DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

University of Padua – Department of Information Engineering

Bachelor's Degree in Ingegneria dell'Informazione

Final Report

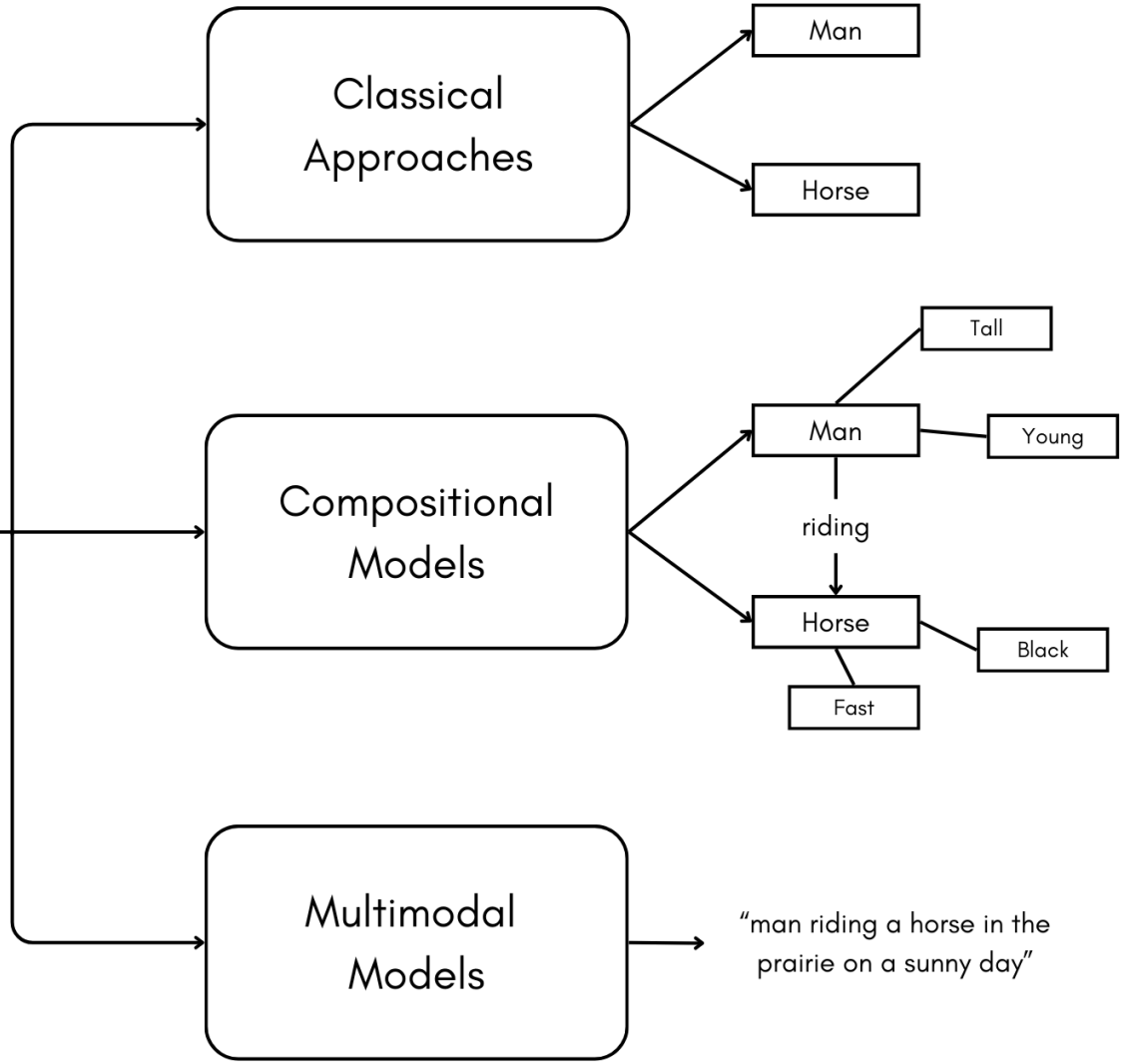
«Image Semantics Understanding»

Supervisor: Prof. Federica Battisti
Co-supervisor: Annalisa Gallina

Student's name: Francesco Roder
Student ID: 2074097

Padua, 13/03/2026

INTRODUCTION



CLASSICAL APPROACHES TO IMAGE SEMANTICS

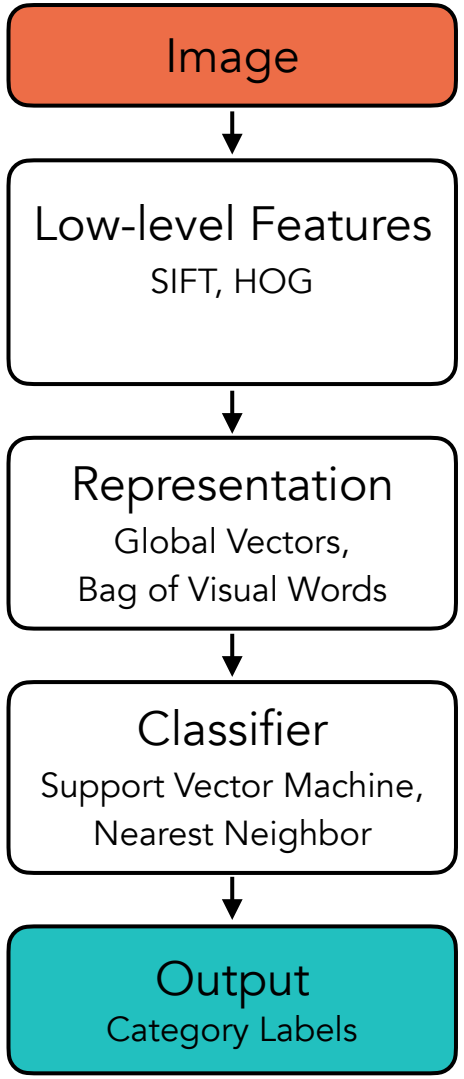
Goal: bridging the semantic gap

Classical Pipeline:

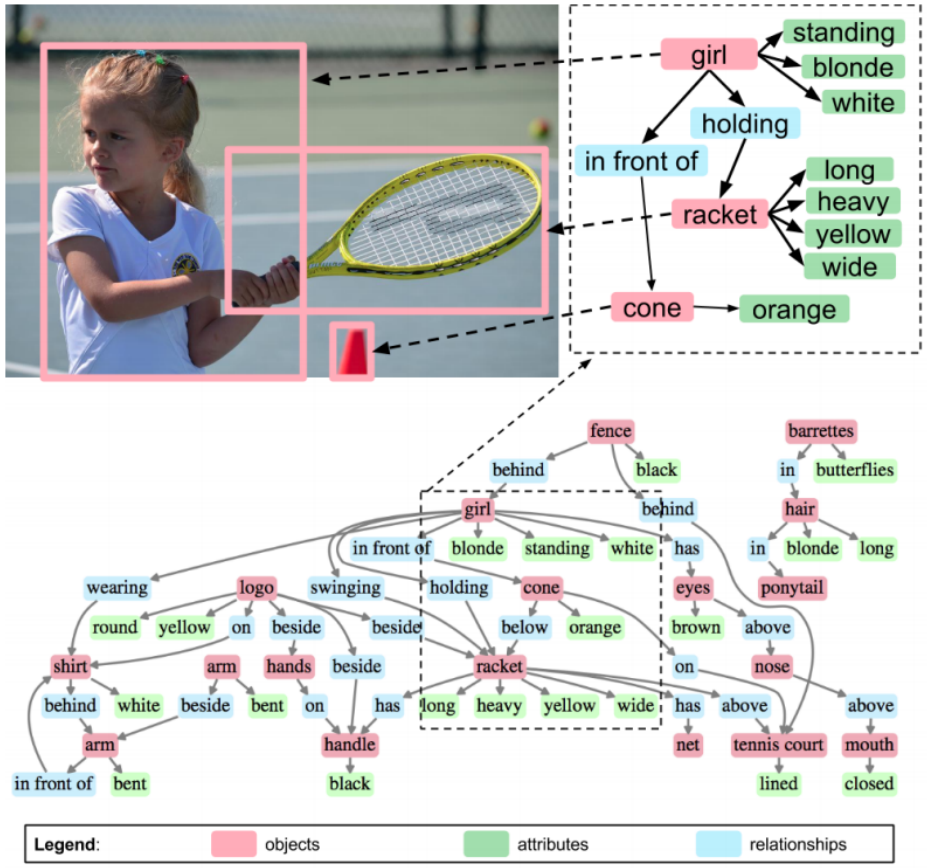
1. Hand-crafted low-level features
2. Standardized image representation
3. Supervised classifier
4. Category label(s)

Key development: Bag of Visual Words

Limitation: semantics ≠ object categorization



FROM OBJECT RECOGNITION TO COMPOSITIONAL SCENE UNDERSTANDING



Scene graph representation from Johnson et al., CVPR 2015.

Visual understanding requires:

- Attributes
- Relationships

Attribute-Based Modeling

Supports:

- Describing unseen objects
- Few-shot learning

Scene Graphs

Scene graphs encode:

- Objects as nodes
- Relationships as edges
- Attributes as nodes metadata

ADVANTAGES AND LIMITATIONS OF STRUCTURED SEMANTIC MODELS

STRENGTHS

- Fine-grained relational understanding
- Compositional structure
- Task transferability

LIMITATIONS

- Dense supervision required
  →Expensive and difficult to scale
- Graph matching is computationally expensive

LANGUAGE AS MEDIUM FOR SEMANTICS REPRESENTATION

LABELS	→	**NATURAL LANGUAGE**
Closed vocabulary		Open vocabulary
Limited semantic granularity		Rich, contextual text embeddings
Seen-class restricted		Zero-shot capable

VISION-LANGUAGE MODELS

Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models

Global Alignment → Selective Exposure → Full Multimodal Integration

VISION-LANGUAGE MODELS

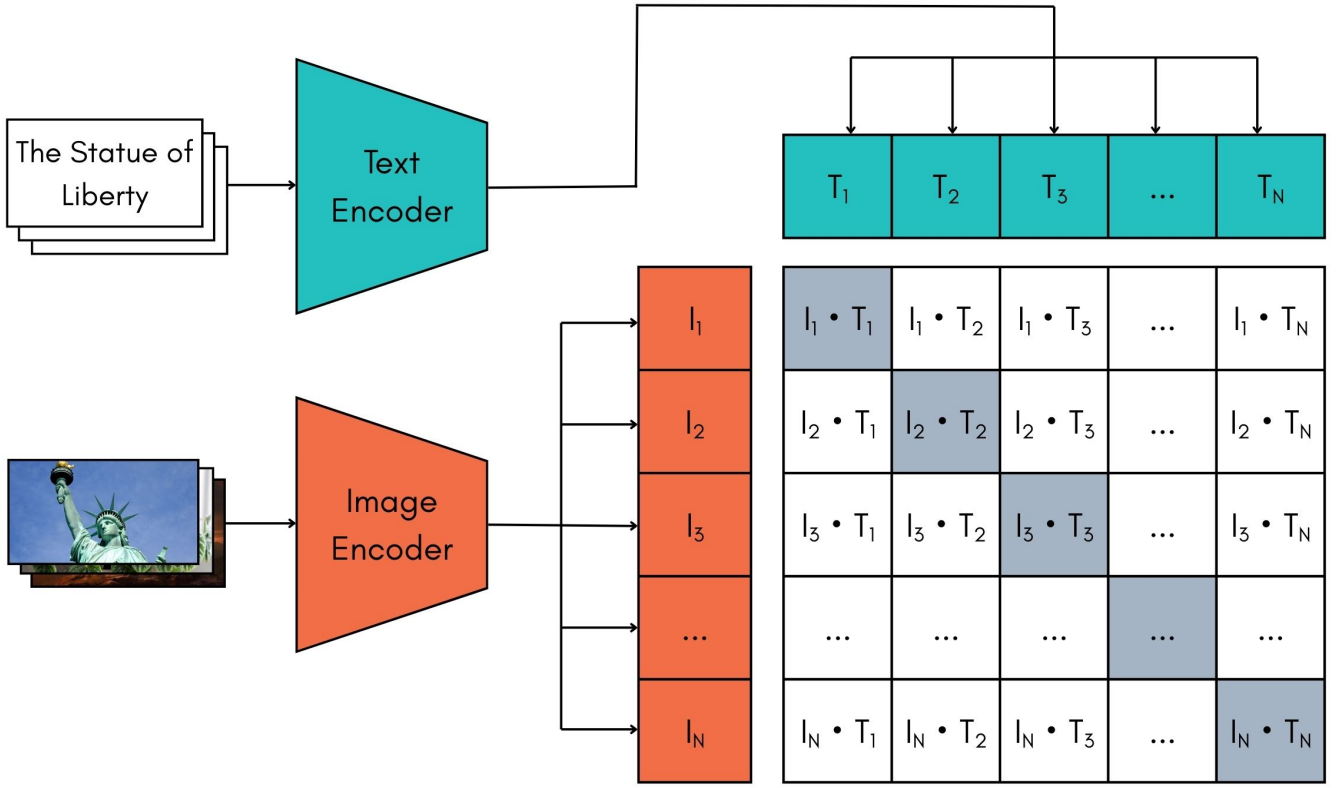
Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models

Operational Principle

- Independent encoders
- Shared latent space
- Contrastive learning
- Embedding geometry shaped by language



VISION-LANGUAGE MODELS

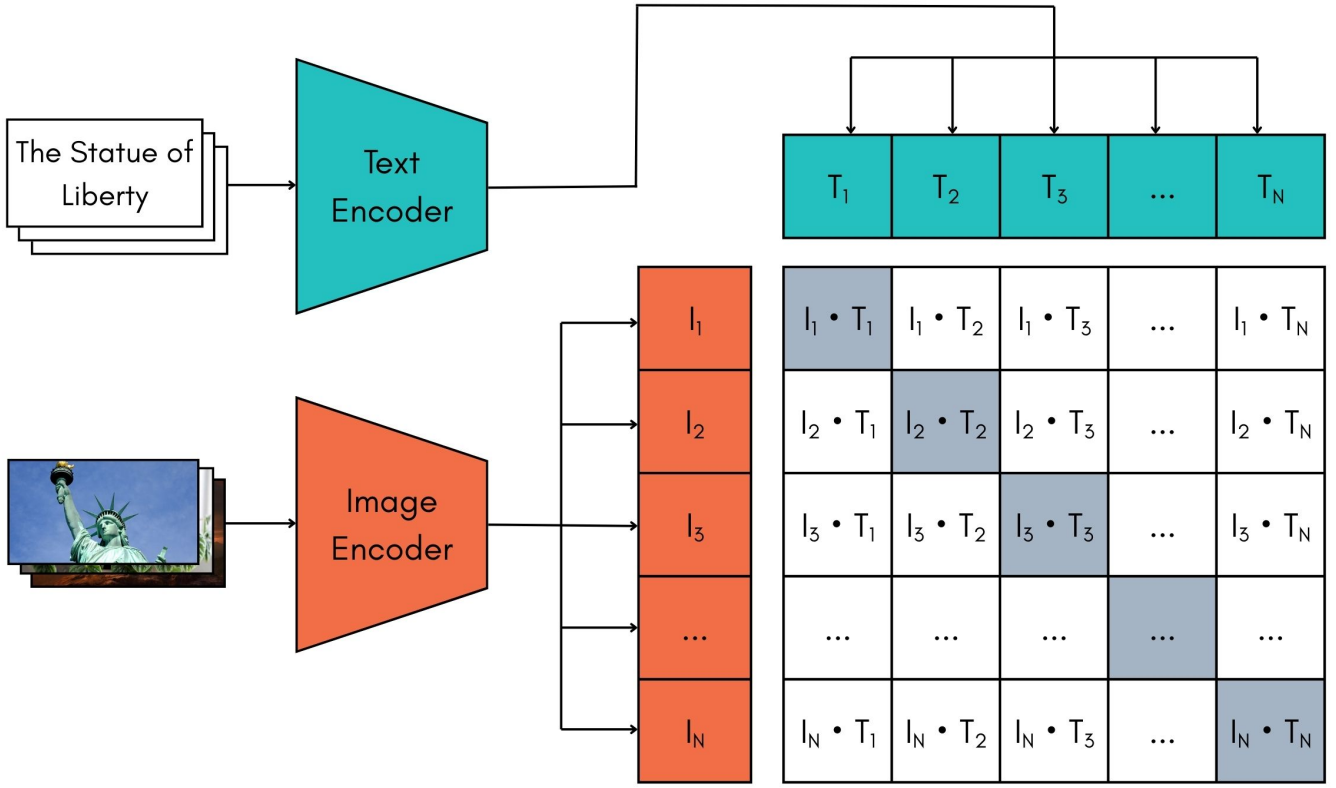
Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models

Semantic consequences

- Single global embedding
- Cross-modal interaction occurs after compression
- Limited relational and spatial reasoning

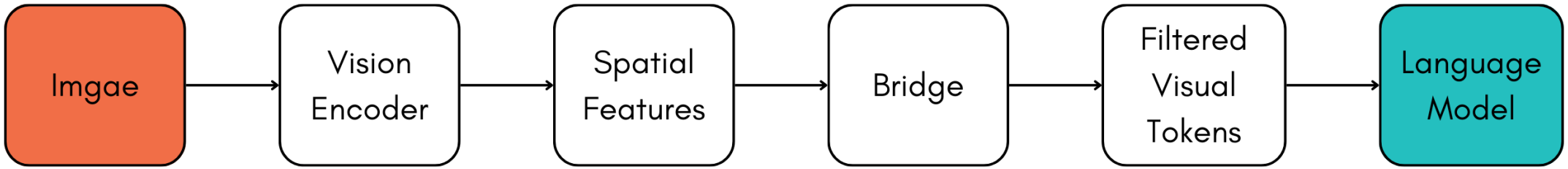


VISION-LANGUAGE MODELS

Embedding-Level Alignment Models | **Bridged Feature Interaction Models** | Large Multimodal Models

Operational Principle

- Frozen Components
- Bridging mechanism
- Curated set of visual tokens available to the LLM



VISION-LANGUAGE MODELS

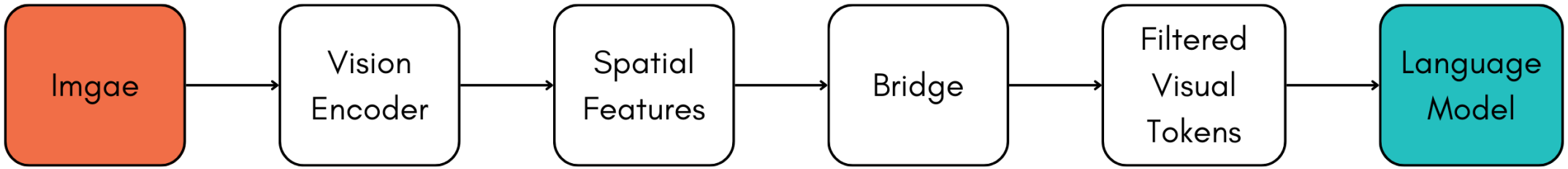
Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models

Semantic consequences

- LLM reasons over filtered representations
- Improved compositional grounding
- Trade-off between semantic richness and computational efficiency

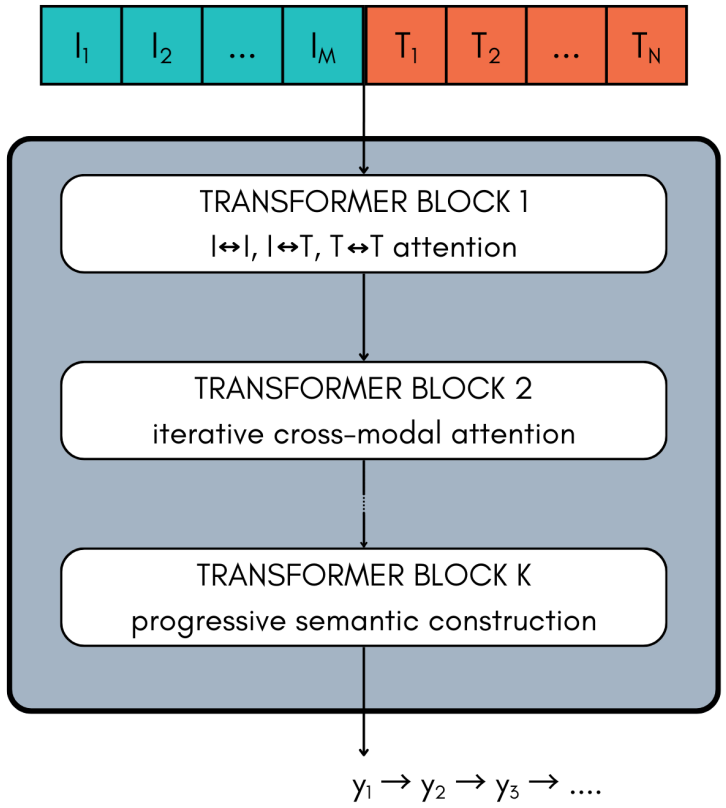


VISION-LANGUAGE MODELS

Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models



Unified Input Token Sequence

Generative Multimodal Semantic Space

Autoregressive Generation
Each token attends to inputs and previous tokens

Operational Principle

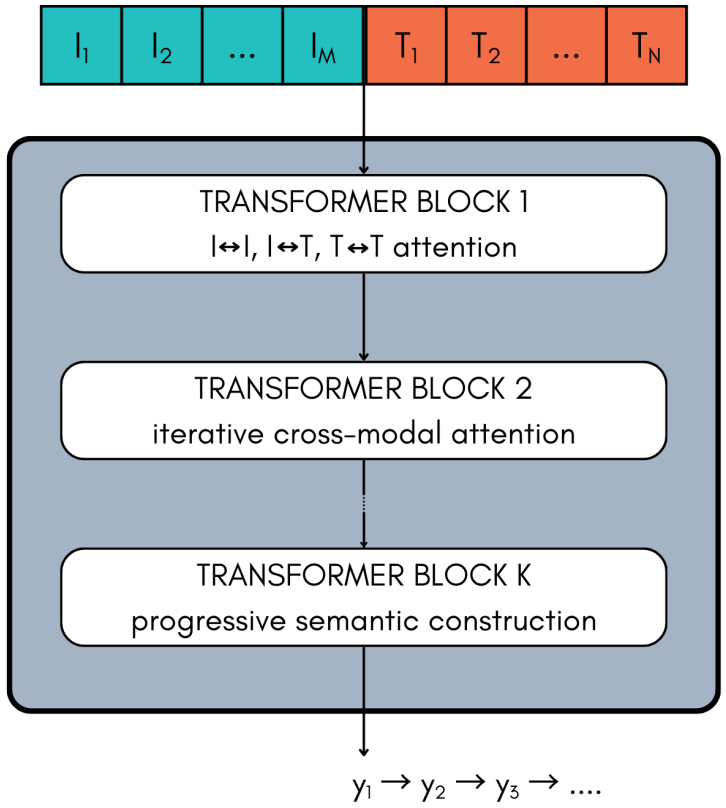
- Shared self-attention across modalities
- Semantics constructed during generation
- No fixed semantic bottleneck

VISION-LANGUAGE MODELS

Embedding-Level Alignment Models

Bridged Feature Interaction Models

Large Multimodal Models



Unified Input Token Sequence

Generative Multimodal Semantic Space

Autoregressive Generation
Each token attends to inputs and previous tokens

Semantic consequences

- Fine-grained relational reasoning
- Multi-stage semantic grounding
- Highest expressiveness, but increased training and computational cost

VISION-LANGUAGE MODELS: COMPARISON

	Embedding-Level Alignment Models	Bridged Feature Interaction Models	Large Multimodal Models
Cross-modal Interaction	• Global embedding	• Curated set of visual tokens	• Unified multimodal sequence
Semantic Granularity	• Global only	• Spatially structured	• Fine-graied, dynamically attended
Compositional Reasoning	• Limited	• Improved	• Strong