Novelty Assessment Report

Paper: Visual symbolic mechanisms: Emergent symbol processing in Vision Language Models
PDF URL: https://openreview.net/pdf?id=3RQ863cRbx
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

To accurately process a visual scene, observers must bind features together to represent individual objects. This capacity is necessary, for instance, to distinguish an image containing a red square and a blue circle from an image containing a blue square and a red circle. Recent work has found that language models solve this 'binding problem' via a set of symbol-like, content-independent indices, but it is unclear whether similar mechanisms are employed by Vision Language Models (VLM). This question is especially relevant, given the persistent failures of VLMs on tasks that require binding. Here, we identify a previously unknown set of emergent symbolic mechanisms that support binding specifically in VLMs, via a content-independent, spatial indexing scheme. Moreover, we find that binding errors, when they occur, can be traced directly to failures in these mechanisms. Taken together, these results shed light on the mechanisms that support symbol-like processing in VLMs, and suggest possible avenues for reducing the number of binding failures exhibited by these models.

Core Task Landscape

This paper addresses: **Visual Feature Binding in Vision Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Visual Encoding and Representation Architectures**
- **Multimodal Alignment and Integration Strategies**
- **Binding Mechanisms and Symbolic Processing**
- **Task-Specific VLM Applications and Capabilities**
- **Few-Shot Learning and Adaptation in VLMs**
- **Evaluation, Robustness, and Model Analysis**
- **Survey and Review Literature**

Complete Taxonomy Tree

- Visual Feature Binding in Vision Language Models Survey Taxonomy
- Visual Encoding and Representation Architectures
  - Visual Encoder Design and Comparison (2 papers)
  - [3] Vinvl: Revisiting visual representations in vision-language models (Pengchuan Zhang, 2021) View paper
  - [5] From clip to dino: Visual encoders shout in multi-modal large language models (Jiang Dongsheng, 2023) View paper
  - Visual Token Processing and Efficiency (3 papers)
  - [10] LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models (Shang, 2024) View paper
  - [20] Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models (Luo, 2024) View paper
  - [29] TokenPacker: Efficient Visual Projector for Multimodal LLM (Wentong Li, 2024) View paper
  - Visual Information Flow and Representation Analysis (2 papers)
  - [4] Visual representations inside the language model (Liu, 2025) View paper
  - [21] Towards interpreting visual information processing in vision-language models (Clement Neo, 2024) View paper
- Multimodal Alignment and Integration Strategies
  - Vision-Language Pre-training Methods (4 papers)
  - [2] Vila: On pre-training for visual language models (Ji Lin, 2024) View paper
  - [33] Bridgetower: Building bridges between encoders in vision-language representation learning (Xu Xiao, 2023) View paper
  - [38] A survey of vision-language pre-trained models (Yifan Du, 2022) View paper
  - [41] Pyramidclip: Hierarchical feature alignment for vision-language model pretraining (Gao Yuting, 2022) View paper
  - Adapter-Based Alignment (2 papers)
  - [34] Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment (Cijo Jose, 2025) View paper
  - [37] CLIP-Adapter: Better Vision-Language Models with Feature Adapters (Peng Gao, 2021) View paper
  - Cross-Modal Mapping and Grounding (3 papers)
  - [1] Generating images with multimodal language models (Koh, 2023) View paper
  - [27] Language is not all you need: Aligning perception with language models (Huang, 2023) View paper
  - [32] Grounding language models to images for multimodal inputs and outputs (Koh, 2023) View paper
  - Hierarchical and Multi-Level Feature Alignment (2 papers)
  - [35] Lion: Empowering multimodal large language model with dual-level visual knowledge (Chen Gongwei, 2024) View paper
  - [43] Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models (LI Xin, 2024) View paper

- Binding Mechanisms and Symbolic Processing
  - Binding Problem Analysis in VLMs ★ (3 papers)
  - [0] Visual symbolic mechanisms: Emergent symbol processing in Vision Language Models (Anon et al., 2026) View paper
  - [36] Investigating Mechanisms for In-Context Vision Language Binding (Darshana Saravanan, 2025) View paper
  - [45] Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem (Rane, 2024) View paper
  - Cross-Domain Binding and Neural Decoding (3 papers)
  - [23] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features (Changde Du, 2023) View paper
  - [24] Feature binding in biological and artificial vision (Pieter R. Roelfsema, 2025) View paper
  - [31] Revealing vision-language integration in the brain with multimodal networks (Subramaniam Vighnesh, 2024) View paper
- Task-Specific VLM Applications and Capabilities
  - Region-Level Understanding and Grounding (4 papers)
  - [6] Contextual object detection with multimodal large language models (Yuhang Zang, 2025) View paper
  - [7] Cogvlm: Visual expert for pretrained language models (Keqin Chen, 2024) View paper
  - [12] Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models (Ma, 2024) View paper
  - [40] Vip-llava: Making large multimodal models understand arbitrary visual prompts (Mu Cai, 2024) View paper
  - Multimodal Generation and Synthesis (2 papers)
  - [30] LMFusion: Adapting Pretrained Language Models for Multimodal Generation (Shi, 2024) View paper
  - [46] Multimodal neural language models (Ryan Kiros, 2014) View paper
  - Video Understanding and Temporal Reasoning (3 papers)
  - [14] Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces (Jihan Yang, 2024) View paper
  - [26] Cogvlm2: Visual language models for image and video understanding (Hong Wenyi, 2024) View paper
  - [42] MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens (Shen, 2024) View paper
  - Visual Reasoning and Chain-of-Thought (2 papers)
  - [11] Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models (Hu, 2024) View paper
  - [17] Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models (Yuhao Dong, 2024) View paper
  - Specialized Visual Understanding Domains (2 papers)
  - [44] Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition (Zhang Pan, 2023) View paper
  - [50] SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension (LI Bohao, 2024) View paper
- Few-Shot Learning and Adaptation in VLMs (3 papers)
  - [16] Multimodal few-shot learning with frozen language models (Maria Tsimpoukelli, 2021) View paper
  - [22] Integrated Image-Text Augmentation for Few-Shot Learning in Vision-Language Models (Ran Wang, 2025) View paper
  - [48] Flamingo: a Visual Language Model for Few-Shot Learning (Alayrac, 2022) View paper
- Evaluation, Robustness, and Model Analysis
  - Benchmark Development and Evaluation Frameworks (3 papers)
  - [15] Vlm2vec: Training vision-language models for massive multimodal embedding tasks (Jiang Ziyan, 2024) View paper
  - [28] NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples (Li, 2024) View paper
  - [39] Evaluating text-to-visual generation with image-to-text generation (Lin, 2024) View paper
  - Robustness and Adversarial Analysis (2 papers)
  - [19] Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models (Li Yifan, 2024) View paper
  - [25] ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models (Hao Yin, 2025) View paper
  - Model Interpretation and Representation Analysis (2 papers)
  - [8] Delving into out-of-distribution detection with vision-language representations (Ming, 2022) View paper
  - [13] ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference (Lan, 2024) View paper
- Survey and Review Literature (4 papers)
  - [9] Vision-language models for vision tasks: A survey (Jingyi Zhang, 2024) View paper
  - [18] The revolution of multimodal large language models: a survey (Davide Caffagni, 2024) View paper
  - [47] Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models (Xiaodan Wang, 2023) View paper
  - [49] Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (Wang Wen-hui, 2022) View paper

Narrative

Core task: visual feature binding in vision language models. The field addresses how VLMs associate visual features with linguistic descriptions, a challenge that spans multiple dimensions. The taxonomy reflects this breadth through branches covering Visual Encoding and Representation Architectures (which explore how raw images are transformed into embeddings, as in Vila Pretraining[2] and Vinvl Visual Representations[3]), Multimodal Alignment and Integration Strategies (focusing on cross-modal fusion techniques like those in BridgeTower[33] and LMFusion[30]), and Binding Mechanisms and Symbolic Processing (examining the core binding problem itself). Additional branches address Task-Specific VLM Applications, Few-Shot Learning and Adaptation (e.g., Multimodal Few-Shot[16]), and Evaluation, Robustness, and Model Analysis (including works like NaturalBench[28] and Out-of-Distribution Detection[8]). Survey and Review Literature provides overarching perspectives, such as Vision Language Survey[9] and Multimodal LLM Revolution[18].

Within Binding Mechanisms and Symbolic Processing, a particularly active line of work investigates the fundamental limits and capabilities of VLMs in correctly associating visual entities with their attributes. Visual Symbolic Mechanisms[0] sits squarely in this cluster, analyzing how models handle symbolic reasoning over visual features. Nearby, In-Context Vision Binding[36] explores whether binding can emerge from in-context learning, while Binding Problem Limits[45] examines inherent constraints in current architectures. These studies contrast with works in adjacent branches that focus on improving representations (e.g., CLIP to DINO[5] or ClearCLIP[13]) or enhancing integration strategies (e.g., Groma[12] or Visual Sketchpad[11]). The original paper's emphasis on symbolic mechanisms places it at the intersection of theoretical analysis and practical diagnosis, complementing empirical studies like Feature Binding Vision[24] and interpretability efforts such as Revealing Vision-Language Integration[31].

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Investigating Mechanisms for In-Context Vision Language Binding

Authors: Darshana Saravanan, Makarand Tapaswi, Vineet Gandhi | **Year/Venue**: 2025 | **URL**: View paper

Abstract

To understand a prompt, Vision-Language models (VLMs) must perceive the image, comprehend the text, and build associations within and across both modalities. For instance, given an â□□image of a red toy carâ□□, the model should associate this image to phrases like â□□carâ□□, â□□red toyâ□□, â□□red objectâ□□, etc. Feng and Steinhardt [4] propose the Binding ID mechanism in LLMs, suggesting that the entity and its corresponding attribute tokens share a Binding ID in the model activations. We invest...

Relationship Analysis

Both papers investigate binding mechanisms in Vision Language Models within the same taxonomy category, examining how VLMs associate visual features with attributes and analyzing binding failures. The original paper identifies emergent symbolic mechanisms using spatial position IDs as content-independent indices for binding, while the candidate paper investigates the Binding ID mechanism from LLMs extended to VLMs, demonstrating that VLMs use binding ID vectors to associate 3D objects in images with their textual references through causal interventions on a synthetic Shapes task. The key difference is that the original paper focuses on spatial indexing schemes and their role in binding failures across multiple VLM architectures, whereas the candidate paper adapts the LLM Binding ID framework to demonstrate factorizability and position independence of binding vectors in VLMs.

---

2. Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem

Authors: Rane, Sunayana, Declan Campbell, Giallanza, Tyler, et al. (24 authors total) | **Year/Venue**: 2024 • Neural Information Processing Systems | **URL**: View paper

Abstract

Recent work has documented striking heterogeneity in the performance of state-of-the-art vision language models (VLMs), including both multimodal language models and text-to-image models. These models are able to describe and generate a diverse array of complex, naturalistic images, yet they exhibit surprising failures on basic multi-object reasoning tasks -- such as counting, localization, and simple forms of visual analogy -- that humans perform with near perfect accuracy. To better understand...

Relationship Analysis

Both papers belong to the 'Binding Problem Analysis in VLMs' category, investigating how vision language models associate visual features with attributes and the mechanisms underlying binding failures. The original paper focuses on identifying emergent symbolic mechanisms (position IDs and three-stage processing architecture) that VLMs use to solve binding problems through mechanistic interpretability techniques including causal mediation analysis and interventions. The candidate paper takes a broader theoretical approach, analyzing VLM binding failures through the lens of cognitive science's binding problem framework to explain the heterogeneous performance patterns across tasks, without delving into specific internal mechanisms or conducting interventional studies.

Contributions Analysis

Overall novelty summary. The paper investigates how Vision Language Models solve the visual feature binding problem through emergent symbolic mechanisms, specifically identifying a three-stage, content-independent spatial indexing scheme. It resides in the 'Binding Problem Analysis in VLMs' leaf, which contains only three papers total, making this a relatively sparse research direction within the broader taxonomy. The sibling papers examine in-context learning approaches to binding and architectural limits, suggesting this leaf focuses on mechanistic understanding rather than architectural improvements or task-specific applications.

The taxonomy reveals that binding research sits within a larger 'Binding Mechanisms and Symbolic Processing' branch, adjacent to 'Cross-Domain Binding and Neural Decoding' which bridges biological and artificial vision. Neighboring branches address visual encoding architectures, multimodal alignment strategies, and task-specific applications like region-level grounding. The paper's focus on internal symbolic mechanisms distinguishes it from alignment-focused work in branches like 'Vision-Language Pre-training Methods' or 'Cross-Modal Mapping and Grounding', and from application-oriented studies in 'Region-Level Understanding and Grounding'. The scope_note clarifies this leaf excludes general alignment methods, concentrating instead on feature-attribute association failures.

Among 26 candidates examined across three contributions, none were found to clearly refute the paper's claims. The first contribution (three-stage symbolic mechanisms) examined 10 candidates with zero refutable matches; the second (position ID validation across VLMs) also examined 10 with zero refutations; the third (linking failures to mechanism breakdowns) examined 6 with zero refutations. This suggests that within the limited search scope, the specific mechanistic analysis of spatial indexing schemes and their failure modes appears relatively unexplored, though the broader binding problem has received attention from sibling papers in the same taxonomy leaf.

Based on the top-26 semantic matches examined, the paper's mechanistic focus on spatial indexing and three-stage symbolic processing appears to occupy a distinct niche within binding research. The sparse population of its taxonomy leaf and absence of refuting candidates suggest novelty in its specific analytical approach, though the limited search scope means potentially relevant work outside these candidates remains unexamined. The contribution's emphasis on diagnosing failure modes through position ID mechanisms differentiates it from architectural or training-focused approaches in neighboring taxonomy branches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Identification of three-stage visual symbolic mechanisms for binding in VLMs

Description: The authors identify a three-stage architecture in Vision Language Models that uses position IDs as content-independent spatial indices for binding object features. The three stages consist of ID retrieval heads, ID selection heads, and feature retrieval heads, which are defined using causal mediation analyses.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Object-based attention requires monocular visual pathways.

URL: View paper

Brief Assessment

Object-based Attention Pathways[56] focuses on monocular visual pathways for object-based attention in biological vision systems, not on computational mechanisms in Vision Language Models or position ID-based binding architectures.

---

2. Object-based visual attention for computer vision

URL: View paper

Brief Assessment

Object-based Visual Attention[57] focuses on attention mechanisms for spatial location hypotheses in machine vision systems, not on the three-stage architecture (ID retrieval, ID selection, feature retrieval) using position IDs as content-independent spatial indices in Vision Language Models.

---

3. Binding, spatial attention and perceptual awareness

URL: View paper

Brief Assessment

Binding Spatial Attention[53] focuses on human visual perception and the role of spatial attention in feature binding, not on computational mechanisms in Vision Language Models or attention head architectures.

---

4. Bindings in working memory: The role of object-based attention

URL: View paper

Brief Assessment

Bindings Working Memory[60] examines object-based attention in human working memory for feature binding, not computational mechanisms in Vision Language Models. The paper focuses on psychological experiments with human participants rather than neural network architectures.

---

5. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism

URL: View paper

Brief Assessment

Online Multi-Object Tracking[51] focuses on multi-object tracking using CNN-based single object trackers with spatial-temporal attention for video sequences, not on visual symbolic mechanisms or binding in Vision Language Models.

---

6. Spatial Attention in Visual Working Memory Strengthens Feature-Location Binding

URL: View paper

Brief Assessment

Spatial Attention Working[52] investigates spatial attention's role in working memory feature binding through behavioral experiments with human participants, not computational mechanisms in Vision Language Models. The candidate focuses on memory-driven attention capture paradigms and does not address attention head architectures or position IDs in neural networks.

---

7. Linguistic and conceptual control of visual spatial attention

URL: View paper

Brief Assessment

Linguistic Conceptual Control[54] focuses on spatial attention mechanisms in human vision, not Vision Language Models. The candidate discusses how attention is directed to targets in human perception, which is fundamentally different from the original paper's analysis of emergent computational mechanisms in VLMs.

---

8. Object perception through visual attention

URL: View paper

Brief Assessment

Object Perception Attention[55] focuses on event-based attentional architecture for object learning through sensorimotor sequences, not on identifying attention head mechanisms for visual binding using position IDs in VLMs.

---

9. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model

URL: View paper

Brief Assessment

FINST Spatial-Index[58] focuses on pre-attentive visual indexing mechanisms in human perception using content-independent spatial indices (FINSTs), not on attention head architectures in Vision Language Models or their three-stage processing mechanisms.

---

10. Anchor objects guide spatial attention during visual search.

URL: View paper

Brief Assessment

Anchor Objects Guide[59] focuses on spatial attention during visual search tasks in human perception, not on mechanistic interpretability of Vision Language Models or their internal binding mechanisms.

Contribution 2: Validation of position IDs across diverse VLMs through multiple analysis methods

Description: The authors validate the identified mechanisms across seven different VLM models using representational similarity analysis and intervention experiments, demonstrating that position IDs are a consistent feature across model families and scales.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Deconstructing Spatial Intelligence in Vision-Language Models

URL: View paper

Brief Assessment

Deconstructing Spatial Intelligence[65] focuses on a taxonomy for spatial intelligence evaluation and mentions depth positional encoding, but does not present validation of position ID mechanisms across VLMs using representational similarity analysis and intervention experiments as described in the original paper.

---

2. The Narrow Gate: Localized Image-Text Communication in Native Multimodal Models

URL: View paper

Brief Assessment

Narrow Gate[63] focuses on information flow patterns between image and text modalities in native vs. non-native multimodal VLMs, not on position ID validation mechanisms or intervention experiments across model families.

---

3. Applying Positional Encoding to Enhance Vision-Language Transformers

URL: View paper

Brief Assessment

Applying Positional Encoding[70] focuses on adding positional encoding schemes (DETR and IRPE) to vision-language transformers for image captioning tasks, not on validating position ID mechanisms across models through representational similarity analysis and intervention experiments.

---

4. OMEGA: Optimized Multimodal Position Encoding Index Derivation with Global Adaptive Scaling for Vision-Language Models

URL: View paper

Brief Assessment

OMEGA[69] focuses on position encoding strategies for multimodal alignment, not on validating position ID mechanisms through representational similarity analysis and intervention experiments across VLM architectures.

---

5. Revisiting Multimodal Positional Encoding in Vision-Language Models

URL: View paper

Brief Assessment

Revisiting Multimodal Positional[61] focuses on multimodal rotary positional embedding (RoPE) design and optimization for vision-language models, not on validating position ID mechanisms through representational similarity analysis and intervention experiments as described in the original paper.

---

6. Mitigating Coordinate Prediction Bias from Positional Encoding Failures

URL: View paper

Brief Assessment

Mitigating Coordinate Prediction[68] focuses on coordinate prediction bias from positional encoding failures in high-resolution inputs, not on validating position ID mechanisms across VLMs through representational similarity and intervention experiments as in the original work.

---

7. Reading Images Like Texts: Sequential Image Understanding in Vision-Language Models

URL: View paper

Brief Assessment

Reading Images Texts[67] focuses on position embeddings in VLMs but does not validate position IDs across multiple models using representational similarity analysis and intervention experiments as described in the original paper. The candidate examines 2D RoPE and absolute position embeddings theoretically and empirically, but does not demonstrate the same systematic validation approach across seven different VLM models with the specific mechanisms (ID retrieval, selection, and feature retrieval heads) identified in the original work.

---

8. Advancing General Multimodal Capability of Vision-language Models with Pyramid-descent Visual Position Encoding

URL: View paper

Brief Assessment

Pyramid-descent Visual Position[62] focuses on a novel position encoding method (PYPE) to improve visual token perception, not on validating position ID mechanisms across models through representational similarity analysis and intervention experiments.

---

9. Positional Preservation Embedding for Multimodal Large Language Models

URL: View paper

Brief Assessment

Positional Preservation Embedding[66] focuses on preserving positional information during visual token compression in MLLMs, not on validating position ID mechanisms across VLMs through representational similarity and intervention experiments as done in the original paper.

---

10. Beyond Semantics: Rediscovering Spatial Awareness in Vision-Language Models

URL: View paper

Brief Assessment

Beyond Semantics Spatial[64] focuses on embedding norm imbalances and RoPE sensitivity in VLMs, not on validating position ID mechanisms through representational similarity analysis and intervention experiments as described in the original paper's contribution.

Contribution 3: Linking binding failures to position ID mechanism failures

Description: The authors demonstrate that persistent binding errors in VLMs can be directly traced to failures in the identified symbolic mechanisms, particularly showing that position IDs are less accurately represented in conditions that typically lead to binding errors, such as high feature entropy scenes.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Accelerating primer design for amplicon sequencing using large language model-powered agents

URL: View paper

Brief Assessment

Primer Design Agents[71] focuses on accelerating primer design for amplicon sequencing using large language models, not on vision language models or binding error mechanisms in visual processing.

---

2. GABInsight: Exploring Gender-Activity Binding Bias in Vision-Language Models

URL: View paper

Brief Assessment

GABInsight[73] focuses on gender-activity binding bias in vision-language models during retrieval tasks, not on the general binding problem mechanisms or position ID failures that the original paper investigates.

---

3. Seeeeg: Semantic-aware eeg-based multi-modal retrieval-augmented generation for high-fidelity visual brain decoding

URL: View paper

Brief Assessment

Seeeeg[72] focuses on EEG-based visual brain decoding using retrieval-augmented generation for reconstructing visual stimuli from brain signals. This is fundamentally different from analyzing binding error mechanisms in vision-language models.

---

4. Temporal dynamics of unimodal and multimodal feature binding

URL: View paper

Brief Assessment

Temporal Dynamics Binding[75] investigates temporal dynamics of feature binding in human perception using behavioral experiments, not computational mechanisms in VLMs. The paper focuses on how binding effects decay over time in human subjects, not on position ID mechanisms or retrieval failures in vision language models.

---

5. Multimodal feature binding in object memory retrieval using event-related potentials: Implications for models of semantic memory.

URL: View paper

Brief Assessment

Multimodal Feature Binding[74] focuses on object memory retrieval using event-related potentials in semantic memory models, not on vision language models' binding mechanisms or position ID failures.

---

6. What is left after an error? Towards a comprehensive account of goal-based binding and retrieval

URL: View paper

Brief Assessment

Goal-Based Binding Retrieval[76] focuses on episodic binding between stimuli and intended correct responses in error processing contexts, not on visual language model binding errors or position ID mechanisms in multi-object visual scenes.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Visual symbolic mechanisms: Emergent symbol processing in Vision Language Models View paper
- [1] Generating images with multimodal language models View paper
- [2] Vila: On pre-training for visual language models View paper
- [3] Vinvl: Revisiting visual representations in vision-language models View paper
- [4] Visual representations inside the language model View paper
- [5] From clip to dino: Visual encoders shout in multi-modal large language models View paper
- [6] Contextual object detection with multimodal large language models View paper
- [7] Cogvlm: Visual expert for pretrained language models View paper
- [8] Delving into out-of-distribution detection with vision-language representations View paper
- [9] Vision-language models for vision tasks: A survey View paper
- [10] LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models View paper
- [11] Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models View paper
- [12] Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models View paper
- [13] ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference View paper
- [14] Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces View paper
- [15] Vlm2vec: Training vision-language models for massive multimodal embedding tasks View paper
- [16] Multimodal few-shot learning with frozen language models View paper
- [17] Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models View paper
- [18] The revolution of multimodal large language models: a survey View paper
- [19] Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models View paper
- [20] Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models View paper
- [21] Towards interpreting visual information processing in vision-language models View paper
- [22] Integrated Image-Text Augmentation for Few-Shot Learning in Vision-Language Models View paper
- [23] Decoding visual neural representations by multimodal learning of brain-visual-linguistic features View paper
- [24] Feature binding in biological and artificial vision View paper
- [25] ClearSight: Visual Signal Enhancement for Object Hallucination Mitigation in Multimodal Large Language Models View paper
- [26] Cogvlm2: Visual language models for image and video understanding View paper
- [27] Language is not all you need: Aligning perception with language models View paper
- [28] NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples View paper
- [29] TokenPacker: Efficient Visual Projector for Multimodal LLM View paper
- [30] LMFusion: Adapting Pretrained Language Models for Multimodal Generation View paper
- [31] Revealing vision-language integration in the brain with multimodal networks View paper
- [32] Grounding language models to images for multimodal inputs and outputs View paper
- [33] Bridgetower: Building bridges between encoders in vision-language representation learning View paper
- [34] Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment View paper
- [35] Lion: Empowering multimodal large language model with dual-level visual knowledge View paper
- [36] Investigating Mechanisms for In-Context Vision Language Binding View paper
- [37] CLIP-Adapter: Better Vision-Language Models with Feature Adapters View paper
- [38] A survey of vision-language pre-trained models View paper
- [39] Evaluating text-to-visual generation with image-to-text generation View paper
- [40] Vip-llava: Making large multimodal models understand arbitrary visual prompts View paper
- [41] Pyramidclip: Hierarchical feature alignment for vision-language model pretraining View paper
- [42] MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens View paper
- [43] Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models View paper

- [44] Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition View paper
- [45] Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem View paper
- [46] Multimodal neural language models View paper
- [47] Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models View paper
- [48] Flamingo: a Visual Language Model for Few-Shot Learning View paper
- [49] Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks View paper
- [50] SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension View paper
- [51] Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism View paper
- [52] Spatial Attention in Visual Working Memory Strengthens Feature-Location Binding View paper
- [53] Binding, spatial attention and perceptual awareness View paper
- [54] Linguistic and conceptual control of visual spatial attention View paper
- [55] Object perception through visual attention View paper
- [56] Object-based attention requires monocular visual pathways. View paper
- [57] Object-based visual attention for computer vision View paper
- [58] The role of location indexes in spatial perception: A sketch of the FINST spatial-index model View paper
- [59] Anchor objects guide spatial attention during visual search. View paper
- [60] Bindings in working memory: The role of object-based attention View paper
- [61] Revisiting Multimodal Positional Encoding in Vision-Language Models View paper
- [62] Advancing General Multimodal Capability of Vision-language Models with Pyramid-descent Visual Position Encoding View paper
- [63] The Narrow Gate: Localized Image-Text Communication in Native Multimodal Models View paper
- [64] Beyond Semantics: Rediscovering Spatial Awareness in Vision-Language Models View paper
- [65] Deconstructing Spatial Intelligence in Vision-Language Models View paper
- [66] Positional Preservation Embedding for Multimodal Large Language Models View paper
- [67] Reading Images Like Texts: Sequential Image Understanding in Vision-Language Models View paper
- [68] Mitigating Coordinate Prediction Bias from Positional Encoding Failures View paper
- [69] OMEGA: Optimized Multimodal Position Encoding Index Derivation with Global Adaptive Scaling for Vision-Language Models View paper
- [70] Applying Positional Encoding to Enhance Vision-Language Transformers View paper
- [71] Accelerating primer design for amplicon sequencing using large language model-powered agents View paper
- [72] Seeeeg: Semantic-aware eeg-based multi-modal retrieval-augmented generation for high-fidelity visual brain decoding View paper
- [73] GABInsight: Exploring Gender-Activity Binding Bias in Vision-Language Models View paper
- [74] Multimodal feature binding in object memory retrieval using event-related potentials: Implications for models of semantic memory. View paper
- [75] Temporal dynamics of unimodal and multimodal feature binding View paper
- [76] What is left after an error? Towards a comprehensive account of goal-based binding and retrieval View paper