MIRROR: Aligning Semantic Relations from Language to Image via Gromov–Wasserstein

Hong-Han Wang*, Yuntao Wang*, and Hu Ding**

University of Science and Technology of China, Hefei, China
hh9999@mail.ustc.edu.cn, wangyuntao@mail.ustc.edu.cn, huding@ustc.edu.cn

Abstract. Multimodal Large Language Models (MLLMs) inherit rich relational priors from their language backbones, yet often fail when asked to apply these relationships in visual contexts. We trace this failure to a structural blind spot: projection-based alignment trains each visual token to carry the right semantics, but never asks whether the relationships between concepts survive the crossing from language to vision. To address this, we propose **MIRROR** (**M**apping **I**nter-concept **R**elations from language to visual **R**epresentation via **O**ptimal-transport-based **R**egularization), a geometric regularization framework that transfers relational priors from language to vision by exploiting the rich relational structure encoded in language representations. Specifically, we derive a surrogate loss from the proposed **Semi-Inverse Gromov–Wasserstein (SI-GW)** problem, an inverse geometric problem that aligns visual representations with language-derived relational priors. We show that this formulation admits a unique closed-form solution that prescribes the ideal visual relational structure implied by language geometry and cross-modal coupling. The structure of the formulation also enables efficient computation, making it applicable to long token sequences. Applying SI-GW inside decoder-only Transformers requires careful design. We introduce targeted strategies at the layer, head, and token levels to ensure stable extraction without additional parameters or inference cost. MIRROR improves relational consistency while preserving performance on general vision-language tasks.

Keywords: Alignment · Multimodal Large Language Model · Vision-Language Model

1 Introduction

Multimodal Large Language Models (MLLMs) have recently achieved substantial progress in visual understanding. The models such as GPT-4V [1], LLaVA [31], and Qwen-VL [6] can perform complex visual question answering, comprehend scene relationships, and engage in multi-turn visual reasoning. These abilities arise largely from their powerful language model backbones, which encode extensive world knowledge and structured relational reasoning.

* Equal contribution.
** Corresponding author.

Yet, when MLLMs must apply such relational knowledge visually, a challenge emerges (Fig. 1). A Large Language Model (LLM) that can textually reason that "sitting requires the center of gravity above the support surface" or "Huskies differ from Malamutes in ear shape and body proportion" may still fail to determine spatial support relations or confuse fine-grained categories in vision. This observation raises a fundamental question: Why do **semantic relationships**, that is, relational priors already encoded in language models such as spatial configuration and logical dependency, fail to transfer to visual relation understanding?

Specifically, mainstream MLLMs establish cross-modal connections through projection adapters [2, 23, 31] optimized under standard language modeling objectives. While this ensures that each visual token carries sufficient semantic information, it enforces alignment at the level of identity rather than relational structure [25, 32], as we formally analyze in Sec. 3.

This observation led us to a key insight: effective cross-modal understanding requires aligning not just what concepts mean, but how they relate to each other, that is, the geometric structure encoding semantic relationships. Recent theoretical works support this perspective. Huh [21] proposed the Platonic Representation Hypothesis, suggesting that representations across modalities converge toward shared statistical structures, while other research [19] shows that large language models spontaneously develop visual reasoning priors from structured text, particularly relational structures between concepts. Together, these findings suggest that text, due to its symbolic and compositional nature, encodes semantic relations in a more structured geometry, making it a valuable "teacher" for improving visual relational reasoning.

Fig. 1: Vision can override correct language priors. A GQA [20] example where Qwen2-VL-7B predicts "leafy" from "green and brown" without the image, but salient bare branches bias it toward an incorrect answer. See Appendix A for more examples.

To model the transfer of relational structure across modalities, we consider Optimal Transport (OT) [44], a principled framework for comparing probability distributions while preserving the geometry of their underlying spaces. An important extension is the Gromov–Wasserstein (GW) distance [33], which aligns two metric measure spaces $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ and $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ by minimizing the distortion between their pairwise distance structures under a joint coupling. For discrete distributions μ and ν supported on n_t text and n_v visual tokens respectively, the admissible coupling set is

$$\Pi(\mu, \nu) = \{C \in \mathbb{R}_+^{n_t \times n_v} \mid C\mathbf{1}_{n_v} = \mu,\ C^\top \mathbf{1}_{n_t} = \nu\},$$

where C is a joint probability matrix whose row and column marginals match μ and ν. The squared GW distance is then defined as

$$\mathrm{GW}^2(\mu,\nu) = \min_{C\in\Pi(\mu,\nu)} \sum_{i,j=1}^{n_t} \sum_{k,\ell=1}^{n_v} \left|d_{\mathcal{X}}(i,j) - d_{\mathcal{Y}}(k,\ell)\right|^2 C_{ik}C_{j\ell}. \tag{1}$$

We denote $D^X = (d_{\mathcal{X}}(i,j))_{n_t\times n_t}$ and $D^Y = (d_{\mathcal{Y}}(k,\ell))_{n_v\times n_v}$ as the pairwise distance matrices on $\mathcal{X}$ and $\mathcal{Y}$.

While GW provides a principled measure of structural discrepancy, the problem we face in MLLMs is fundamentally different from the one GW solves. Standard GW takes two fixed metric spaces and searches for an optimal coupling between them. In our setting, however, the cross-attention mechanism of the MLLM already provides a natural, data-dependent coupling between text and visual tokens at no extra cost. What is missing is not the correspondence, but rather a precise specification of what relational structure the visual space should exhibit given the well-organized language geometry. This reframes cross-modal relational alignment as an **inverse geometric problem**:

Given the text geometry D^t and the coupling C supplied by cross-attention, what is the ideal visual geometry $\widehat{D}^v$ that would make the two spaces relationally consistent under C?

We formalize this question as the **Semi-Inverse Gromov–Wasserstein (SI-GW)** problem, which optimizes over the visual distance matrix D^v within the GW objective while anchoring both D^t and C. We prove that this inverse problem admits a unique closed-form solution, which maps text-space relational structure onto the visual token space through the cross-attention coupling. Moreover, directly optimizing a visual geometry to preserve pairwise relational consistency requires jointly reasoning over all pairs of text tokens and all pairs of visual tokens, incurring a quartic cost that is prohibitive for the long token sequences typical in MLLMs. We show that the structure of the inverse problem can be carefully exploited, so that its solution can be decomposed into a sequence of basic matrix operations, reducing the cost of computing the objective of SI-GW loss from $O(n_t^2 n_v^2)$ to $O(n_t^2 n_v + n_v^2 n_t)$, where n_t and n_v denote the numbers of text and vision tokens, respectively.

Deploying SI-GW within large-scale MLLMs is not straightforward, however, as the fidelity of its output depends critically on the quality of both the anchored text geometry and the coupling between modalities. Naively sourcing these from arbitrary Transformer layers and heads may introduce noise and gradient instability. We therefore design a set of principled implementation strategies, including layer decoupling, selective head aggregation, and token suppression, that ensure each component of SI-GW receives clean, semantically meaningful input from the Transformer architecture.

The resulting framework, **MIRROR**, **M**apping **I**nter-concept **R**elations from language to visual **R**epresentation via **O**ptimal-transport-based **R**egularization, introduces no additional parameters or inference cost. Moreover, the experiments across relational and general vision-language benchmarks show that MIRROR

consistently improves inter-concept reasoning while maintaining promising performance on general vision-language tasks.

2 Related Work

Multimodal Large Language Models and Cross-Modal Alignment. Recent multimodal large language models (MLLMs) [6, 13, 27, 31, 48, 49] integrate large language models [11,22,43] with vision encoders [24,35] through learnable projection modules [9, 23, 31]. While this adapter-based paradigm effectively maps visual tokens into the language embedding space, the alignment is typically driven by next-token prediction alone, which supervises each visual token independently without explicitly preserving inter-concept relational structure [25, 32]. Meanwhile, emerging evidence suggests that representations across modalities converge toward shared geometric structures [21], and that LLMs develop structured visual reasoning priors from relational patterns in text [19]. These observations motivate our approach of aligning relational geometry, rather than token-level semantics alone, between the two modalities.

Optimal transport and Gromov–Wasserstein Distance. OT is a classic topic in machine learning [38]. Cuturi's entropically regularized Sinkhorn distance enables parallel and significantly faster approximations of discrete Wasserstein computation [12], inspiring a series of improved Sinkhorn-type algorithms in subsequent work [3,4,7,28]. An important variant of OT is the Gromov–Wasserstein (GW) distance [33], rooted in the Gromov–Hausdorff framework [18] and compares metric measure spaces by minimizing the distortion of pairwise distance structures under an optimal coupling, without requiring point-to-point correspondence. Efficient solvers include Frank–Wolfe methods [14,15], entropic regularization [37,39,40], and semidefinite relaxations [8]. Recent work has further improved GW along complementary directions, including scalable data-dependent optimization for large-scale GW computation and robust variants for structural noise [10, 41].

3 Methodology: MIRROR via Semi-Inverse Gromov–Wasserstein

As illustrated in Fig. 2, a standard MLLM maps visual tokens into the LLM input space via a projection adapter $f : \mathcal{V} \to \mathbb{R}^d$ and is trained with the autoregressive objective

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{t} \log p(w_t \mid f(v_1), \ldots, f(v_{n_v}),\, w_{<t})\,, \tag{2}$$

where $\{v_i\}_{i=1}^{n_v}$ are visual tokens from the vision encoder and $\{w_t\}$ are text tokens in the target sequence. Optimizing Eq. (2) encourages identity-level alignment: each visual token is trained to be individually predictable from language context. However, as shown in Fig. 1, many MLLM failures on relational reasoning stem not from missing concept identities, but from missing relational structure.

We capture this intuition through a relational consistency requirement: relational geometry should be preserved across modalities. Specifically, pairs of

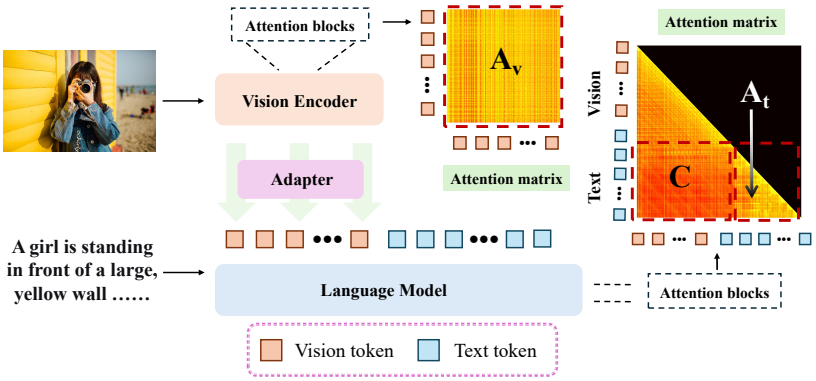


Fig. 2: Overview of the MLLM architecture. An image is encoded into visual tokens via a vision encoder and adapter, then concatenated with text tokens and fed into the language model. A_v denotes visual self-attention from the vision encoder, C the cross-modal attention from text to visual tokens, and A_t the causal text self-attention within the language model.

concepts that are close (or far) in language space should exhibit a consistent relational configuration in visual space. Crucially, Eq. (2) provides no explicit supervision for such relational consistency. Even when individual concepts transfer correctly, their pairwise relations may remain geometrically inconsistent across modalities.

This observation motivates **MIRROR**, a geometric regularization framework that transfers relational priors from language representations to vision representations in MLLMs. We derive the **Semi-Inverse Gromov–Wasserstein (SI-GW)** loss by posing and solving an inverse geometric alignment problem within the MLLM architecture (Sec. 3.1), describe three instantiation strategies that make SI-GW loss stable at scale (Sec. 3.2), and present the overall training procedure (Sec. 3.3).

3.1 Semi-Inverse Gromov–Wasserstein (SI-GW) Loss

Enforcing the relational consistency requirement requires answering three questions: *(i)* how to measure relational geometry within each modality, *(ii)* how to establish correspondences between visual and text tokens, and *(iii)* given the language geometry and the correspondence, what the visual geometry *should* look like. To address those problem, MIRROR augments the standard generation loss with a geometric regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{SI-GW}}, \tag{3}$$

where $\lambda > 0$ controls the strength of structural alignment. The derivation of $\mathcal{L}_{\text{SI-GW}}$ is divided in four steps: *Steps 1* and *2* construct the necessary geometric ingredients from the MLLM's own attention maps. The key contribution comes in *Step 3*, where we pose the *semi-inverse GW problem*, which asks what the ideal visual geometry is, and prove that it admits a closed-form solution. *Step 4* turns this solution into an efficient training loss. Below we elaborate on those steps.

Step 1: Intra-modal geometry. Since relational alignment relies on pairwise structural relations, the primary requirement is an intra-modality distance metric, for which self-attention is a natural proxy: attention in language models is known to encode inter-token relations, from syntactic dependencies to knowledge-graph links [36], so high mutual attention between two tokens can be read as semantic proximity and low attention as distance. Let $A_t \in \mathbb{R}_+^{n_t \times n_t}$ and $A_v \in \mathbb{R}_+^{n_v \times n_v}$ denote self-attention matrices extracted from the text branch and the vision encoder, respectively (Fig. 2). Raw self-attention is generally asymmetric and concentrated near zero, so we symmetrize via $A \leftarrow (A + A^\top)/2$ and apply a negative log-transform to amplify differences:

$$D^t = -\log(A_t + \varepsilon), \qquad D^v = -\log(A_v + \varepsilon), \tag{4}$$

where $\varepsilon \in \mathbb{R}_+$ is a sufficient small amount that prevents numerical instability. Therefore, we obtain text-wise discrepancy $d_t(i,j) = D^t_{i,j}$, and vision-wise discrepancy $d_v(i,j) = D^v_{i,j}$. Under this monotone transformation, semantically related token pair yields small discrepancy, while unrelated pair is mapped to large discrepancy.

These discrepancies quantify geometry within each modality, but enforcing relational consistency requires comparing across modalities, specifically, knowing which visual token corresponds to which text token.

Step 2: Cross-modal correspondence. Comparing two geometries defined on different index sets requires a coupling—a soft assignment that maps tokens in one space to tokens in the other. In our setting, such a correspondence need not be estimated externally: the text-to-vision cross-attention at an intermediate LLM layer already provides a data-dependent soft assignment. Concretely, let $C \in \mathbb{R}^{n_t \times n_v}$ denote the (aggregated) cross-attention matrix, whose entry C_{ij} measures how strongly text token i attends to visual token j.

With the language geometry D^t, the visual geometry D^v, and the coupling C all in hand, the relational consistency requirement can now be stated precisely: *the visual distance structure D^v should be consistent with D^t under the correspondence C*. The question that remains is what "consistent" means quantitatively, so that $\mathcal{L}_{\text{SI-GW}}$ can be derived.

Step 3: The semi-inverse GW problem. We formalize the consistency requirement above as an optimization problem. Treating the language geometry D^t as a fixed teacher and the coupling C as a frozen correspondence, we ask: *what is the visual geometry D^v that best preserves the relational structure of D^t under C?* We call this the **semi-inverse GW problem**:

$$\widehat{D}^v \;=\; \underset{D^v \in \mathbb{R}^{n_v \times n_v}}{\arg\min} \sum_{i,j=1}^{n_t} \sum_{k,\ell=1}^{n_v} \left| d_t(i,j) - d_v(k,\ell) \right|^2 C_{ik}\, C_{j\ell}. \tag{5}$$

The objective is derived from the Gromov–Wasserstein problem (Eq. (1)) in which both D^t and C are held fixed and only D^v is optimized. The name "semi-inverse" is because the roles of the two spaces are asymmetric: the language side specifies the target structure, and the visual side must conform.

Theorem 1 (Closed-form solution of the semi-inverse GW problem). *Let $D^t \in \mathbb{R}^{n_t \times n_t}$ and $C \in \mathbb{R}^{n_t \times n_v}$ with strictly positive column marginals $b = C^\top \mathbf{1}_{n_t} > 0$. The semi-inverse GW problem (Eq. (5)) has the unique minimizer*

$$\widehat{D}^v = \big(C^\top D^t\, C\big) \oslash \big(b\, b^\top\big), \tag{6}$$

where $\oslash$ denotes entrywise division.

Proof (Sketch). Expanding Eq. (5) and collecting terms in $d_v(k,\ell)$ yields a weighted least-squares problem in each entry. The first-order optimality condition gives $\widehat{D}^v_{k,\ell} = \frac{(C^\top D^t\, C)_{k\ell}}{b_k b_\ell}$, which is the unique minimizer since the objective is strictly convex in D^v. The full derivation is provided in Appendix B. □

How to interpret $\widehat{D}^v$? $\widehat{D}^v$ serves as the coupling-induced target vision distance matrix. This target admits an intuitive interpretation as a conditional expectation. For any fixed visual token pair (k,ℓ), we use the coupling weights to induce a probability distribution over text token pairs (i,j):

$$p(i,j \mid k,\ell) \;\triangleq\; \frac{C_{ik}\, C_{j\ell}}{b_k\, b_\ell}, \qquad \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} p(i,j \mid k,\ell) = 1, \tag{7}$$

where the normalization condition follows from $\sum_i C_{ik} = b_k$ and $\sum_j C_{j\ell} = b_\ell$. Under this conditional distribution, $\widehat{D}^v$ admits the following expectation form:

$$\widehat{D}^v_{k,\ell} = \sum_{i=1}^{n_t}\sum_{j=1}^{n_t} d_t(i,j)\, p(i,j \mid k,\ell) = \mathbb{E}_{(i,j)\sim p(\cdot,\cdot\mid k,\ell)}[d_t(i,j)]\,, \tag{8}$$

i.e., the entry $\widehat{D}^v_{k,\ell}$ is the expected text-space distance between the text tokens that attend to visual tokens k and ℓ, respectively. In other words, $\widehat{D}^v$ describes the geometric configuration of visual representation that would emerge if all language relational structure were faithfully inherited through the coupling, which is the precise answer to the question posed at the beginning of this subsection.

Step 4: Training loss. With $\widehat{D}^v$ as a concrete target, we define the SI-GW loss as the squared Frobenius discrepancy between the current visual geometry and the ideal one:

Definition 1 (Semi-Inverse Gromov–Wasserstein loss).

$$\boxed{\mathcal{L}_{SI\text{-}GW} = \Big\| D^v - \widehat{D}^v \Big\|_F^2 = \Big\| D^v - \frac{C^\top D^t\, C}{b\, b^\top} \Big\|_F^2 .} \tag{9}$$

By Theorem 1, minimizing $\mathcal{L}_{\text{SI-GW}}$ over D^v has the same effect as minimizing the quartic semi-inverse objective in Eq. (5)The critical advantage lies in computational efficiency: whereas direct evaluation of Eq. (5) requires $O(n_t^2 n_v^2)$ operations from the sum over all pairs of token pairs, the closed-form reduction to Eq. (9) admits $O(n_t^2 n_v + n_v^2 n_t)$ evaluation. Specifically, in Eq. (9), computing the denominator term $b = C^\top \mathbf{1}_{n_t}$ costs $O(n_t n_v)$; the dominant term $C^\top D^t\, C$ requires $O(n_t^2 n_v + n_v^2 n_t)$ via two matrix multiplications; and the entrywise division and Frobenius norm each cost $O(n_v^2)$.

3.2 Extracting Stable Geometric Ingredients

Computing the SI-GW loss (Eq. (9)) requires extracting D^t, D^v, and C from the multi-layer, multi-head Transformer. Straightforward extraction, e.g., from a single layer with uniform head averaging, yields noisy couplings and unstable gradients in practice (see Sec. 4.4 for empirical evidence). As shown in Fig. 3, we introduce three complementary strategies, each targeting a specific source of instability; the systematic ablations are reported in Sec. 4.4.

Layer Decoupling for Stable Geometry Transfer In decoder-only MLLMs, A_t and C are sub-matrices of the same causal attention map, jointly normalized by a single softmax. Extracting both from the same layer couples their gradients and destabilizes training. We resolve this by *layer decoupling*, extracting each ingredient from a separate stage best suited to its role: A_t is extracted from an early LLM layer, where textual relational geometry is relatively pure before strong multimodal mixing; A_v from the final ViT layer, which most directly governs downstream visual reasoning; and C from an intermediate LLM layer, where representations balance semantic richness and concentration [42, 45].

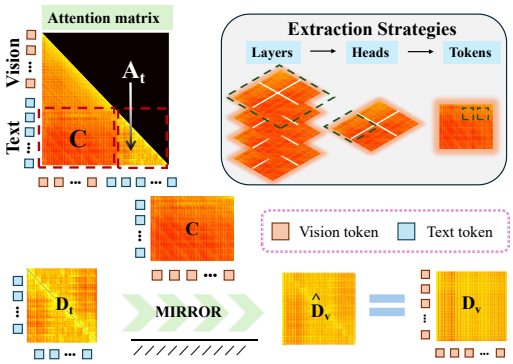


Fig. 3: MIRROR extraction and alignment. A_t and C are extracted from separate LLM layers, and A_v from the final ViT layer, with selection applied at the layer, head, and token levels. MIRROR constructs the target geometry $\widehat{D}^v$ from D^t and C, and drives D^v toward it via the SI-GW loss.

In practice, the full causal attention map at each LLM layer spans all token types. We separate the text–text and text–vision sub-blocks via the padding mask to yield A_t and the raw cross-attention, respectively.

Selective Head Aggregation for Reliable Coupling At the coupling layer, multi-head attention yields H cross-attention maps $\{A^h\}_{h=1}^H$. Prior work has shown that attention heads exhibit highly uneven importance [34, 46]; many produce near-uniform distributions that dilute sharp correspondences when naively averaged. To identify the most informative heads, we compute the entropy $\mathcal{H}_h$ of each head h's cross-attention map A^h:

$$\mathcal{H}_h = -\sum_{i,j} A^h_{ij} \log A^h_{ij}, \tag{10}$$

where lower entropy indicates a more concentrated, and thus more discriminative, cross-modal assignment. We retain the top-k lowest-entropy heads and form

the coupling by averaging over them:

$$C = \frac{1}{k} \sum_{h \in \mathcal{S}_k} A^h, \qquad \mathcal{S}_k = \big\{ h : \mathcal{H}_h \text{ is among the } k \text{ smallest} \big\}. \tag{11}$$

Non-semantic Token Filtering for Cleaner Text Geometry The text stream contains tokens that carry no visual semantics, including system prompts, image placeholders (*e.g.*, `<image>`), and formatting markers. These tokens appear in A_t yet do not participate in meaningful relational structure, introducing noise into the derived distance matrix D^t. To mitigate this, we apply a soft, position-aware suppression. For each text token j, we define a weight $w_j \in [0, 1]$ based on its position and its coupling mass $\sum_k C_{jk}$: tokens that appear early in the sequence (where system prompts typically reside) and attend weakly to visual tokens receive large w_j. The attention entries are then attenuated as:

$$\widetilde{A}_t(i, j) = \big(1 - \alpha \cdot w_j\big)\, A_t(i, j), \tag{12}$$

where $\alpha \in [0, 1]$ controls the overall suppression strength, followed by row-wise re-normalization before computing D^t via Eq. (4). Compared with hard removal, this soft filtering preserves potentially useful context while reducing the influence of non-semantic tokens. Details on the weighting scheme are provided in Appendix C.

3.3 Training Procedure

The complete MIRROR training procedure is summarized in Algorithm 1. At each training step, the standard forward pass through the ViT and LLM produces all required attention matrices (line 2). Layer decoupling, head selection, and token suppression (Sec. 3.2) are then applied to extract clean D^t, D^v, and C (lines 3–6), from which $\widehat{D}^v$ and the SI-GW loss are computed (lines 7–10). The combined objective (Eq. (3)) drives parameter updates (line 11).

No additional inference overhead. MIRROR introduces no additional parameters and inference cost. All SI-GW loss computations use attention matrices already produced during the forward pass and are active only during training. At test time, the model architecture and computational cost remain identical to the baseline MLLM.

4 Experiments

We first describe the implementation details (Sec. 4.1) and evaluation benchmarks (Sec. 4.2), then present main results (Sec. 4.3) followed by ablation and design analysis (Sec. 4.4).

4.1 Implementation Details

Model and Training Configuration. We conduct experiments on two representative MLLM families, LLaVA-1.5 [29] and LLaVA-NeXT [30], each at the 7B and 13B scales. Both families are selected for (1) fully open-sourced training data that enables rigorous reproduction and evaluation, and (2) clean architectural designs that eliminate confounding factors from complex structures. For

Algorithm 1 MIRROR Training

Input: Training data $\mathcal{D}$, MLLM parameters θ, SI-GW weight λ, top-k heads, stability constant ε, learning rate η
Output: Fine-tuned parameters θ^*

for each mini-batch $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ **do**
  Perform forward pass through ViT and LLM
  // Decoupled attention extraction
  Extract A_v (final ViT), A_t (early LLM), $\{A^h\}_{h=1}^{H}$ (intermediate LLM)
  Symmetrize: $A_v \leftarrow (A_v + A_v^\top)/2, \quad A_t \leftarrow (A_t + A_t^\top)/2$
  // Entropy-based head selection and coupling
  Compute per-head entropy: $\mathcal{H}_h \leftarrow -\sum_{i,j} A_{ij}^h \log A_{ij}^h$
  Select $\mathcal{S}_k \leftarrow \{h : \mathcal{H}_h$ is among the k smallest$\}$
  Compute coupling matrix: $C \leftarrow \frac{1}{k} \sum_{h \in \mathcal{S}_k} A^h$
  // Non-semantic token filtering
  Filter non-semantic tokens in A_t to obtain $\tilde{A}_t$ (Eq. (12))
  // Semi-Inverse Gromov–Wasserstein loss
  $D^t \leftarrow -\log(\tilde{A}_t + \varepsilon), \quad D^v \leftarrow -\log(A_v + \varepsilon)$
  $\mathbf{b} \leftarrow C^\top \mathbf{1}_{n_t}, \quad \widehat{D}^v \leftarrow (C^\top D^t C) \oslash (\mathbf{b}\,\mathbf{b}^\top)$
  $\mathcal{L}_{\text{SI-GW}} \leftarrow \|D^v - \widehat{D}^v\|_F^2$
  // Parameter update
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{LM}} + \lambda \cdot \mathcal{L}_{\text{SI-GW}}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}$
end for
return $\theta^* \leftarrow \theta$

each model, we perform MIRROR fine-tuning for 1 epoch on 8 NVIDIA RTX 6000 Ada GPUs (48GB), following the original training protocol of each family. All evaluations use greedy decoding, and baseline results are reproduced from official checkpoints.

MIRROR Hyperparameter Settings. The SI-GW loss weight λ in Eq. (3) is set to 0.002. For layer decoupling (Sec. 3.2), we extract the text self-attention from Layer 1 of the LLM, the visual self-attention from the final vision encoder layer, and the coupling matrix C from Layer 16. For head selection (Eq. (11)), we retain the k=8 lowest-entropy heads. For non-semantic token filtering (Eq. (12)), we set the initial attenuation strength $\alpha = 0.5$.

4.2 Evaluation Benchmarks

We evaluate our model on two categories of tasks:

Relational Reasoning Tasks. GQA [20] is a large-scale scene graph reasoning dataset requiring multi-hop reasoning and spatial relation understanding. It provides fine-grained annotations across five semantic categories: *Attribute*, *Category*, *Global*, *Object*, and *Relation*. BLINK [16] evaluates fine-grained visual perception and multimodal reasoning, spanning three difficulty levels: *low-level* pattern matching (visual correspondence, reflectance), *mid-level* spatial reasoning (spatial relations, multi-view, jigsaw), and *high-level* understanding (localization, counting, forensics, similarity).

Table 1: Structured relational reasoning on GQA. Shaded rows denote MIRROR-enhanced models. **Bold** highlights notable gains ($\Delta \geq 1.0$).

Method	Attr.	Categ.	Global	Obj.	Rel.	Overall
LLaVA-1.5-7B	67.52	51.17	62.66	87.79	53.15	61.16
+ MIRROR	68.08 (+0.56)	**53.61** (+2.44)	**65.82** (+3.16)	88.05 (+0.26)	53.78 (+0.63)	61.93 (+0.77)
LLaVA-1.5-13B	68.85	53.79	63.29	88.30	54.30	62.47
+ MIRROR	69.31 (+0.46)	54.57 (+0.78)	**64.33** (+1.04)	88.56 (+0.26)	54.43 (+0.13)	62.81 (+0.34)
LLaVA-NeXT-7B	71.18	56.73	69.34	89.42	57.38	64.74
+ MIRROR	71.63 (+0.45)	57.49 (+0.76)	**71.47** (+2.13)	89.61 (+0.19)	**58.82** (+1.44)	65.56 (+0.82)
LLaVA-NeXT-13B	71.92	57.51	70.17	89.85	58.14	65.43
+ MIRROR	72.27 (+0.35)	58.14 (+0.63)	**71.68** (+1.51)	90.01 (+0.16)	**59.21** (+1.07)	66.05 (+0.62)

General Vision-Language Tasks. VQAv2 [17] is a widely-used benchmark covering diverse everyday scenarios. POPE [26] evaluates hallucination issues by testing object existence judgment. RealWorldQA [47] contains real-world commonsense reasoning questions.

4.3 Main Results

Our main experiments address three questions:
RQ1. Can MIRROR effectively enhance relational structure understanding in multimodal models?
RQ2. Does it generalize across reasoning levels, from spatial to semantic relations?
RQ3. Can these benefits be achieved without degrading general vision-language performance?

Structured Relational Reasoning (RQ1) We evaluate MIRROR on structured relational reasoning using GQA. As shown in Tab. 1, MIRROR yields steady overall gains across all four model configurations: +0.77% and +0.34% for LLaVA-1.5 (7B/13B), and +0.82% and +0.62% for LLaVA-NeXT (7B/13B), suggesting that its effectiveness generalizes across model scales and architectures.

A cross-category analysis reveals a coherent pattern across all four configurations. The largest gains appear in **Global** tasks (+1.04% to +3.16%), which require integrating relationships among multiple objects (*e.g.*, "how many red objects are on the table"). **Category** tasks involving fine-grained discrimination also benefit (+0.63% to +2.44%). Interestingly, **Relation** tasks show more pronounced improvements on the stronger LLaVA-NeXT baselines (7B: +1.44%, 13B: +1.07%) than on LLaVA-1.5 (7B: +0.63%, 13B: +0.13%). This suggests that a better-aligned feature space allows the model to more effectively leverage its relational reasoning capacity. In contrast, **Object** and **Attribute** tasks exhibit only modest gains, as expected: these categories depend more on local visual feature recognition than on inter-concept relational structures.

This consistent pattern across four model configurations, where Global and Relation tasks improve notably while Object tasks remain near-baseline, supports the hypothesis that MIRROR primarily operates at the level of relational

Table 2: Cross-level reasoning generalization on BLINK. Tasks are grouped by reasoning level: **L**ow-level visual, **M**id-level spatial, and **H**igh-level semantic. Shaded rows denote MIRROR-enhanced models; Δ rows show absolute improvement; "–" indicates no change. **Bold** highlights notable gains ($\Delta \geq 3.0$) and corresponding scores.

Method	Low-level			Mid-level				High-level							Avg.
	Vis.	Refl.	Dep.	Spat.	M-v.	Jig.	Art	Loc.	Cnt.	For.	IQ	Sim.	Sem.	Func.	
LLaVA-1.5-7B	25.6	26.9	51.6	59.4	44.4	52.7	47.5	55.7	44.2	25.0	21.3	47.4	30.9	33.9	40.5
+ MIRROR	**30.8**	26.9	**54.8**	**64.3**	**48.1**	52.7	47.5	**60.7**	**48.3**	25.0	**24.7**	47.4	**36.0**	33.9	42.9
Δ	**+5.2**	–	**+3.2**	**+4.9**	**+3.8**	–	–	**+4.9**	**+4.2**	–	**+3.3**	–	**+5.0**	–	+2.5
LLaVA-1.5-13B	23.8	42.5	54.8	67.1	44.4	52.7	47.5	42.6	44.2	28.0	22.0	47.4	33.8	28.5	41.4
+ MIRROR	25.0	42.5	54.8	69.2	44.4	52.7	47.5	43.4	46.7	**32.6**	**26.7**	47.4	36.7	29.2	42.8
Δ	+1.2	–	–	+2.1	–	–	–	+0.8	+2.5	**+4.6**	**+4.7**	–	+2.9	+0.8	+1.4
LLaVA-NeXT-7B	8.7	40.3	54.8	69.9	44.4	50.0	47.0	31.1	46.7	20.5	17.3	46.7	29.5	18.5	36.9
+ MIRROR	**18.6**	40.3	55.7	69.9	44.4	52.0	48.7	31.1	47.5	**23.5**	**22.0**	48.2	30.2	**23.9**	39.2
Δ	**+9.9**	–	+0.8	–	–	+2.0	+1.7	–	+0.8	**+3.0**	**+4.7**	+1.5	+0.7	**+5.4**	+2.4
LLaVA-NeXT-13B	14.0	39.6	48.4	69.2	46.6	47.3	46.2	54.9	49.2	24.2	2.0	47.4	16.5	6.9	35.8
+ MIRROR	14.0	41.8	48.4	69.2	**50.4**	**53.3**	46.2	**59.8**	49.2	**28.8**	2.7	47.4	**24.5**	**12.3**	38.3
Δ	–	+2.2	–	–	**+3.8**	**+6.0**	–	**+4.9**	–	**+4.6**	+0.7	–	**+7.9**	**+5.4**	+2.5

geometry, addressing the cross-modal misalignment formalized in Sec. 3 rather than uniformly enhancing feature quality. Appendix D visualizes this restructuring, showing how MIRROR drives the visual geometry D^v toward the target $\widehat{D}^v$.

Cross-Level Reasoning Generalization (RQ2) We next evaluate MIRROR on BLINK to assess whether the improvements generalize across reasoning levels. As shown in Tab. 2, all four configurations achieve consistent overall gains (+1.4% to +2.5%), with clear patterns across BLINK's difficulty spectrum.

Mid-level spatial reasoning tasks show the most consistent improvements, particularly spatial relation understanding (LLaVA-1.5-7B: +4.9%) and multi-view reasoning (+3.8%). These tasks require modeling relative positions and spatial configurations between objects, which directly aligns with the relational geometry strengthened by structural alignment. **High-level semantic reasoning** tasks also benefit substantially: localization (+4.9%), counting (+4.2%), semantic correspondence (+5.0% on LLaVA-1.5-7B, +7.9% on LLaVA-NeXT-13B), and functional reasoning (+5.4% on both LLaVA-NeXT variants). Tasks that depend more on holistic perception (*e.g.*, image similarity) remain largely stable. In contrast, **low-level visual tasks** such as reflection and depth estimation yield mixed results. This is expected, as these tasks primarily rely on low-level feature extraction rather than the inter-concept relational structure encoded in language.

Overall, these results indicate that MIRROR generalizes across reasoning levels, with gains concentrated in mid- and high-level tasks that involve spatial and semantic correspondences, while low-level visual tasks remain largely unaffected.

General Vision-Language Performance (RQ3) Finally, we evaluate general VQA tasks to verify that MIRROR preserves overall vision-language un-

Table 3: General vision-language performance on standard benchmarks. MIRROR preserves or improves general VQA while enhancing relational reasoning. Baseline results from [29] and respective publications. Shaded rows denote MIRROR-enhanced models. **Bold** = best within same scale.

Method	LLM	VQAv2	GQA	POPE	RWQA
7B-class models					
InstructBLIP-7B [13]	Vicuna-7B	–[h]	49.2	74.4	–
Qwen-VL-Chat [5]	Qwen-7B	78.2[†]	57.5[†]	–	–
LLaVA-1.5-7B [29]	Vicuna-7B	79.3	61.2	85.7	56.6
+ MIRROR	Vicuna-7B	81.1	61.9	86.3	56.8
LLaVA-NeXT-7B [30]	Mistral-7B	**82.2**	64.8	86.7	**57.8**
+ MIRROR	Mistral-7B	82.1	**65.5**	**87.1**	57.7
13B-class models					
BLIP-2 [23]	Vicuna-13B	65.0	41.0	85.3	–
InstructBLIP-13B [13]	Vicuna-13B	–[h]	49.5	78.9	–
LLaVA-1.5-13B [29]	Vicuna-13B	81.3	62.5	85.5	56.1
+ MIRROR	Vicuna-13B	81.7	62.8	86.0	**56.4**
LLaVA-NeXT-13B [30]	Vicuna-13B	82.8	65.4	86.2	56.2
+ MIRROR	Vicuna-13B	**82.9**	**66.1**	**86.5**	56.1

[†] Eval-set images observed during pre-training. [h] Held-in during instruction tuning; zero-shot score not applicable. RWQA = RealWorldQA.

derstanding. As shown in Tab. 3, MIRROR maintains or improves performance across all four model configurations.

On VQAv2, LLaVA-1.5-7B achieves a +1.8 gain, while LLaVA-NeXT remains stable (7B: −0.1, 13B: +0.1). On POPE, all configurations improve (+0.3 to +0.6), suggesting that better cross-modal distance structures help reduce object hallucination. RealWorldQA remains stable throughout. Notably, LLaVA-NeXT with MIRROR achieves the best results within each scale on both GQA and POPE, suggesting that structural alignment can complement architectural improvements.

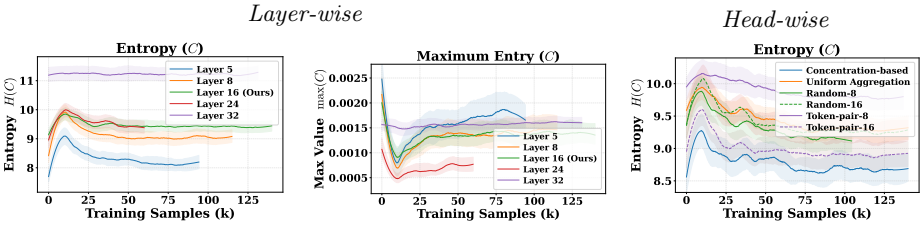
4.4 Ablation and Design Analysis

To understand the efficacy of MIRROR and its underlying mechanisms, we conduct comprehensive ablations and validate our specific design choices.

Component Ablation Tab. 4 examines the contribution of each component via progressive integration and leave-one-out removal. Cumulatively, layer decoupling and head selection together improve the baseline, and token filtering provides a further gain. The leave-one-out results reveal a clear hierarchy of importance. Layer decoupling is indispensable: removing it causes training to diverge, as the SI-GW gradients corrupt the model's attention distributions. Removing head selection also leads to severe degradation, confirming that noisy cross-modal couplings undermine the geometric signal.

Table 4: Component ablation. Top: cumulative addition. Bottom: leave-one-out. †: training diverges.

Layer Decoup.	Head Select.	Token Filter.	GQA	BLINK
			61.2	40.5
Progressive integration				
✓	✓		61.7 (+0.5)	41.9 (+1.4)
✓	✓	✓	**61.9** (+0.2)	**42.9** (+1.0)
Leave-one-out				
✗			diverged†	
	✗		59.8 (-2.1)	35.6 (-7.3)



(a) Entropy **(b)** Max attention **(c)** Entropy

Fig. 4: Cross-attention analysis from two perspectives. **(a,b) Layer-wise:** intermediate layers achieve balanced information richness and peak concentration. **(c) Head-wise (Layer 16):** low-entropy heads yield clearer cross-modal alignment.

Design Choice Validation We empirically justify the specific architectural instantiations of each component.

Coupling layer selection. Fig. 4(a,b) presents the layer-wise cross-attention statistics. Intermediate layers (*e.g.*, Layer 16) achieve an effective balance between information diversity (moderate entropy) and signal concentration (high peak attention values), whereas shallow layers exhibit limited diversity and deep layers tend to over-abstract, losing discriminative power.

Head selection strategy. Fig. 4(c) compares different head aggregation strategies. Selecting the top-k lowest-entropy heads yields the sharpest cross-modal correspondences, while uniform or random aggregation dilutes the concentrated patterns that are critical for reliable structural guidance.

Token filtering effectiveness. The position-aware filtering mechanism effectively down-weights non-semantic tokens, including function words and early-position context tokens (*e.g.*, system prompts), thereby isolating the semantically meaningful tokens required for constructing an accurate relational geometry.

Table 5: Sensitivity to the SI-GW weight λ on GQA.

λ	GQA
0 (no SI-GW)	60.8
10^{-4}	60.8
10^{-3}	61.5
$\mathbf{2{\times}10^{-3}}$	**61.9**
$5{\times}10^{-3}$	61.8
10^{-2}	59.7
10^{-1}	49.3

SI-GW weight λ. Tab. 5 reports the effect of the regularization strength λ. The method is robust across $\lambda \in [10^{-3}, 5\times10^{-3}]$ and performs best at λ=2×10^{-3}; as λ increases further, the SI-GW term progressively overrides the language-modeling objective and accuracy declines.

5 Conclusion

We have identified a critical gap in multimodal alignment: existing methods align individual concepts but fail to preserve relational structures essential for compositional reasoning. To address this, we propose MIRROR, which applies a lightweight variant of Gromov–Wasserstein distance as a geometric regularizer to explicitly align distance structures between visual and textual representations. Through principled hierarchical design strategies, MIRROR induces systematic geometric restructuring of visual space, transforming it to mirror language's relational geometry. Extensive experiments demonstrate that this approach enhances relational reasoning across multiple task levels while fully preserving general vision-language capabilities without additional inference cost. These results establish geometric structure alignment as a promising direction for robust multimodal understanding.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was partially supported by the National Key Research and Development Program of China (No. 2021YFA1000900), and the National Natural Science Foundation of China (No. 62272432, No. 62432016).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Altschuler, J., Bach, F., Rudi, A., Niles-Weed, J.: Massively scalable sinkhorn distances via the nyström method. Advances in neural information processing systems **32** (2019)
4. Altschuler, J., Niles-Weed, J., Rigollet, P.: Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. Advances in neural information processing systems **30** (2017)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

7. Benamou, J.D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing **37**(2), A1111–A1138 (2015)
8. Chen, J., Nguyen, B.T., Koh, S., Soh, Y.S.: Semidefinite relaxations of the gromov-wasserstein distance. Advances in Neural Information Processing Systems **37**, 69814–69839 (2024)
9. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024)
10. Cheng, K., Huang, J., Song, J., Zhang, W., Han, B., Ding, H.: Achieving structurally robust gromov wasserstein distance via adaptive dual-mask. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=PK0zeaGAJ9
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
13. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36** (2024)
14. Ferradans, S., Papadakis, N., Peyré, G., Aujol, J.F.: Regularized discrete optimal transport. SIAM Journal on Imaging Sciences **7**(3), 1853–1882 (2014)
15. Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., et al.: Pot: Python optimal transport. Journal of Machine Learning Research **22**(78), 1–8 (2021)
16. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
17. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
18. Gromov, M., Katz, M., Pansu, P., Semmes, S.: Metric structures for Riemannian and non-Riemannian spaces, vol. 152. Springer (1999)
19. Han, J., Tong, S., Fan, D., Ren, Y., Sinha, K., Torr, P., Kokkinos, F.: Learning to see before seeing: Demystifying llm visual priors from language pre-training (2025), https://arxiv.org/abs/2509.26625
20. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
21. Huh, M., Cheung, B., Wang, T., Isola, P.: Position: the platonic representation hypothesis. In: Proceedings of the 41st International Conference on Machine Learning. pp. 20617–20642 (2024)
22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
23. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

24. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
25. Li, M., Su, N., Qu, F., Zhong, Z., Chen, Z., Li, Y., Tu, Z., Li, X.: Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization (2025), https://arxiv.org/abs/2505.10917
26. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Singapore (2023), https://openreview.net/forum?id=xozJw0kZXF
27. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pretraining for visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26689–26699 (2024)
28. Lin, T., Ho, N., Jordan, M.: On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In: International Conference on Machine Learning. pp. 3982–3991. PMLR (2019)
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
30. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
32. Masry, A., Rodriguez, J.A., Zhang, T., Wang, S., Wang, C., Feizi, A., Suresh, A.K., Puri, A., Jian, X., Noël, P.A., et al.: Alignvlm: Bridging vision and language latent spaces for multimodal understanding. In: Second Workshop on Representational Alignment at ICLR 2025 (2025)
33. Mémoli, F.: Gromov–wasserstein distances and the metric approach to object matching. Foundations of computational mathematics **11**, 417–487 (2011)
34. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? Advances in neural information processing systems **32** (2019)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
36. Ren, J., Guo, Q., Yan, H., Liu, D., Zhang, Q., Qiu, X., Lin, D.: Identifying semantic induction heads to understand in-context learning. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 6916–6932. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.412, https://aclanthology.org/2024.findings-acl.412/
37. Rioux, G., Goldfeld, Z., Kato, K.: Entropic gromov-wasserstein distances: Stability and algorithms. Journal of Machine Learning Research **25**(363), 1–52 (2024)
38. Rüschendorf, L.: The wasserstein distance and approximation theorems. Probability Theory and Related Fields **70**(1), 117–129 (1985)
39. Séjourné, T., Vialard, F.X., Peyré, G.: The unbalanced gromov wasserstein distance: Conic formulation and relaxation. Advances in Neural Information Processing Systems **34**, 8766–8779 (2021)

40. Solomon, J., Peyré, G., Kim, V.G., Sra, S.: Entropic metric alignment for correspondence problems. ACM Trans. Graph. **35**(4), 72:1–72:13 (2016)
41. Song, J., Huang, J., Cheng, K., Han, B., Ding, H.: LoBCD-GW: A fast and data-dependent algorithm for computing gromov-wasserstein distance via localized block coordinate descent. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=dxwRakfEFm
42. Tao, M., Huang, Q., Xu, K., Chen, L., Feng, Y., Zhao, D.: Probing multimodal large language models for global and local semantic representations. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 13050–13056 (2024)
43. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
44. Villani, C., et al.: Optimal transport: old and new, vol. 338. Springer (2009)
45. Viswanathan, K., Gardinazzi, Y., Panerai, G., Cazzaniga, A., Biagetti, M.: The geometry of tokens in internal representations of large language models (2025), https://arxiv.org/abs/2501.10573
46. Voita, E., Talbot, D., Moiseev, F., Titov, I., Sennrich, R.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 5797–5808 (2020)
47. x.ai: Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v (2025)
48. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
49. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2023)