.49	90.54
Frozen product × probe image	50.07	52.91	93.08
Frozen image × contextualized text^2 × probe image	59.68	60.74	93.08
Soft balanced probe variant	60.11	60.90	93.08

Table 5.5: Transfer of the auxiliary SigLIP2 linear probe to the composed retrieval benchmark. The probe is strong for pure instance retrieval, but directly replacing the frozen image branch reduces mAP@100 from 59.64 to 55.95. A balanced use of the probe as an additional image-verification signal is more stable and slightly improves the first-stage macro result.

The direct replacement is worse than the frozen ICIR first stage: mAP@100 decreases from 59.64 to 55.95, and Recall@100 decreases from 93.08 to 90.54. This shows that the learned ILIAS geometry is not automatically aligned with ICIR's composed retrieval objective. However, the probe is not simply redundant. In an image-only top-10 comparison, the frozen centered image ranking and the probe centered image ranking share only 5.31 candidates on average out of 10. Only 66 of the 1,883 queries have identical top-10 sets, while 33 queries have no overlap at all. The probe therefore changes the image-neighborhood structure substantially.

The key limitation is that this transfer adapts only the image side. The ILIAS probe is trained for instance retrieval on image embeddings, not for composed image-text retrieval. ICIR, however, needs the image branch and the modification branch to be calibrated together. Without a corresponding learned text encoder or a probe trained with modification-aware supervision, the transferred image space can become more instance-focused while the text score remains in the original SigLIP2 space. This mismatch explains why the direct replacement fails even though the ILIAS probe is strong on its own task.

This explains why the most stable transfer variant is not replacement, but balanced augmentation. If the final score contains two image factors and only one text factor, the ranking becomes too identity-heavy and the modification branch is underweighted. Repeating the contextualized text factor,

$$s(d) = s_{\text{frozen},I}(q,d) \cdot s_{\text{frozen},T}(t,d)^2 \cdot s_{\text{probe},I}(q,d),$$

recovers the frozen baseline and slightly improves Macro-mAP@100. A softer balanced variant reaches 60.11 mAP@100 and 60.90 Macro-mAP@100. The gain over the frozen first stage is less than one mAP point, so this branch is not adopted as the final method. Still, the query-level behavior is informative: the text-repeated probe variant improves 825 queries, worsens 740, and ties 318, with a mean absolute query-level change of 6.44 AP@100 points. This explains why the dataset-level mean changes only slightly even though the image embedding neighborhood is substantially altered.

This result is useful because it explains why the later competitive probe route is formulated as a balanced second-stage product rather than as a direct image-branch replacement. The learned probe substantially changes the instance-centered image geometry, but without a properly matched text branch its gains and losses mostly cancel. In other words, the ILIAS probe supports the claim that instance information is extractable from SigLIP2, while also showing that ICIR needs a composed image-text solution rather than an image-only adaptation.

This branch also clarifies a possible misunderstanding. A method can improve pure instance retrieval and still fail to improve composed retrieval. The reason is that composed retrieval is not only identity retrieval. It is identity retrieval under a modification constraint. If the image branch becomes much sharper while the text branch remains unchanged, the product can become imbalanced. The system may retrieve the same instance more aggressively but become less sensitive to the modification. The small improvement of the balanced probe variant shows that the probe contains useful information, but the lack of a matched text-side adaptation limits its value.

For future training-based work, this result is encouraging rather than discouraging. It suggests that the next step should not be a purely image-side probe, but a composition-aware adapter that preserves calibration between the image and text branches. Such an adapter could use the same instance-level signal discovered in ILIAS while also training against composed positives and hard negatives. The current thesis already shows a practical version of this idea: when the probe is used as one factor in a four-variable second-stage product, and is paired with a contextualized text score from the same second backbone, it becomes competitive with the stronger RZEn route.

5.5 Second-stage reranking results

Why the second stage is more than a pure reranker. The second stage starts from the top-100 candidates returned by the first stage, but its role is not only to reorder those candidates. In every four-variable formulation, the first-stage image score and the first-stage contextualized-text score remain part of the final product. In other words, the first stage contributes both the shortlist and two of the final four multiplicative factors. The second stage then adds two new factors from another source: either a dual-encoder/probe branch or the RZEn multimodal branch.

Dual-encoder and multimodal reranking. The main reranking question is whether the second stage helps only because it introduces another embedding backbone, or whether an instruction-guided multimodal reranker contributes something more specific. Table 5.6 compares these routes for the two first-stage backbones emphasized in the thesis. Each block starts from the best first-stage baseline, then adds a second-stage dual encoder, an ILIAS-trained linear-probe variant of that dual-encoder branch, a plain RZEn branch without descriptor guidance, and finally the descriptor-guided RZEn multimodal reranker.

First-stage backbone	Step	Added signal	mAP@100	Macro-mAP@100	Recall@100	Δ mAP
SigLIP2 SO400M Patch16 384 first stage						
SigLIP2 SO400M	Best first-stage baseline	Centered image score × contextualized modification score	59.64	60.24	93.08	–
SigLIP2 SO400M	Baseline + dual-encoder reranker	Frozen SigLIP1 centered contextualized product on the top-100	63.76	61.10	93.08	+4.13
SigLIP2 SO400M	Baseline + ILIAS linear-probe reranker	SigLIP1 ILIAS-probe image score + frozen SigLIP1 contextualized text score	66.05	64.14	93.08	+6.41
SigLIP2 SO400M	Baseline + plain RZEn reranker	RZEn image and text scores without descriptor-guided focus instruction	62.54	63.56	93.08	+2.90
SigLIP2 SO400M	Baseline + RZEn reranker	Descriptor-guided RZEn image and text scores	66.42	65.65	93.08	+6.78
CLIP ViT-L/14 first stage						
CLIP ViT-L/14	Best first-stage baseline	Centered image score × contextualized modification score	41.00	35.85	83.74	–
CLIP ViT-L/14	Baseline + dual-encoder reranker	Frozen SigLIP2 centered contextualized product on the top-100	54.93	50.99	83.74	+13.93
CLIP ViT-L/14	Baseline + ILIAS linear-probe reranker	SigLIP2 ILIAS-probe image score + frozen SigLIP2 contextualized text score	56.51	53.03	83.74	+15.51
CLIP ViT-L/14	Baseline + plain RZEn reranker	RZEn image and text scores without descriptor-guided focus instruction	51.26	45.57	83.74	+10.26
CLIP ViT-L/14	Baseline + RZEn reranker	Descriptor-guided RZEn image and text scores	53.35	47.80	83.74	+12.35

Table 5.6: Vertical second-stage ablation for the two first-stage backbones emphasized in this thesis. Each block begins with the best first-stage baseline and then adds one extra reranking source on top of the same top-100 candidate set. The dual-encoder rows test whether a different frozen backbone provides a useful second opinion. The ILIAS linear-probe rows test whether instance-supervised image adaptation transfers as an additional reranking signal when paired with the corresponding contextualized text branch. The plain RZEn rows test whether a stronger multimodal embedding backbone helps without an explicit instance-focus instruction. The descriptor-guided RZEn rows then add the instance descriptor as a focused reranking cue. For fixed top-100 reranking, Recall@100 stays equal to the first-stage candidate recall; improvements therefore reflect better ordering of already retrieved positives.

Several conclusions follow immediately.

First, RZEn is not useful as a naive replacement for the first stage. The plain RZEn pair alone falls below the SigLIP2 baseline, which means the reranker should not be interpreted as a stronger standalone first-stage retriever.

Second, dual-encoder reranking already gives a substantial gain. This shows that part of the second-stage improvement comes from heterogeneous evidence: a second backbone makes different mistakes from the first-stage backbone, and the product rewards candidates that are supported by both. The ILIAS linear-probe branch is also informative. It improves the SigLIP2 and CLIP ViT-L/14 reranking settings, suggesting that instance-focused supervision can make the second image signal more useful, even though it is still limited by the need to remain calibrated with the modification text branch.

Third, descriptor guidance with RZEn improves further in the strongest SigLIP2 setting. This shows that the final gain is not only a matter of backbone diversity. The reranker also benefits from being pointed toward the literal target instance. Relative to the first-stage baseline, the descriptor-guided RZEn route improves mAP@100 by 6.78 points and Macro-mAP@100 by 5.40 points while keeping Recall@100 at 93.08. The ILIAS linear-probe route is close behind at 66.05 mAP@100 and 64.14 Macro-mAP@100, making it a second competitive pipeline rather than only a side analysis.

It is also important to note what the thesis *does not* emphasize as the main result. Several larger reranking branches were explored, but the thesis keeps the reported systems simple: a four-variable ILIAS-probe route and a four-variable descriptor-guided RZEn route. Both isolate the contribution of the two original first-stage factors and the two added verifier factors, which makes the results easier to interpret.

The numerical difference between the plain and descriptor-guided four-variable products is one of the cleanest pieces of evidence in the thesis. The plain four-variable product reaches 62.54 mAP@100. This is already above the first-stage baseline, which means that simply adding a heterogeneous RZEn view of the same top-100 candidates is useful. However, the descriptor-guided version reaches 66.42 mAP@100. The additional gain indicates that the instruction and descriptor are not decorative prompt text; they change the candidate representation in a way

that better matches the instance-level task.

This distinction matters for the interpretation of multimodal large language model based embeddings. It would be tempting to conclude that a larger or more expressive embedder automatically improves retrieval when multiplied with SigLIP2. The results do not support that simplistic view. RZEn alone is worse than the SigLIP2 first stage, and the plain RZEn pair is not sufficient. The improvement comes from a controlled use of RZEn: keep the strong SigLIP2 evidence, restrict RZEn to the shortlist, and guide its candidate embeddings toward the target instance. The probe route supports a related but cheaper conclusion: instance-focused supervision can also provide a strong second opinion when it is not allowed to replace the composed first-stage evidence.

Both competitive second-stage routes also improve the macro metric. This is important because it means the gain is not only concentrated in query groups with many examples. The ILIAS linear-probe route raises Macro-mAP@100 from 60.24 to 64.14, and the descriptor-guided RZEn product raises it to 65.65. These macro improvements are close to the global mAP improvements, which supports the claim that the gains are broad rather than caused by a small subset of frequent instances. In the next subsection, the per-query analysis focuses on the RZEn family because it is the final multimodal reranker used in the strongest pipeline.

How broadly reranking helps. Mean mAP is useful, but it does not show how broadly a method helps across queries. Table 5.7 therefore compares the main reranking families against the same SigLIP2 first-stage baseline at the query level.

Method	Improved	Worsened	Tied	Mean Δ mAP@100 (pp)
Plain RZEn pair only	730	1010	143	-5.51
Best plain 4-variable product	925	729	229	+2.81
Final descriptor-guided 4-variable product	1134	487	262	+6.68

Table 5.7: Per-query comparison against the SigLIP2 first-stage baseline on all 1,883 benchmark queries. The final descriptor-guided reranker improves 1,134 queries and worsens 487, with an average query-level gain of +6.68 percentage points in average precision at 100. This confirms that the improvement is broad across the benchmark rather than being caused by a few isolated large wins.

This table sharpens the interpretation of the reranking stage. The plain RZEn pair alone hurts more queries than it helps. In contrast, once RZEn is used as an additional verifier rather

than a replacement, the picture changes. The no-instruction four-variable branch already improves more queries than it worsens. The descriptor-guided RZEn branch improves the broadest set of queries and produces the largest average gain among the RZEn variants.

This behavior is exactly what one would expect if the descriptor-guided RZEn branch is doing two things at once:

1. acting as a heterogeneous ensemble partner to the first stage, and
2. using the descriptor to focus that second opinion more tightly on the literal target instance.

The improved-versus-worsened counts are especially useful because a higher mean can sometimes hide an unstable method. A method could improve a few queries dramatically while hurting many others slightly. That is not the pattern here. The descriptor-guided RZEn reranker improves more than twice as many queries as it worsens. It still has failure cases, and those are discussed below, but the aggregate behavior is clearly positive. This supports using it as one of the main reported pipelines rather than treating it as an isolated lucky sweep.

The tied queries also matter. A tied query usually means that the top-100 ordering or the relevant positives were not affected in a way that changes average precision. In some cases the first-stage ranking was already strong, so the reranker has little room to improve. In other cases the relevant positives were absent from the top-100 candidate set, so a fixed-candidate reranker cannot recover them. This is why the first stage remains crucial: reranking can improve ranking quality only among candidates it receives.

Qualitative improvement and failure cases. The aggregate tables show that descriptor-guided reranking improves the mean result, but they do not show the shape of the ranking changes. I therefore also generated qualitative top-10 figures. The examples were selected to cover both strengths and failures caused by the pipeline: cases where additional instance verification brings more positives into the visible ranks, and cases where the reranker makes the ordering worse by emphasizing the wrong evidence. In each figure, candidates marked in green are positives and candidates marked in red are negatives. The score rows show how the ranking evolves from the first-stage baseline to the later reranked setting.

Figures 5.2, 5.3, and 5.4 show successful and diagnostically useful good-case examples. These examples illustrate the intended behavior of the pipeline: the centered/contextualized first stage already retrieves a meaningful candidate set, and the later reranking signals move more true positives into the visible top ranks. The third example is especially useful because it also shows how a multimodal reranker can be misled when the modification phrase is visually ambiguous. Figures 5.5 and 5.6 show the opposite behavior: cases where the RZEn-hybrid reranker substantially hurts the ranking. These failures are useful because they prevent the method from being interpreted as universally beneficial. The reranker can overemphasize descriptor-compatible visual evidence while suppressing candidates that were correctly balanced by the original SigLIP2 product. This is why the analysis reports both the number of improved and worsened queries, not only the mean mAP.

WHY THE DESCRIPTOR-GUIDED BRANCH WORKS. The best reranking result supports the main methodological claim of the thesis. Instance-level CIR is better handled when identity verification and modification verification remain coupled but not collapsed. In the final four-variable system:

1. the first-stage SigLIP2 image branch contributes identity-oriented visual evidence,
2. the first-stage SigLIP2 contextualized-text branch contributes modification-oriented evidence,
3. the descriptor-guided RZEn image score contributes a second, more expressive opinion about literal identity, and
4. the descriptor-guided RZEn text score checks how compatible that same descriptor-guided candidate representation is with the query modification.

The descriptor itself is important because it narrows the model's attention toward the correct object, product, character, or landmark, especially in multi-object scenes and fine-grained near-miss cases. But the improvement is not only about the descriptor. It is also about combining two genuinely different model families: a strong dual encoder and a promptable multimodal embedder. The instructed branch works best because it turns that ensemble into an *instance-guided* ensemble.

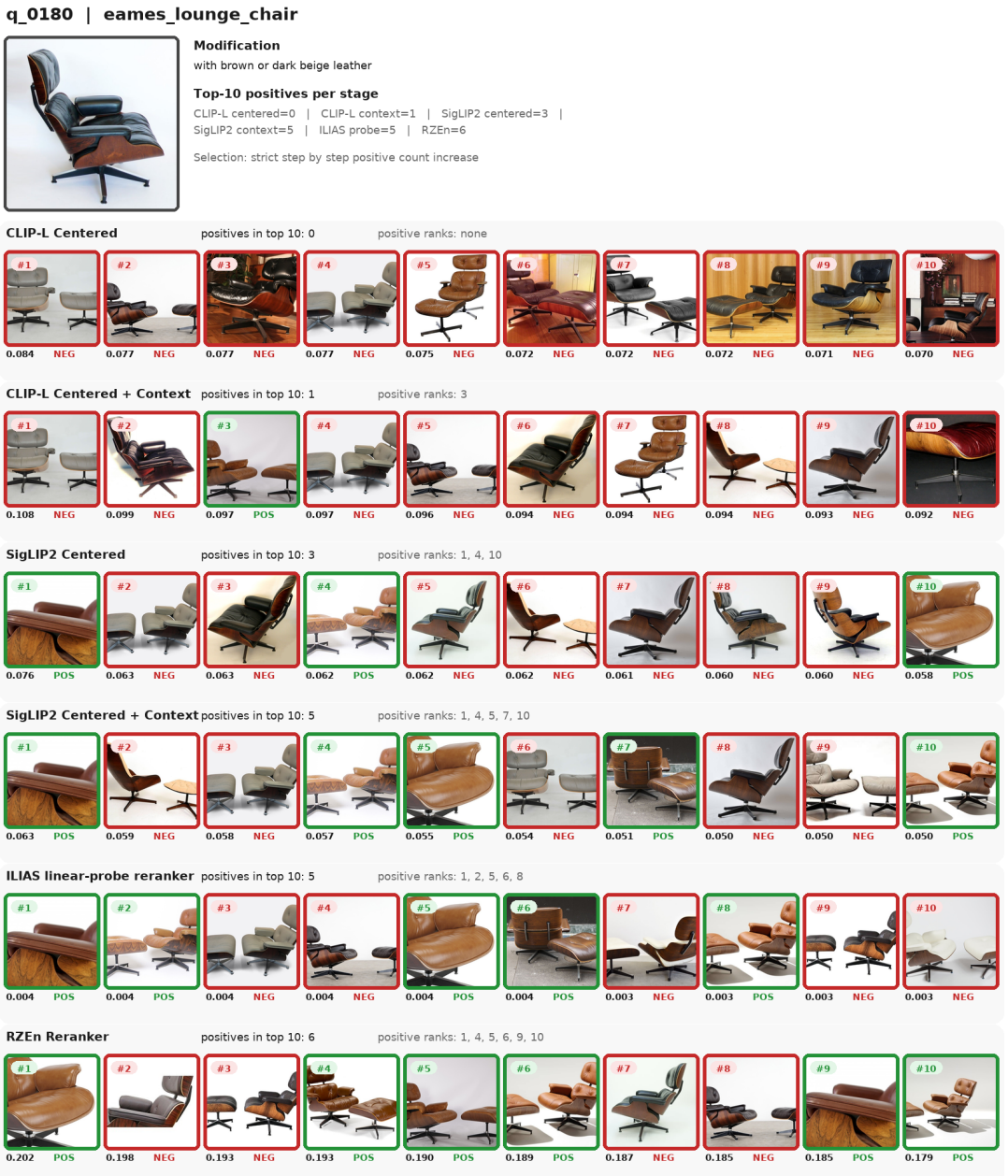

Figure 5.2: Qualitative improvement example for `eames_lounge_chair`. Candidate images marked in green are positives and candidates marked in red are negatives. The progression shows how later pipeline stages move more true positives into the visible top-10. The added ILIAS linear-probe row appears immediately before the RZEn row and shows the alternative learned instance-verification route, while the final RZEn row shows the descriptor-guided multimodal route.

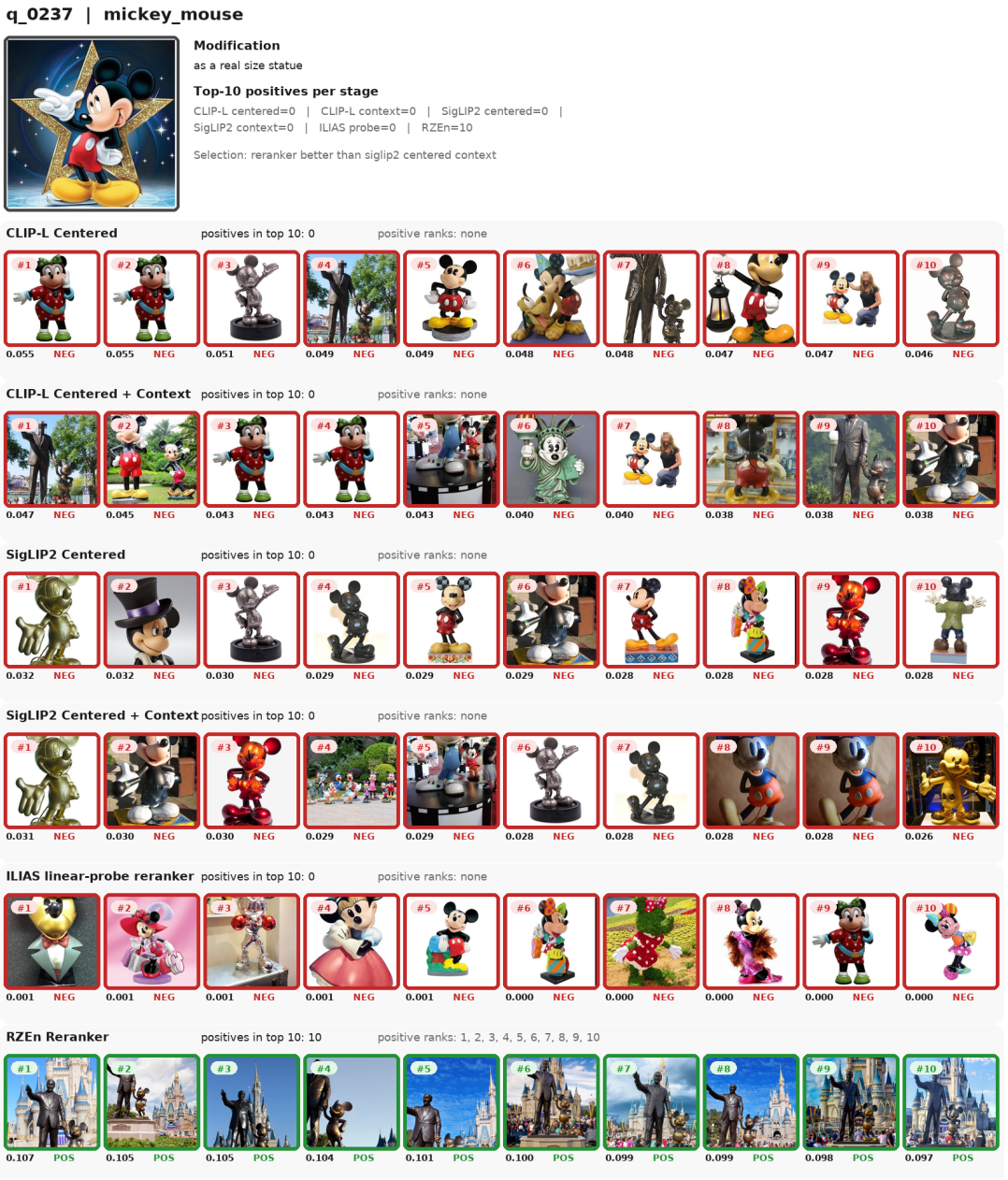

Figure 5.3: Qualitative improvement example for `mickey_mouse`. The rows compare the first-stage baselines, the ILIAS linear-probe reranker, and the descriptor-guided RZEn reranker. The example shows that different second-stage routes can change the visible top-10 in different ways: the learned probe tests whether external instance supervision helps, while the RZEn route tests whether descriptor-guided multimodal evidence improves exact-instance ranking.

Figure 5.4: Qualitative good-case example for `spongebob_squarepants` with the modification "on a yellow t-shirt". The six rows compare progressively stronger first-stage and second-stage rankings: CLIP-L image retrieval, CLIP-L with contextualized text, SigLIP2 image retrieval, SigLIP2 with contextualized text, the descriptor-guided RZEn reranker, and the ILIAS linear-probe reranker. This example is useful because it is not a purely monotonic success case. The ILIAS linear-probe route places more positives in the visible top-10, while the RZEn reranker is partly misled by the modification. Since the RZEn route combines image-side and text-side evidence in the final product, it requires a precise understanding of the phrase "on a yellow t-shirt". Here the word "yellow" can overlap with the identity of SpongeBob himself, so the modification branch can overweight a misleading cue. The example therefore illustrates both the benefit of extra instance-focused evidence and the remaining difficulty of balancing identity and modification signals.

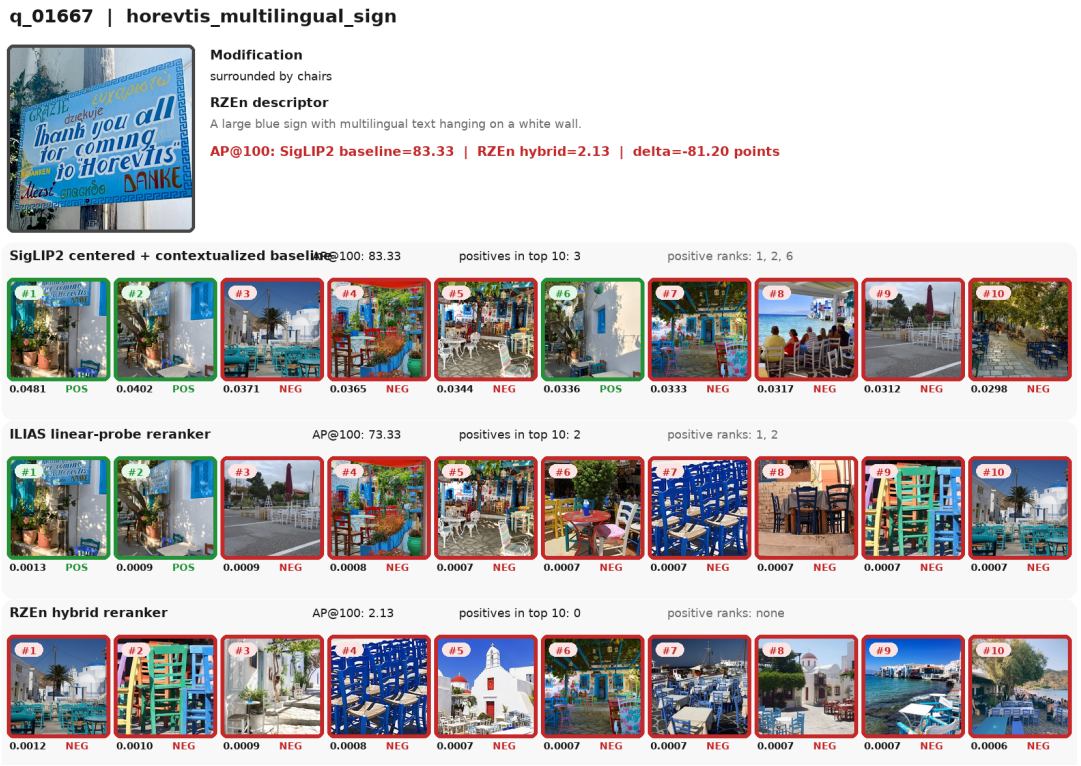

Figure 5.5: Qualitative failure example for `horevtis_multilingual_sign`. The rows show the SigLIP2 first-stage ranking, the ILIAS linear-probe reranker, and the descriptor-guided RZEn reranker. The RZEn route sharply reduces average precision for this query, while the ILIAS row gives a useful comparison point: not every second-stage verifier fails in the same way. Such cases reveal where descriptor-guided identity evidence can overpower the balance achieved by the first-stage product.

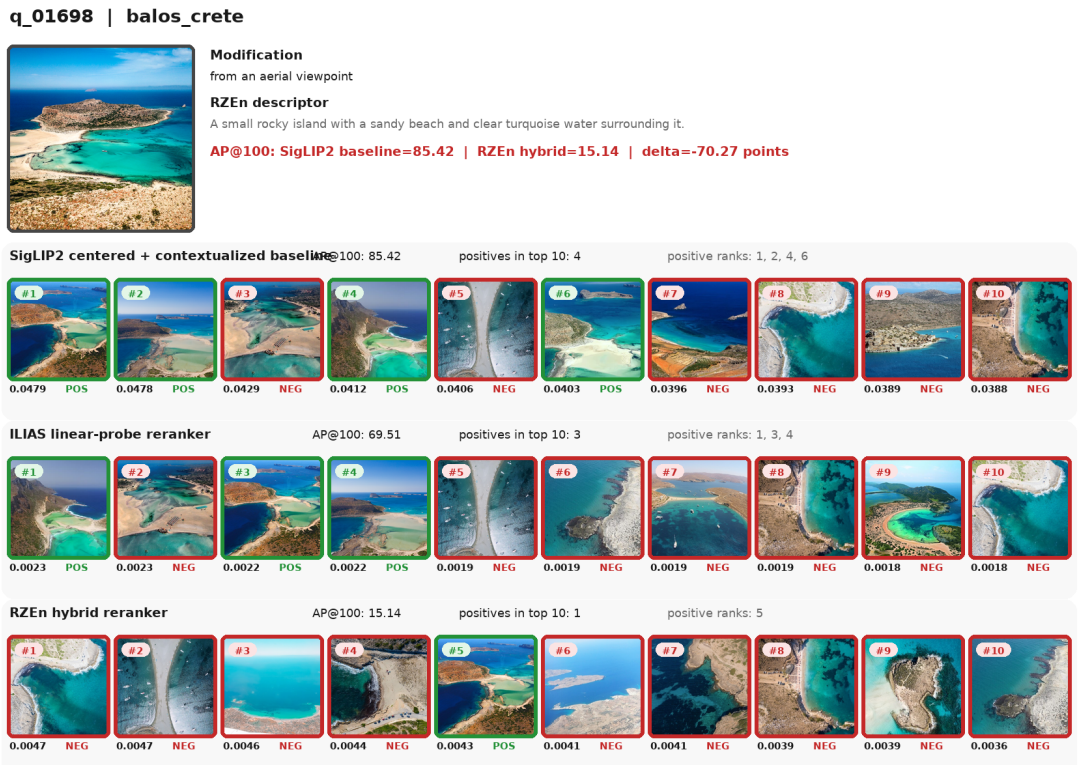

Figure 5.6: Qualitative failure example for `balos_crete`. The figure compares the first-stage ranking with both second-stage routes. The ILIAS linear-probe row is inserted before the RZEn row to separate the behavior of a learned dual-encoder verifier from the behavior of the descriptor-guided multimodal reranker. This example motivates the limitations discussed in the conclusion: reranking can improve the mean substantially while still failing on queries where the descriptor or candidate representation emphasizes the wrong visual evidence.

The descriptor-guided branch can also be understood as a way to reduce ambiguity in the candidate embedding. A candidate image may contain several objects or visual regions. A generic image embedding must compress the whole image into one vector, and the resulting representation may emphasize the most captionable or most salient elements rather than the query-relevant instance. The descriptor gives the reranker an explicit attention guide: represent the candidate with respect to this object, character, product, or landmark identity. This is why the descriptor is generated from the query image and kept fixed across the top-100 candidates.

At the same time, the descriptor-guided RZEn route deliberately avoids using the descriptor as a target-state description. The descriptor should not tell the model that the target must be "at night", "as an emoji", or "in red". That information belongs to the modification branch. If the descriptor leaked the modification, the reranker would collapse identity and modification evidence into a single generated phrase. The thesis instead keeps the descriptor as an identity guide and lets the modification text provide the state evidence. This preserves the decomposition that motivates the whole pipeline.

Other reranking experiments. Several other reranking ideas were explored during the project. They were valuable for analysis, but none replaced the final four-variable descriptor-guided product.

Difference and reverse-difference variables. These formulas compared target-conditioned and source-conditioned state embeddings in order to measure how much a candidate representation changed under different textual conditions. They were analytically useful, but they made the system harder to explain and were not necessary for the main thesis formulation.

The difference variables were motivated by a reasonable hypothesis: if a candidate image already satisfies a modification, conditioning the candidate embedding on that modification should not change the embedding much; if the modification is absent, the conditioning may move the embedding more. In practice this signal was noisy. Some correct candidates changed substantially because the model reinterpreted the image under the new text, while some incorrect candidates changed little because the modification was generic or weakly visible. The signal could help on particular queries, but it was not stable enough to replace the clearer four-variable

formulation.

Text rewriting and future-state captioning. I also explored rewriting modifications into richer sentences and generating future-state captions with open-source multimodal models. These experiments were helpful for understanding language-side failure modes, but richer text often injected extra variability rather than cleaner retrieval evidence. In the reported pipelines, the simpler contextualized first-stage text and the controlled second-stage verifiers were more reliable.

These experiments were valuable because they revealed a limitation of language-side enrichment. A more detailed text is not automatically a better retrieval query. If the generated text adds background assumptions, changes the scope of the modification, or describes the current query image rather than the desired target state, it can move the text embedding away from the intended condition. This was especially visible for short modifications whose meaning depends on the image. The RZEn route therefore uses generated language only where it is most useful: to describe the stable instance for the reranker, not to replace the modification signal in the first stage.

Another failure mode is repetition of the same evidence across branches. If multiple text or image branches encode nearly the same descriptor, the final product can over-amplify that repeated cue. This may look like stronger confidence, but it is not necessarily stronger evidence. For example, a descriptor repeated through both the image-conditioned and text-conditioned branches can make the model overfit to the most salient described attribute while ignoring subtler modification evidence. This is one reason the final thesis model keeps only two second-stage variables and preserves the original SigLIP2 image and text scores as guard rails.

Grounding and crop-based reranking. Grounding-based crops were tested to isolate the most relevant object region before reranking. This direction remains promising, but it did not outperform the main whole-image reranker in the present experiments. A likely reason is that some instances are identified not only by a tight bounding box, but also by scale, configuration, or surrounding cues. At the same time, this branch still seems worth investigating further, because many manually checked boxes were imperfect and more reliable grounding models than Grounding DINO could still unlock better crop-based verification.

The crop branch is a good example of a method that is conceptually attractive but operationally fragile. For object-centric queries, a crop can remove irrelevant clutter and help the embedding model focus on the target. For landmarks, scenes, artworks, signs, or objects whose identity depends on surrounding structure, a crop can remove evidence that is actually useful. The quality of the grounding model also matters. If the box misses part of the object or selects the wrong object in a multi-object scene, the reranker receives a worse input than the original image. This explains why the branch is left as future work rather than included in the final pipeline.

Modification-aware prompting. I also tested reranking branches that explicitly instructed the reranker to focus on whether the modification itself was present. These branches were informative, but they did not outperform the simpler descriptor-guided candidate embedding that was compared directly against both the query image and the query text. The reported RZEn route therefore keeps the modification signal implicit in the four-variable product rather than introducing an additional prompt-specific branch.

This result also clarifies the role of the RZEn text score in the reported model. The candidate embedding is generated once under the instance-focused descriptor instruction, and that embedding is compared with the modification text embedding. A separate modification-aware candidate embedding was not used in the final system. This keeps the reranker cheaper and avoids forcing the candidate representation to answer two different prompts. The product still contains modification evidence because the first-stage contextualized text score and the RZEn text comparison both depend on the modification.

Taken together, these non-final experiments strengthen the final interpretation. The best-performing systems are not simply the most complex ones tested. They are the systems that preserve the cleanest decomposition while improving the strongest baseline. The unsuccessful branches are therefore not wasted work: they show that richer language, extra variables, and localized crops must be evaluated carefully in instance-level composed retrieval, where additional information can easily shift the system away from the exact identity or modification evidence that matters.

6

Conclusion

This chapter synthesizes the findings of the thesis, discusses the limitations of the current system, and outlines future research directions. The main conclusion is that instance-level composed image retrieval benefits from treating identity preservation and modification satisfaction as separate but coupled forms of evidence. This view leads to a strong two-stage system: a SigLIP2 first stage for efficient candidate retrieval followed by an additional verifier. Two verifier routes are effective. The descriptor-guided RZEn route gives the strongest final result, while the ILIAS-trained linear-probe route is a lighter and practical alternative with nearly comparable precision. Both routes show that second-stage evidence is most useful when it preserves the original composed score instead of replacing it.

6.1 Synthesis of findings

The thesis studied instance-level composed image retrieval in a setting where a correct target must satisfy two conditions at the same time. It must depict the same literal instance as the query image, and it must visually satisfy the requested modification. This is stricter in its rele-

vance definition than ordinary semantic retrieval, although not every individual query is necessarily harder: the instance constraint can sometimes help reject candidates that are obviously different. The difficult cases are those in which identity and modification evidence pull the ranking in different directions. A candidate from the correct category is not sufficient. A candidate that matches only the requested style, color, state, or context is also not sufficient. The task is conjunctive, and the experiments show that the scoring strategy should reflect that structure.

The first important finding is that explicit multiplicative fusion is very strong for this problem. The product of image and text evidence is not only a simple baseline; it is aligned with the task definition. Additive fusion is too permissive because one branch can compensate for the other. In instance-level composed retrieval, such compensation is often exactly the failure case: a wrong instance can have high text compatibility, or a correct instance can have low modification compatibility. Multiplication is stricter and therefore better suited to the benchmark.

The second important finding is that frozen embedding backbones can be much stronger than the original baseline when the retrieval pipeline is adapted carefully. Across the audited first-stage models, SigLIP2 SO400M Patch16 384 was the strongest backbone. Its best first-stage recipe was not the heaviest BASIC-style pipeline, but the centered contextualized product. Leave-one-instance-out image centering improved identity-sensitive visual comparison, and text contextualization made short modification phrases more compatible with a caption-trained text encoder. Together, these operations reached 59.64 mAP@100 and 60.24 Macro-mAP@100 on the full benchmark.

This first-stage result is important because it shows that a large part of the gain comes from using the right frozen backbone in the right way. The model does not need to be fine-tuned on ICIR to become strong. Instead, the embedding space must be corrected and queried in a way that matches the benchmark. Centering addresses the image side by reducing common dataset directions. Contextualization addresses the text side by making short modifications more natural for the text encoder. The final product then combines these two corrected branches.

The third important finding is that not every plausible BASIC-style component helps the strongest backbone. Projection, local query expansion, min-score scaling, and several score-

normalization variants did not improve the final SigLIP2 setting. This is a useful negative result. It shows that retrieval pipelines should not be treated as a checklist of refinements. A component that helps a weaker or differently trained backbone can hurt a stronger one, especially in a benchmark where fine-grained identity cues matter. The final first-stage pipeline is therefore deliberately simple.

The fourth important finding is that reranking provides a substantial additional improvement when it is used as verification rather than replacement. The thesis shows this with two complementary routes. The ILIAS-trained linear-probe route uses supervised instance-retrieval evidence as a cheap second-stage verifier, while the RZEn route uses descriptor-guided multimodal embeddings as a heavier instruction-aware verifier. RZEn alone is not a better first-stage retriever than the strong SigLIP2 baseline. However, when RZEn contributes two additional scores on top of the SigLIP2 image and text scores, the ranking improves. The final descriptor-guided four-variable product reaches 66.42 mAP@100 and 65.65 Macro-mAP@100. This is the strongest result reported in the thesis and the clearest evidence that the proposed second stage adds value beyond the first-stage baseline.

The distinction between plain and descriptor-guided reranking is especially important. A plain no-instruction RZEn branch already improves over the first stage, which indicates an ensemble effect between two different embedding families. The descriptor-guided branch improves further, which shows that the instruction and instance descriptor contribute more than generic backbone diversity. The ILIAS route supports the same broad lesson from a different direction: a trained probe can shift the representation toward instance evidence, but it works best when it remains one factor in a composed score rather than replacing the modification-aware first stage. In other words, the rerankers help because they add focused evidence while preserving the original identity-versus-modification decomposition.

The fifth important finding is diagnostic and practical. The ILIAS experiments show that frozen SigLIP2 embeddings already contain substantial instance-level information. Under leave-one-domain-out evaluation, a lightweight linear probe raises Macro-mAP@100 from 31.39 to 73.26. Directly replacing the image branch in ICIR is not enough, because the adapted image space must remain calibrated with the modification text. However, the probe becomes

valuable when used as a second-stage verifier inside the composed product. In that role, the ILIAS-based route reaches competitive ICIR performance while being much lighter than the RZEn reranker. This shows that the thesis has two viable identity-preservation mechanisms: train a small instance-focused probe, or prompt a multimodal embedder to focus on a generated instance descriptor.

These findings support the central thesis claim. Instance-level composed image retrieval is best treated as structured verification. The system should preserve a visible identity branch and a visible modification branch, and it should combine them with an operation that reflects the fact that both conditions must hold. The final two-stage method implements this idea first with SigLIP2 and then with descriptor-guided RZEn reranking.

6.2 Limitations

The first limitation is dataset scope. The experiments are focused on the ICIR benchmark and on ILIAS as an auxiliary diagnostic source. This is appropriate for the research question, but it means that the conclusions should be tested on additional instance-level composed retrieval datasets when such datasets become available. ICIR is carefully constructed and difficult, but broader evaluation would clarify how well the proposed decomposition generalizes across image sources, object domains, languages, and modification types.

The benchmark also contains many short modifications. These are useful because they isolate a clear visual change, but real user requests can be more complex: they may combine multiple attributes, relations, style changes, object interactions, and contextual constraints in one sentence. A future benchmark with more compositional and multi-step modifications would test whether the same multiplicative decomposition remains sufficient or whether stronger language-model reasoning is required.

The second limitation is dependence on the embedding backbone. The strongest first-stage result depends on SigLIP2 SO400M Patch16 384. Other backbones follow the same general methodology, but their absolute behavior differs. Some components that help weaker models do not help the strongest one. This means that the final recipe should be understood as an

empirical result for the audited backbones, not as a universal rule that every future embedding model will follow.

The third limitation is reranking cost. The second stage processes only the top 100 candidates per query, so it is far more feasible than running a large multimodal embedder over the entire database. Nevertheless, RZEn reranking is still substantially more expensive than the SigLIP2 first stage. In practice, throughput depends on graphics memory, batch size, image resolution, and caching. This cost is acceptable for an offline benchmark and for many high-value retrieval settings, but it remains a practical constraint for very large-scale or low-latency deployment.

The fourth limitation concerns descriptor generation. The final reranker relies on automatically generated instance descriptors. These descriptors are useful because they guide the reranker toward the literal target instance, but they can still be imperfect. A descriptor may emphasize a non-discriminative visual cue, omit an important detail, or describe a scene element too generically. The qualitative failure cases show that the reranker can sometimes overemphasize descriptor-compatible evidence and harm a query that the first stage ranked better.

The fifth limitation is that the final reranker remains an embedding-based reranker, not a full cross-encoder verifier. This choice is intentional because it keeps the method scalable and allows score-factor analysis. However, a cross-encoder that jointly processes the query image, modification, candidate image, and descriptor might capture finer interactions. Such a model would be more expensive and harder to cache, but it could reduce errors where the embedding representation loses relational information.

The sixth limitation is that localization remains unresolved. Crop-based reranking and grounding experiments were explored, but they did not become the final method. Some failures clearly involve clutter or multiple objects, where localization should help. Other failures involve landmarks, signs, or scenes where tight crops can remove useful context. A better localization strategy may need to decide adaptively when to crop, when to keep context, and how to combine cropped and full-image evidence.

The seventh limitation is calibration. The final product works well because its factors are useful and reasonably compatible, but it is still a hand-designed score. The thesis intentionally avoids learned fusion for interpretability and because the benchmark setting is limited. A

future method could learn calibrated weights or transformations from a validation set, but it would need to preserve the identity-versus-modification decomposition and avoid overfitting to specific instances.

6.3 Future work

One direction for future work is stronger instance-aware training. The ILIAS experiments show that a small head can extract instance structure from frozen SigLIP2 embeddings, but the transfer back to composed retrieval is not yet fully reliable. A more complete training objective could adapt both the image and text sides, or train a residual adapter that preserves the original composed retrieval geometry while increasing instance sensitivity. Such training should be evaluated carefully because making the image branch too instance-dominant can reduce the influence of the modification branch.

A second direction is composition-aware adaptation. The ILIAS probe is image-side and instance-focused. ICIR requires an adapted representation that remains compatible with text modifications. Future work could train an adapter with triplets or local hard negatives that explicitly include both same-instance positives and modification-aware negatives. The objective should reward candidates that satisfy both conditions and penalize candidates that satisfy only one. This would align the training objective more directly with the multiplicative scoring principle used in the thesis.

A third direction is better reranking. The current RZEn reranker uses a single descriptor-guided candidate embedding and compares it to both the query image and the modification text. Future work could explore rerankers trained specifically for instance-level composed retrieval, better prompt-conditioned embeddings, or cross-encoder style rerankers for the final top candidates. The key requirement is that the reranker should improve verification without collapsing the distinction between identity and modification evidence.

A closely related direction is explicit modification verification. The current final reranker focuses mainly on instance-guided candidate representations and keeps modification evidence through the first-stage text score and the RZEn text comparison. Future systems could add a

specialized branch that asks whether the modification is visually applied, especially for subtle states such as material changes, viewpoint, lighting, action, or context. Such a branch should be designed carefully so that it does not simply repeat the identity descriptor or over-amplify generic text cues.

A fourth direction is improved descriptor generation. The descriptor should identify the stable instance without leaking the requested modification. This is a narrow generation task, and it may benefit from better prompt engineering, specialized fine-tuning, or automatic quality filters. A useful future system could generate several candidate descriptors, score their discriminativeness, and choose the one that best separates positives from hard negatives on a small validation set.

A fifth direction is improved localization. Grounding and crop-based experiments remain promising because many retrieval failures involve clutter, multiple objects, or candidate images where the relevant instance occupies only part of the image. More reliable grounding models could help isolate the object of interest before reranking. At the same time, future work must preserve enough context for instances whose identity depends on surrounding structure, scale, or configuration.

A sixth direction is larger-scale evaluation. The thesis uses the intended query-local hard-negative protocol, which is the correct protocol for this benchmark. However, practical systems may need to combine global retrieval, local hard-negative ranking, and reranking. Studying how the proposed decomposition behaves when candidate pools are generated dynamically at larger scale would make the method easier to transfer into applied retrieval systems.

A seventh direction is deeper analysis of score factors. The final method keeps the four scores explicit, which makes it possible to study which queries improve because of image evidence, which improve because of text evidence, and which require agreement between model families. Future work could use this factor-level information to build adaptive rerankers. For example, if the first-stage image score is already highly confident but the text score is uncertain, the system could allocate more computation to modification verification. If the text score is strong but the image score is ambiguous, it could allocate more computation to instance verification.

Finally, future instance-level composed retrieval systems will likely need language models as

part of the pipeline. Real modifications are often compositional and context-dependent, not just short attribute phrases. A language model can help decompose a request into stable identity cues, required visual changes, and irrelevant background details. The key challenge is to use this reasoning without leaking target information or allowing generated text to dominate the visual evidence. The thesis results suggest that language-model assistance is most useful when it is constrained to a clear role, such as descriptor generation or final verification, rather than allowed to replace the whole retrieval score.

6.4 Closing remarks

The main contribution of the thesis is both empirical and methodological. Empirically, the final two-stage system improves a strong SigLIP2 first-stage baseline from 59.64 to 66.42 mAP@100, outperforming the previous BASIC state of the art for this setting. A second competitive route based on ILIAS-trained linear probes shows that instance-focused supervision can also provide a strong and efficient verifier. Methodologically, the thesis shows that instance-level composed image retrieval should be treated as a structured verification problem rather than as generic semantic matching. The most successful pipelines keep identity evidence and modification evidence visible throughout the scoring process, first with a centered contextualized product and then with either a trained instance-aware probe or descriptor-guided multimodal reranking.

The broader lesson is that strong frozen embeddings are not passive black boxes. They contain useful identity information, but the retrieval pipeline determines whether that information is exposed, suppressed, or miscalibrated. The final system succeeds because it respects the structure of the task: the same instance must remain the same, the modification must be visible, and both conditions must be supported at the same time. This decomposition provides a clear foundation for future work on more scalable, more trainable, and more instance-aware composed retrieval systems.

A

Supplementary Methodological Details

A.1 Text contextualization details

The first-stage text contextualization used in the main SigLIP2 pipeline follows the intuition of BASIC: short modification phrases are often poorly matched to the caption-like text distribution used to train CLIP/SigLIP-style dual encoders. The aim is therefore not to add new semantic evidence, but to place the same modification inside a set of neutral sentence-like contexts before embedding.

Let t_q denote the original ICIR modification, for example "in black color" or "from an aerial viewpoint". Let

$$C = \{c_1, \ldots, c_M\} \tag{A.1}$$

be a corpus of generic context strings. In the implementation, C is read from the BASIC generic-subject corpus distributed with the ICIR codebase. For the SigLIP2 SO400M experiments I used the first $M = 100$ generic entries. These entries are deliberately broad and non-instance-specific. They are used only as linguistic scaffolding, not as labels for the image or as additional

ground-truth information.

For every modification t_q, two contextualized orders are generated:

$$g_i^{\text{pre}}(t_q) = c_i\ t_q, \qquad g_i^{\text{post}}(t_q) = t_q\ c_i. \tag{A.2}$$

The final contextualized text embedding is the normalized mean of all $2M$ embeddings:

$$z_q^{T,\text{ctx}} = \text{norm}\left(\frac{1}{2M}\sum_{i=1}^{M}\left[f_T(g_i^{\text{pre}}(t_q)) + f_T(g_i^{\text{post}}(t_q))\right]\right). \tag{A.3}$$

This is the embedding used in the final first-stage score. Importantly, the contextualization corpus does not contain the query instance name, the image filename, or the target candidate identity. It is only a set of generic phrases used to make the modification text look more like a natural caption. This distinction matters because the ICIR task should remain instance-level and should not be solved by leaking the answer through text.

As a concrete example, suppose the raw ICIR modification is "at night". The contextualization procedure does not rewrite this as a target-specific sentence such as "the Eiffel Tower at night" or "the same chair photographed at night". Instead, it produces generic variants such as "a photo of an object at night" and "at night in a natural image". Similarly, for a modification such as "made of wood", the contextualized set contains sentence-like wrappers around the same phrase rather than descriptions of any particular candidate. The average of these embeddings gives the text encoder a more caption-like input distribution while preserving the original modification semantics.

The reason this helps CLIP/SigLIP-style models is that their text encoders are usually trained on complete captions rather than isolated fragments. A two-word phrase can be meaningful to a human, but its embedding may be less stable than the embedding of the same phrase inside a short caption-like context. Averaging many neutral contexts reduces sensitivity to any one wrapper and avoids adding answer-specific information.

I also tested alternative text-centering variants. One direction used an external caption distribution, including Flickr-style natural captions, as the text-side center. Another direction

centered the raw and contextualized text using dataset-derived text statistics. These variants were useful diagnostically, but they did not improve the final SigLIP2 SO400M result. The final pipeline therefore keeps image centering but does not apply text centering.

A.2 First-stage fusion operators

The main thesis uses multiplication because ICIR relevance is conjunctive: the candidate should preserve the query instance and satisfy the modification. During development, several alternative fusion operators were tested or considered. Let $a \geq 0$ denote the image-side score and $b \geq 0$ denote the text-side score after any required score shifting or normalization. The operators are:

Product.

$$F_{\text{prod}}(a, b) = ab. \tag{A.4}$$

This is the main operator. It is high only when both branches are high, and it strongly suppresses candidates that match only identity or only modification.

Additive fusion.

$$F_{\text{add}}(a, b) = a + b. \tag{A.5}$$

Addition is more permissive. It allows a very high image score to compensate for a poor text score, or the reverse. This is not desirable for ICIR because a candidate that preserves identity but ignores the modification should not be ranked highly.

Geometric mean.

$$F_{\text{geo}}(a, b) = \sqrt{ab}. \tag{A.6}$$

The geometric mean preserves the agreement structure of the product but softens its dynamic range. It was useful as an analysis point, but the final thesis tables focus on the plain operator comparison without additional transformations.

Harmonic mean.

$$F_{\text{harm}}(a, b) = \frac{2ab}{a + b + \varepsilon}. \tag{A.7}$$

This behaves similarly to an F-score: it is dominated by the weaker branch. It is stricter than addition, but in practice it was less effective than the product in the SigLIP2 first-stage setting.

Minimum.

$$F_{\min}(a, b) = \min(a, b). \tag{A.8}$$

The minimum is an extreme weak-link operator. It completely ignores the stronger branch once the weaker branch is known. This was too harsh empirically.

Soft minimum.

$$F_{\text{softmin}}(a, b) = \frac{ab}{a + b - ab + \varepsilon}. \tag{A.9}$$

For scores in $[0, 1]$, this is a fuzzy-logic-style strict agreement operator. It performed close to the product in the first-stage fusion comparison, but did not clearly surpass it.

Harris-style disagreement penalty.

$$F_{\text{Harris}}(a, b) = ab - \lambda(a + b)^2. \tag{A.10}$$

This operator follows the BASIC-style Harris response after branch-wise score scaling. The product rewards joint evidence from the image and text branches, while the quadratic term penalizes broad one-sided dominance through a trace-like penalty. In the final SigLIP2 SO400M setting it did not improve over the simpler centered contextualized product, although it remains an interesting calibration idea for weaker or less stable backbones.

A.3 Instance descriptor generation

The RZEn reranker uses an instance descriptor to focus the candidate embedding on the literal query instance. These descriptors are generated automatically rather than manually. The query image is given to an open-source multimodal language model, and the textual modification is used only to help the model identify the salient instance in the query image. The output is then constrained to describe the stable identity of the instance, not the requested target modification.

Multiple prompting families were tested. The project explored Qwen 3.5 9B and InternVL 3.5 8B, short five-line descriptors, longer five-line descriptors, strict slot-based descriptors, future-state captions, and modification-conditioned state descriptions. The selection process used a validation subset: several instruction variants were run on held-out examples, their outputs were inspected for leakage and descriptiveness, and the most reliable variant was then used for the larger ICIR evaluation. The final reranking branch uses an exact identity-oriented descriptor line, referred to in the code as the Line-2 descriptor.

The descriptor-generation prompt followed three principles:

1. describe the literal instance, not the background;
2. avoid leaking the desired target modification into the identity descriptor;
3. produce a standalone phrase or sentence that can guide a reranker toward discriminative identity cues.

The instruction used for the identity-description generation can be summarized as follows:

> You are given a query image and a modification. Use the modification only to identify which instance in the image is the target. Do not describe the modification itself. Describe the salient literal instance as a standalone identity description. Focus on discriminative visual cues of the object, product, character, landmark, artwork, scene element, or other instance type. Avoid unnecessary background details. Do not mention image filenames, dataset names, query identifiers, or hidden labels. Do not include target-state information that would leak the answer. Produce several diverse descriptions, each describing the same instance from a different visual aspect.

For the stricter five-line descriptor variants, the prompt additionally requested diversity across lines: one line for object/category and overall form, one line for color/material/texture, one line for distinctive parts or markings, one line for viewpoint or composition, and one final standalone line emphasizing another identifying cue. During validation, outputs were checked for three common failure modes: generic captions that ignored the instance, leakage of the modification into the identity descriptor, and descriptions dominated by background or scene context. The final reranker used the most stable exact identity line rather than the full multi-line caption.

The final descriptor asset is stored as one descriptor per ICIR query. When a generated descriptor line is missing or empty, the pipeline uses a deterministic fallback: first a valid descriptor from the same query image when available, then a descriptor from the same instance when available, and finally a cleaned instance identifier only as a last resort. This fallback is used to prevent silent missing-text failures during reranking.

A.4 Reranker prompts

The final thesis reranker uses the instance-focused RZEn prompt. The full prompt used for candidate-side descriptor-guided embedding is:

> Represent the candidate image for exact instance-level retrieval. Use the provided instance descriptor as an attention guide for the literal object, character, product, or landmark identity. Prioritize visual evidence that the candidate image depicts the described instance. Ignore background, style, lighting, pose, scene context, and modification/state changes unless they help identify the instance. Do not treat the descriptor as a generic caption; use it to focus on discriminative identity cues. Instance descriptor:

A shorter thesis-facing paraphrase of the same instruction is:

> Represent the image for exact instance retrieval using the descriptor as an identity guide; prioritize discriminative object, product, character, or landmark cues.

The following modification-aware prompt was also tested in ablation experiments:

> Represent the candidate image for modification-aware retrieval. Use the provided modification text as a visual condition. Focus on whether this modification, state, action, style, location, or context is actually visible in the candidate image. Give high similarity when the candidate image visually satisfies the modification, and lower similarity when it does not. Do not focus on instance identity here; focus on the modification evidence. Modification to check:

The modification-aware branch was useful for analysis, but it is not used in the final thesis model. The final four-variable result embeds each candidate with the instance-focused instruction and compares that same descriptor-guided candidate embedding to both the RZEn query-image embedding and the RZEn modification-text embedding. This choice keeps the reranker simpler and avoids adding a second candidate embedding pass for every top-100 image.

References

[1] B. Psomas, G. Retsinas, N. Efthymiadis, P. Filntisis, Y. Avrithis, P. Maragos, O. Chum, and G. Tolias, "Instance-level composed image retrieval," *arXiv preprint arXiv:2510.25387*, 2025.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," *arXiv preprint arXiv:2212.07143*, 2022.

[4] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," *arXiv preprint arXiv:2303.15343*, 2023.

[5] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[6] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval: An empirical odyssey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6439–6448.

[7] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Pic2word: Mapping pictures to words for zero-shot composed image retrieval," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19 305–19 314.

[8] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, "Zero-shot composed image retrieval with textual inversion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 338–15 347.

[9] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "Vision-by-language for training-free compositional image retrieval," *arXiv preprint arXiv:2310.09291*, 2023.

[10] N. Efthymiadis, B. Psomas, G. Retsinas, P. Filntisis, P. Maragos, O. Chum, and G. Tolias, "Composed image retrieval for training-free domain conversion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

[11] G. Kordopatis-Zilos, V. Stojnić, A. Manko, P. Šuma, N.-A. Ypsilantis, N. Efthymiadis, Z. Laskar, J. Matas, O. Chum, and G. Tolias, "Ilias: Instance-level image retrieval at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[12] W. Jian, Y. Zhang, D. Liang, C. Xie, Y. He, D. Leng, and Y. Yin, "Rzenembed: Towards comprehensive multimodal retrieval," *arXiv preprint arXiv:2510.27350*, 2025.

[13] M. Li, Y. Zhang, D. Long, K. Chen, S. Song, S. Bai, Z. Yang, P. Xie, A. Yang, D. Liu, J. Zhou, and J. Lin, "Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking," *arXiv preprint arXiv:2601.04720*, 2026.

[14] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen, "Vlm2vec: Training vision-language models for massive multimodal embedding tasks," *arXiv preprint arXiv:2410.05160*, 2024.

[15] R. Meng, Z. Jiang, Y. Liu, M. Su, X. Yang, Y. Fu, C. Qin, Z. Chen, R. Xu, C. Xiong, Y. Zhou, W. Chen, and S. Yavuz, "Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents," *arXiv preprint arXiv:2507.04590*, 2025.

[16] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[17] O. Simeoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jegou, P. Labatut, and P. Bojanowski, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[18] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion iq: A new dataset towards retrieving images by natural language feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 307–11 317.

[19] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2125–2134.

[20] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," *arXiv preprint arXiv:2303.11916*, 2023.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Lamberto Ballan, for his guidance, feedback, and trust throughout this thesis. His suggestions helped me turn a broad set of experiments into a clearer research direction and pushed me to make the work more rigorous, better motivated, and easier to communicate.

I am also grateful to my co-supervisor, Prof. Marco Fiorucci, for his support and for the valuable perspective he brought to the project. I thank the Department of Mathematics and the Data Science Master's program at the University of Padova for providing the academic environment in which this work was developed. I am especially thankful to the Department of Mathematics for providing access to the computing cluster and related resources that made the large-scale experimental work in this thesis possible.

This thesis involved many long experimental campaigns, failed ideas, reruns, checks, and revisions. I am thankful to everyone who helped me discuss results, debug experiments, and keep perspective during that process. I also thank my family for their patience, encouragement, and constant support. Their belief in me made the difficult parts of this work much easier to carry.

Finally, I am grateful for the opportunity to work on a problem that sits at the intersection of computer vision, language, and retrieval. The process taught me not only how to build and evaluate retrieval systems, but also how to think more carefully about what a model is actually learning and what a retrieval result really means.