REPRESENTATIONAL SPACE ALIGNMENT FOR VISION-LANGUAGE MODELS

Tian Yun[1,2*] **Tarun Kalluri**[2] **Xuefei Cao**[2] **Dat Huynh**[2] **Jiacheng Zhu**[2]
Bor-Chun Chen[2] **Yifan Wu**[2] **Yuhao Wang**[2] **Ning Zhang**[2] **Miao Liu**[2] **Hengduo Li**[2]
[1]Brown University, [2]Meta

ABSTRACT

While recent vision–language models (VLMs) achieve impressive performance across diverse benchmarks, a substantial modality gap persists due to the distinct inductive biases of their visual and textual backbones in data, architecture, and training objectives. Prior efforts primarily enforce cross-modal alignment by aligning visual and textual representations of the same semantics (e.g., "a green apple" as an image or as a caption) to exhibit high angular similarity. However, such exact alignment can suppress modality-specific information and limit the flexibility or expressiveness of the learned representations.

In this work, we relax such alignment constraint and instead focus on aligning the geometric structure in vision latent space and language latent space. Inspired by vector arithmetic phenomena in word embeddings and linear function vectors in large language models, we propose a straightforward but effective Representational Space Alignment (RSA) loss that encourages the relative geometry of the vision latent space to mirror that of the language latent space. Empirically, we show that (1) unimodal backbones in existing VLMs exhibit weak structural alignment, particularly across layers; Plus, we find that unimodal backbones of VLMs align the best in their last layers. (2) VLMs trained with RSA loss not only reach better cross-modal alignment, but also reach high alignment faster; and (3) VLMs trained with RSA loss achieve consistent gains on fine-grained visual reasoning and perception benchmarks, including MME, MMBench, RealWorldQA, and OK-VQA. Moreover, RSA enhances data efficiency, enabling strong performance under limited training data. These results highlight representational structure alignment as a promising new direction for building more coherent and manipulable vision–language representations.

1 INTRODUCTION

Vision–language models (VLMs) have demonstrated remarkable performance across a wide range of tasks. By jointly training on paired image–text data, these models learn to bridge visual and linguistic semantics within a shared embedding space (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Xu et al.). Despite this progress, a substantial modality gap persists between the visual and textual representations (Liang et al., 2022; Schrodi et al.; Huh et al., 2024). This gap arises because the vision and language backbones are exposed to fundamentally different inductive biases—in data distributions, network architectures, and training objectives—which may have hindered the emergence of strong cross-modal representations (Liang et al., 2022; Schrodi et al.).

Most existing alignment strategies address this issue by enforcing point-wise alignment during the encoder pretraining stage, such as CLIP (Radford et al., 2021). They constrain an image and its corresponding caption (e.g., "a green apple") to map to the vectors with high angular similarity in the joint space (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; 2023; Tschannen et al., 2025; Xu et al.). Such instance-level supervision underlies the success of CLIP (Radford et al., 2021) and its successors (Zhai et al., 2022; 2023; Tschannen et al., 2025; Xu et al.). Yet this approach aligns semantics only at the surface level and overlooks the geometric structure of the latent spaces (Liang et al., 2022; Schrodi et al.). When the training objective collapses different modalities to be aligned

[*]This work was done during an internship of Tian Yun at Meta.

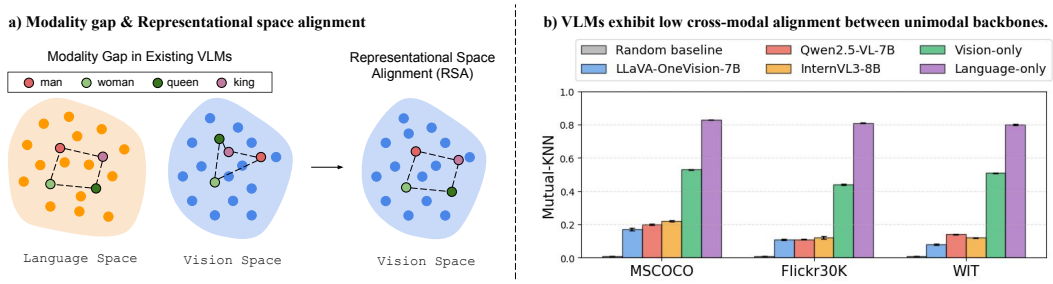

Figure 1: **Motivation and empirical evidence for modality gap in vision-language models (VLMs).** (a) Existing VLMs exhibits a modality gap: the relational geometry of visual and textual representations is misaligned (Liang et al., 2022; Schrodi et al.; Huh et al., 2024). Our proposed RSA loss addresses this by aligning the internal geometries of image and text representations. (b) We quantitatively show that VLMs show poor cross-modal alignment between their unimodal backbones. Vision-only and Language-only illustrate the inherent within-modality structure, while random baselines show expected noise-level alignment.

in direction, it may obscure modality-specific information crucial for downstream reasoning (Jiang et al., 2023; Liang et al., 2022).

In this work, we argue that effective cross-modal understanding requires not only semantic alignment, but also structural alignment between modalities. Rather than enforcing direction similarity, we aim to align the relative geometry of visual and linguistic representation spaces. Our motivation is twofold: First, as shown in Figure 1(a), the structural organization of the language space—shaped by distributional semantics and syntactic regularities—captures a rich manifold of semantics such as royalty (e.g., "king" - "man")(Mikolov et al., 2013; Merullo et al., 2024). If the vision space could mirror this structure, the adapter or projection layer connecting the two modalities could preserve more visual information and enable more faithful grounding. Second, structural alignment opens new possibilities for representation-level manipulation, analogous to the linguistic relation "King - Man + Woman = Queen" in word embedding space (Mikolov et al., 2013) or to the linear function vectors in large language models (LLMs) (Merullo et al., 2024). Such operations would facilitate controllable, language-conditioned image editing and interpretable reasoning in a unified latent space.

To instantiate this idea, we introduce Representational Space Alignment (RSA) loss, a simple yet effective objective that can be incorporated into common VLM (post-)training paradigms such as LLaVA-style VLMs (Liu et al., 2023; 2024a; Li et al., 2024; Li et al.). RSA loss encourages the pairwise relational geometry of vision representations to match that of language representations (Figure 1(a)). Concretely, we minimize the discrepancy between pairwise distance matrices of image and text representations within a batch, thereby transferring the structural regularities of the linguistic manifold into the visual one. The RSA loss can be seamlessly integrated into existing multimodal training pipelines and applied across different architectures and datasets.

Through comprehensive analysis, we first show that existing VLMs exhibit weak structural alignment between unimodal backbones Figure 1(b), particularly in early and mid-level layers, and that alignment peaks near the last layers. We then demonstrate that incorporating RSA loss significantly enhances cross-modal structural coherence: models converge faster to higher alignment scores, achieve consistent improvements on fine-grained reasoning and perception benchmarks such as MME (Fu et al., 2025), MMBench (Liu et al., 2024b), RealWorldQA (xAI, 2024), and OK-VQA (Marino et al., 2019), and generalize better under limited data. Specifically, we make the following contributions:

1. We evaluate the cross-modal structural alignment between vision and language backbones of VLMs, and find that (1) the alignment is weak; (2) the alignment peaks near the last layers of vision and language backbones.

2. We relax the constraint of aligning the representations of the same semantics but different modalities to the same vectors, and introduce Representational Space Alignment (RSA)

loss to enforce the alignment between the geometric structure of vision latent space and language latent space.

3. We demonstrate our RSA loss not only brings alignment between unimodal backbones in VLMs, but also better downstream performance on a suite of downstream tasks. Last, we also show that RSA loss improves data efficiency for VLMs.

2 Related Works

2.1 Vision-Language Models

Vision–language models (VLMs) seek to integrate visual perception and natural language understanding within a unified architecture, progressing from early contrastive dual-encoder methods to increasingly LLM-centric multimodal systems. Initial works (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022; 2023; Tschannen et al., 2025; Xu et al.) relied on large-scale image–text contrastive pretraining to learn robust cross-modal correspondences and support zero-shot recognition. With the success of large language models (LLMs), multimodal LLMs like Flamingo (Alayrac et al., 2022), PaLM-E (Driess et al., 2023), and BLIP-2 (Li et al., 2023b) pair a frozen vision encoder with a pretrained LLM through lightweight projectors, demonstrating that strong multimodal capabilities can be achieved with minimal visual adaptation. Building on this foundation, a new generation of models—including the LLaVA family (Liu et al., 2023; 2024a; Li et al., 2024; Li et al.), MiniGPT-4 (Zhu et al.), Qwen-VL family (Wang et al., 2024; Bai et al., 2025), InternVL family (Chen et al., 2024b; Zhu et al., 2025; Wang et al., 2025), Cambrian (Tong et al., 2024)—employ end-to-end instruction tuning and scaled vision backbones to enhance open-ended multimodal reasoning. In our work, we mainly focus on the new generation of VLMs, since these multimodal LLMs have strong capabilities to support chain-of-though visual reasoning and to solve various downstream tasks.

2.2 Cross-modal Alignment for Vision-Language Models

A key challenge in vision–language modeling is the persistent representational gap between visual and textual encoders. Prior works (Liang et al., 2022; Schrodi et al.) observes *modality gap*: in CLIP (Radford et al., 2021), even though images and texts are encoded into a joint latent space, in practice their embeddings remain separated and do not form a truly unified representation; they attribute this modality gap to model initialization and the temperature hyperparameter in the loss function. Platonic representation hypothesis (Huh et al., 2024) argues that neural networks, regardless of different training objectives or architectures or modalities, are converging to a shared representational space in terms of latent space geometry, and shows that the cross-modal alignment between vision models (Oquab et al.; He et al., 2022; Radford et al., 2021) and language models (Scao et al., 2022; Touvron et al., 2023) increases when the model increases in number of parameters. Prior works have explored ways to mititgate the modality gap via cross-modal mutual information (Li et al., 2025), concept space alignment (Qiu et al., 2026), and patch-level alignment (Jiang et al., 2025). Recent works (Eslami & de Melo, 2025; Gröger et al., 2025) introduce intra-modality separation to reach better alignment. However, majority of these works focus on aligning image and text embeddings at instance-level. Inspired by Huh et al. (2024), we focus on cross-modal alignment in terms of geometric structure of image and text representations. Related in spirit, Park et al. (2019) transfers pairwise relational structure among examples rather than matching features pointwise. Our work similarly emphasizes geometric structure, but differs in that we align vision and language latent spaces within a VLM rather than distilling a teacher into a student.

2.3 Vector Arithmetics and Linear Representation Hypothesis in Large Language Models

A large body of work has shown that modern language models exhibit highly linear representational structure, where semantic relations and behaviors can be expressed through vector arithmetic (Mikolov et al., 2013; Merullo et al., 2024) and linear function directions (Todd et al.; Park et al.) in the latent space. These findings suggest that semantic meaning in LLMs is encoded not in individual embeddings but in the relational geometry of the representational manifold. These findings has direct implications of cross-modal alignment: if LLMs rely on linear, semantically organized geometry, then visual representations must be structurally compatible with that geometry to support

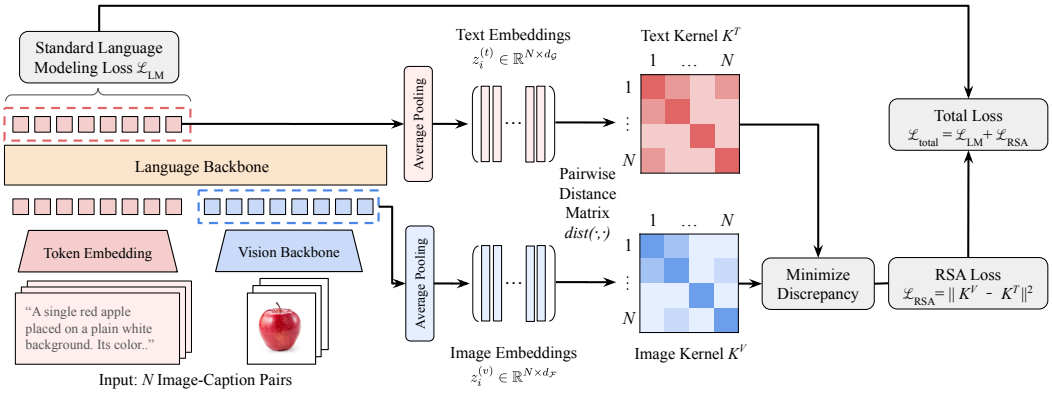

Figure 2: Illustration of the proposed Representational Space Alignment (RSA) auxiliary loss for VLMs. Given N image-caption pairs, the vision backbone encodes image tokens and the language backbone encodes text tokens. Average-pooled representations of each sample are used to construct pairwise distance matrices for both modalities: the image kernel K^V and language kernel K^T. The RSA loss, $\mathcal{L}_{RSA}$ minimizes the discrepancy between these two kernels to encourage the vision and language latent spaces to share similar geometric structures. This objective compliments the standard language modeling loss $\mathcal{L}_{LM}$, promoting cross-modal geometric alignment while preserving modality-specific semantics.

effective cross-modal grounding. Motivated by this connection, our alignment objective aligns the pairwise similarity structure of image and text representations so that visual embeddings inherit the same geometric organization that enables linearity and compositionality in LLMs.

3 METHODS

In this section , we propose Representational Space Alignment (RSA), an auxiliary training objective that aligns the geometric structure of visual and linguistic representations instead of enforcing instance-level vector similarity. RSA encourages the pairwise relations within the vision latent space to mirror those within the language latent space. Figure 2 provides an overview: the RSA loss is to compute the pairwise distances among image embeddings and text embeddings, and minimizes the discrepancy between their relational structures. This preserves modality-specific information while promoting consistent structural organization across modalities.

3.1 KERNEL-BASED REPRESENTATIONAL ALIGNMENT

We mainly focus on vision-language models (VLM) consisting of a vision encoder $\mathcal{F}$, a large language model (LLM) $\mathcal{G}$, and a projection layer $\mathcal{P}$ that maps visual representations into language latent space (Liu et al., 2023; Li et al.; Bai et al., 2025; Wang et al., 2025). Understanding how $\mathcal{F}$ and $\mathcal{G}$ align in their latent structures is central to improving cross-modal representation learning. To this end, we adopt a kernel-based framework for quantifying representational alignment.

Kernels offer a principled framework for analyzing latent representations, as they capture the relative geometric structure among data samples (Kornblith et al., 2019; Klabunde et al., 2025). Such relational structures also constitute the fundamental learning signal in many machine learning algorithms (Aronszajn, 1950; Smola & Schölkopf, 1998; Gupta et al., 2025). Building on this foundation, we follow Huh et al. (2024)'s definition of *representational alignment* – a metric that measures the similarity between the similarity structure of two representations, thus a similarity measure computed over kernels.

For a batch of image-caption pairs $\{I_i, T_i\}_{i=1}^N$, each image will be processed into image tokens and each caption contains multiple language tokens. The encoders then produce **image and caption**

representations by averaging the image and caption token representations:

$$z_i^{(v)} = \mathcal{F}(I_i) \in \mathbb{R}^{d_\mathcal{F}} \tag{1}$$

$$z_i^{(t)} = \mathcal{G}(T_i) \in \mathbb{R}^{d_\mathcal{T}}, \tag{2}$$

where $d_\mathcal{F}$ and $d_\mathcal{T}$ are hidden dimensions of the vision backbone and language backbone.

For each modality, we construct a **kernel**, which reflects the internal geometry of a latent space by measuring the pairwise distance/similarity between data samples:

$$K^V(i,j) = d(z_i^{(v)}, z_j^{(v)}) \tag{3}$$

$$K^T(i,j) = d(z_i^{(t)}, z_j^{(t)}), \tag{4}$$

where $d(\cdot,\cdot)$ denotes a distance metric, such as euclidean distance and cosine distance.

To measure the alignment between the vision and language latent space, we use *mutual nearest-neighbor metric* (mutual-KNN) (Huh et al., 2024), which computes the mean intersection of the k-nearest neighbor sets induced by K^V and K^T, normalized by k. More details of the definition of mutual-KNN can be found in the Appendix.

3.2 Representational Space Alignment

The hypothesis behind **Representational Space Alignment (RSA)** is that effective multimodal alignment emerges when the geometries of vision and language latent space coincide—that is, when the relative relations between images reflect those between their textual counterparts.

RSA loss is to minimize the discrepancy between the two kernels:

$$\mathcal{L}_{RSA} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (K^V(i,j) - K^T(i,j))^2. \tag{5}$$

This loss enforces structural consistency: if two captions are semantically close in language space, their corresponding images are encouraged to be close in vision space, without requiring direct vector equivalence. We further discuss the importance of dense image captions for RSA loss in Section 3.4

For VLMs, RSA can be integrated as an auxiliary objective with the standard language-modeling loss on image-captioning or visual question answering datasets:

$$\mathcal{L}_{total} = \mathcal{L}_{LM} + \lambda_{RSA}\mathcal{L}_{RSA}, \tag{6}$$

where λ_{RSA} controls the strength of representational space alignment. Empirically, we find that using euclidean distance for $d(\cdot,\cdot)$ with $\lambda_{RSA} = 0.01$ works the best.

3.3 Training Strategies

We adopt a multi-stage training paradigm similar to LLaVA-OneVision (Li et al.), which consists of three stages where the training objectives and data diversity gradually increase across stages. Our modification introduces the RSA loss throughout Stage-1 and Stage-1.5 to enhance cross-modal structural consistency:

Stage-1: Language-Image Alignment. The goal of this stage is to align visual features with the token embedding space of the language model. We jointly train both the vision encoder $\mathcal{V}$ and the projection layer $\mathcal{P}$ using a combination of the RSA loss and language modeling loss. This joint optimization explicitly enforces representational space alignment between $\mathcal{V}$ and the language backbone $\mathcal{G}$, allowing the vision backbone to better capture semantic structure from language backbone, while preserving its visual perception capabilities.

Stage-1.5: High-Quality Knowledge Learning. This stage exposes the model to high-quality and knowledge-rich data, including dense image–caption pairs, OCR-based datasets, and pure text corpora. All model components are trainable in this stage, enabling the integration of new multimodal

knowledge while maintaining strong language understanding. Both RSA loss and language modeling loss are optimized, but RSA is applied *only* to image–caption pairs, since only these samples provide semantically aligned visual and textual content suitable for structural alignment.

Stage-2: Visual Instruction Tuning. This stage teaches the model a diverse set of tasks, including image-captioning, mathematical reasoning, various types of visual question answering (VQA) tasks (more details about the data can be found in the Appendix). Since RSA loss is designed specifically for semantically-aligned image–text pairs, many tasks in this stage are not suitable for RSA loss—such as VQA (e.g., the query-answer of "What is the color of the sky? Answer: Blue."), where a query-answer pair only describes small portion of the visual content in the corresponding image.

3.4 Dense Image Captions for RSA Loss

For the image captioning datasets, the quality and granularity of the captions used to compute RSA loss play a critical role in shaping the alignment signal. If captions are overly short or coarse , they fail to reflect the fine-grained semantics of the visual scene, leading to weak or noisy supervision in the relational structure. To obtain a stronger and more informative language geometry, we use dense image captions—detailed textual descriptions that cover multiple visual entities, attributes, and relationships within an image.

4 Modality Gap between Vision Backbone and Language Backbone

Before examining the effectiveness of the auxiliary Representational Space Alignment (RSA) loss, we first examine whether the vision and language backbones in existing vision–language models (VLMs) are already well aligned. Understanding their intrinsic alignment provides insights into where RSA can be most effective.

4.1 Experimental Setup

We evaluate three representative open-sourced VLMs—LLaVA-OneVision (Li et al.), InternVL3 (Zhu et al., 2025), and Qwen2.5-VL (Bai et al., 2025). For each model, we extract the hidden representations from all layers of the vision backbone and the language backbone of each VLM using image–caption samples from MSCOCO (Chen et al., 2015), Flickr30k (Plummer et al., 2015), and Wikipedia caption dataset (WIT) (Srinivasan et al., 2021), which collectively span across everyday scenes, human-centric photos, and encyclopedic content. On each dataset, we randomly sample 1K image-caption pairs to compute the mutual-KNN score. We report the average over three random seeds.

We compute the mutual-KNN score with $k = 10$ to quantitatively measure the representational space alignment between vision and language backbones of a VLM. Mutual-KNN score is ranged from 0 to 1. A higher score indicates stronger alignment between the two modalities.

4.2 Baselines

We mainly consider the following three baselines.

Random baseline is where two sets of 1K hidden representations are randomly sampled from a standard normal distribution, and then compute mutual-KNN score by using these two sets representations. This can be considered as the lower bound.

Vision-only is where the mutual-KNN score is computed by using two sets of representations from the same modality. More specifically, the mutual-KNN reflects the alignment between Qwen2.5-VL's vision backbone (ViT (Dosovitskiy, 2020)) and InternVL3's vision backbone (InternViT (Dosovitskiy, 2020)). Given that (Huh et al., 2024) has shown relatively high alignment across different vision encoders, we thus expect Vision-only baseline to reach high alignment scores as well, since the alignment happens within the vision modality.

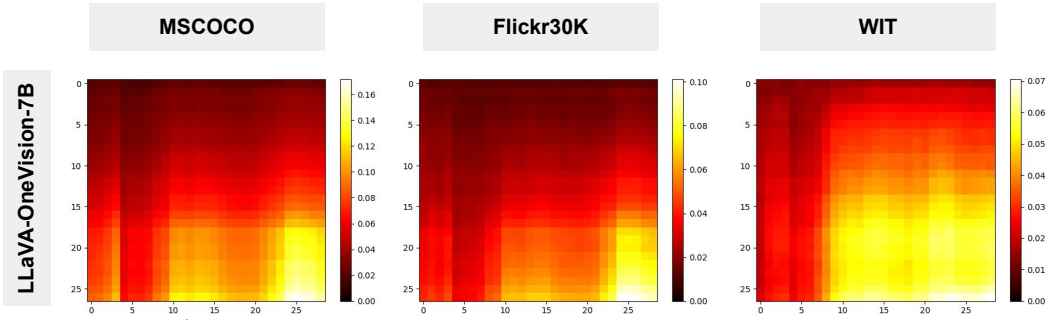

Figure 3: Layer-wise mutual-kNN alignment between vision and language backbones of LLaVA-OneVision-7B. X-axes refer to layers of language backbone. Y-axes refer to layers of vision backbone. Brighter regions indicate higher mutual-KNN scores. Alignment gradually increases with layer depth and peaks near the top layers of both modalities, suggesting that higher layers encode more abstract and semantically aligned features. Evaluation results of all models and all datasets can be found in the Appendix.

Language-only is similar to Vision-only, but the mutual-KNN is computed by representations from Qwen2.5-VL's language backbone (Qwen2-Instruct (Team et al., 2024b)) and InternVL3's language backbone (Qwen2.5-Instruct (Qwen et al., 2025)).

4.3 Observations

Figure 1(b) shows the best mutual-KNN scores across all layer-wise combinations between vision and language backbones. Compared to vision/language baselines, VLMs have much lower mutual-KNN scores, indicating that their latent geometries remain largely unaligned. This observation suggests that current multimodal training strategies fall short of inducing deep representational alignment between unimodal backbones. Interestingly, we also observe that the language baseline consistently achieves substantially higher mutual-KNN scores than the vision baseline across all datasets. This implies that language representations are far more structurally coherent and semantically organized than visual representations.

Figure 3 shows the layer-wise mutual-KNN scores across all layers of vision backbone and language backbone. Among all datasets, the alignment score increases gradually with layer depth and peaks near the final layers of both modalities. We attribute the stronger alignment in the top layers to the fact that both the vision and language models encode higher-level semantic concepts in the final layers.

In all, the observed trend motivates our proposed RSA loss, which explicitly enforces geometric alignment between the vision and language representations. *Given that the alignment reaches the highest at the top layers of vision and language backbone, when computing RSA loss, the image and caption representations are extracted from the last layers of vision and language backbones.*

5 Effectiveness of Representational Space Alignment

In this section, we evaluate the effectiveness of Representational Space Alignment (RSA) as an auxiliary training objective. We summarize our experimental setups, and conduct experiments to demonstrate the effectiveness of RSA loss on improving downstream performance and data efficiency for VLMs.

5.1 Experimental Setups

Model Backbone. We mainly experiment with LLaVA-OneVision-1.5B and LLaVA-OneVision-7B (Li et al.), a recent multimodal model that integrates a SigLIP visual encoder (Zhai et al., 2023) with a Qwen2-Instruct language model (Team et al., 2024b) through a learned MLP projection layer. We refer our model trained with RSA loss as Model-Name-RSA.

Training Data. We employ the official LLaVA-OneVision training mixture, consisting of (1) Stage-1 (Multimodal Pretraining): 558K image–caption pairs with dense captions; (2) Stage-1.5 (High-quality Knowledge Learning): 4M curated multimodal examples where 3.5M samples are image-caption pairs with dense captions, 0.5M OCR samples, and 100K pure language data. (3) Stage-2 (Visual Instruction Tuning): 3.2M multimodal instruction samples from various tasks, such as general knowledge understanding, visual perception, reasoning, etc. We made our best effort to use the same splits and preprocessing steps as the official release to ensure comparability, and only excluded few data splits that are not released.

Baseline Since both the vision backbone and the projection layer will be optimized at Stage 1 and the data is slightly different from the original LLaVA-OneVision, we train a baseline with only the language modeling loss for fair comparisons with our model trained with RSA auxiliary loss.

Implementation Details. All experiments are conducted on a cluster of 28 nodes $\times$ 8 A100 GPUs (224 GPUs in total), each with 80GB of memory.

We follow LLaVA-OneVision training configuration with minor differences on learning rate at Stage-1 since we optimize both the vision backbone and the projection layer.

RSA loss weight is set to $\lambda_{RSA} = 0.01$. For the RSA loss computation, we aggregate image representations by averaging the image representations of the vision encoder before projection into the language space. Given that LLaVA-OneVision adopts AnyRes strategy to increase the resolution of image inputs, each image will be processed into one base image and multiple image patches. Therefore, when aggregating image representations, we only aggregate over the image tokens representing the base image (i.e., the original image) to mitigate the noise and redundancy of this average pooling process. Detailed training configuration for each stage and additional implementation details can be found in Appendix.

5.2 Overall Comparison to Other VLMs

We evaluate on benchmarks covering three categories:

- **Multidisciplinary Reasoning & Knowledge**: MMMU (Yue et al., 2024), MMMU Pro (Yue et al., 2025), MMStar (Chen et al., 2024a), ScienceQA (Lu et al., 2022), MM-Bench (Liu et al., 2024b), and OKVQA (Marino et al., 2019).
- **Chart/Document/Diagram Understanding**: AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), and InfographicVQA (Mathew et al., 2022).
- **General Visual Understanding & Perception**: SeedBench (Li et al., 2023a), RealWorldQA (xAI, 2024), POPE (Li et al., 2023c), and MME (Fu et al., 2025).

Table 1 compares our model trained with RSA loss with other close- and open-sourced VLMs. The results demonstrate that incorporating the proposed RSA loss consistently improves performance across a wide range of benchmarks while maintaining competitive results on all others.

RSA loss improves 1.5B model across the benchmarks. Compared to the baseline, our 1.5B model reaches higher performance on the majority of the reasoning- and knowledge-intensive tasks (e.g., **MMBench(en/cn)** (+2.5/+1.6%), **MMMU Pro** (+1.2%)). RSA further enhances the model's chart/document understanding capability (e.g., **DocQA** (+1.2%), **InfoQA** (+1.7%)), as well as its general visual perception capability (e.g., **MME** (+73.0)). Overall, the 1.5B model remains competitive with, and often surpasses, the baseline, demonstrating that RSA loss provides significant benefits for small VLMs.

RSA loss improves 7B model on fine-grained perception tasks. For the 7B model, RSA yields substantial gains on tasks requiring fine-grained perception, object reasoning, and compositional understanding (e.g., **OK-VQA** (+5.9%), **RealWorldQA** (+1.4%), and **MME** (+30.0)). However, we observe that the baseline is already strong on reasoning-heavy tasks, like MMMU, using larger model, and RSA loss has minimal impact over these tasks. This suggests that the trade-off between

Model	Multidisciplinary Reasoning & Knowledge						
	MMMU	MMMU Pro (standard)	MMStar	ScienceQA	MMBench (en)	MMBench (cn)	OK-VQA
Gemini-1.5-Pro (Team et al., 2024a)	62.2	-	-	-	-	-	-
Claude 3.5 Sonnet (Anthropic, 2024)	68.3	-	-	-	-	-	-
GPT-4V (OpenAI, 2023)	56.8	-	57.1	75.7	75.0	-	-
GPT-4o (OpenAI, 2024)	69.1	-	-	-	-	-	-
Qwen2-VL-7B (Wang et al., 2024)	41.1	-	48.0	-	83.0	80.5	-
Qwen2.5-VL-7B (Bai et al., 2025)	51.2	-	68.2	-	83.5	83.4	-
InternVL-2-8B (Chen et al., 2024c)	49.3	-	59.4	97.0	81.7	-	-
InternVL-3-8B (Zhu et al., 2025)	62.7	-	68.2	-	83.4	82.2	-
Cambrian-34B (Tong et al., 2024)	49.7	-	-	85.6	80.4	79.2	-
LLaVA-OneVision-1.5B	36.2	17.0	44.2	82.8	63.1	56.6	38.7
LLaVA-OneVision-1.5B-RSA	36.4	18.2	44.0	83.4	65.6	58.2	38.7
Δ	+0.2	+1.2	-0.2	+0.6	+2.5	+1.6	0.0
LLaVA-OneVision-7B	45.4	25.1	57.7	93.8	80.2	76.5	46.3
LLaVA-OneVision-7B-RSA	43.7	24.4	56.6	94.1	81.3	77.9	52.2
Δ	-1.7	-0.7	-1.1	+0.3	+1.1	+1.4	+5.9

Model	Chart/Document/Diagram Understanding				General Visual Understanding & Perception			
	AI2D	ChartQA	DocQA	InfoQA	SeedBench (image)	RealWorldQA	POPE	MME (sum)
Gemini-1.5-Pro (Team et al., 2024a)	94.4	87.2	93.1	81.0	-	70.4	-	-
Claude 3.5 Sonnet (Anthropic, 2024)	94.7	90.8	95.2	49.7	-	59.9	-	-
GPT-4V (OpenAI, 2023)	78.2	78.5	88.4	-	49.9	61.4	-	1926
GPT-4o (OpenAI, 2024)	94.2	85.7	92.8	-	76.2	58.6	-	-
Qwen2-VL-7B (Wang et al., 2024)	83.0	83.0	94.5	76.5	-	70.1	88.1	2327
Qwen2.5-VL-7B (Bai et al., 2025)	83.9	87.3	95.7	82.6	-	68.5	-	2347
InternVL-2-8B (Chen et al., 2024c)	83.8	83.3	91.6	74.8	76.0	64.4	-	2210
InternVL-3-8B (Zhu et al., 2025)	85.2	86.6	92.7	76.8	-	70.8	-	2415
Cambrian-34B (Tong et al., 2024)	79.7	75.6	75.5	46.0	-	67.8	-	-
LLaVA-OneVision-1.5B	67.9	66.2	69.4	39.7	69.4	57.0	88.7	1563
LLaVA-OneVision-1.5B-RSA	66.4	66.0	70.6	41.4	69.1	57.3	88.2	1636
Δ	-1.5	-0.2	+1.2	+1.7	-0.3	+0.3	-0.5	+73.0
LLaVA-OneVision-7B	81.1	80.4	85.9	63.6	76.1	66.7	88.7	2042
LLaVA-OneVision-7B-RSA	80.3	80.2	85.4	62.4	76.0	68.1	88.6	2072
Δ	-0.8	-0.2	-0.5	-1.2	-0.1	+1.4	-0.1	+30

Table 1: LLaVA-OneVision-RSA performance across different benchmarks. Δ shows the difference between LLaVA-OneVision-RSA (ours) and LLaVA-OneVision (baseline). RSA objective brings performance improvement in reasoning & knowledge-related and perception-related tasks, while introducing minimal trade-offs on chart/document understanding tasks.

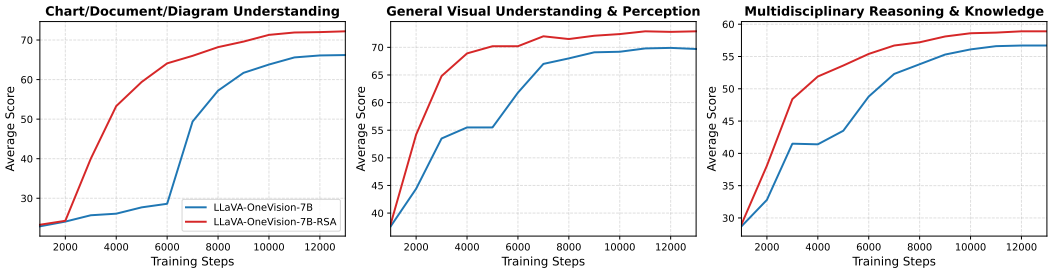

Figure 4: RSA improves data efficiency for VLMs. We remove the intermediate Stage 1.5, thus removing substantial amount of training data (4M) and also removing compute-intensive training of language backbone. Across the three evaluation categories, the model trained with RSA loss learns significantly faster and reaches stronger performance.

cross-modal alignment and downstream performance is minimal, and that RSA primarily enhances capabilities / tasks where detailed visual understanding is most critical.

Overall, these results demonstrate the effectiveness of RSA as a simple yet powerful auxiliary loss for the downstream performance.

Model	Img. Agg.	Stop Grad.	ChartQA	InfoQA	MMMU Pro (standard)	ScienceQA	MMBench (en)	SeedBench (image)	MME (sum)	Average
LLaVA-OneVision-0.5B	-	-	66.2	39.7	17.0	82.8	63.1	69.4	1563	39.4
LLaVA-OneVision-0.5B-RSA	Full	yes	66.5	41.0	17.6	83.5	64.9	69.2	1556	39.8
	Base	yes	66.5	40.9	17.6	82.6	61.0	69.0	1592	39.4
	Base	no	66.0	41.4	18.2	83.4	65.6	69.1	1636	40.2

Table 2: Ablation on (1) image representation aggregation approach and (2) whether RSA loss should update language backbone. We compare averaging over all image tokens (Full) versus only base-image tokens (Base). Using base-image aggregation with no stop-gradient on language backbone yields strong performance across various benchmarks.

5.3 RSA Loss Improves Data Efficiency

Data efficiency is increasingly important as modern VLM training pipelines rely on millions of image–text pairs, making training costly in both compute and time. Improving data efficiency can therefore lead to faster training, reduced compute requirements, and better convergence/performance under limited supervision. Motivated by this, we investigate whether RSA loss can improve data efficiency by removing Stage-1.5 entirely (4M data) and training only with Stage-1 (558K data) and Stage-2. We also train a corresponding baseline model under the same setting for comparison.

Figure 4 shows the training dynamics of models' downstream performance (average performance over all tasks under each category) under the reduced-data setting. Our model trained with RSA loss not only outperforms the baseline in majority of the tasks but also reaches strong performance substantially earlier in training. These results indicate that representational space alignment can meaningfully enahnce the data efficiency of VLMs.

5.4 Ablation on Image Aggregation and More

Table 2 shows an ablation on (1) how image representations are aggregated (i.e., average-pooling over all image tokens versus only the base-image tokens), and (2) whether gradients from RSA loss should update the language backbone. We observe that combining the use of base-image features and optimize the language backbone with RSA loss brings performance improvement on downstream tasks.

In the Appendix, we conduct more experiments about (1) RSA loss improves representational space alignment between the unimodal backbones in VLMs; (2) Ablation on distance metric (euclidean distance versus cosine distance); (3) Ablation on the strength λ_{RSA} of RSA loss; (4) Qualitative examples of VLMs trained with RSA loss.

6 Conclusion

In this work, we introduced Representational Space Alignment (RSA), an auxiliary training objective that aligns the geometric structure of vision and language latent spaces rather than enforcing instance-level similarity. We first demonstrated that, despite the impressive performance of existing VLMs, their unimodal backbones remain weakly aligned across layers, with relatively stronger alignment emerging only at their final layers. By integrating RSA loss into the training pipeline of LLaVA-OneVision, we showed that representational space alignment can be learned efficiently and leads to consistent improvements in a diverse set of benchmarks. Moreover, RSA loss enhances data efficiency. When resource-intensive training stages are removed or a huge amount of training data is unavailable, models trained with RSA not only achieves better results but also reaches strong performance substantially earlier in training.

Overall, our findings highlight representational structure alignment as a promising direction for building performant and data-efficient VLMs. Future work will explore applying RSA objective across additional modalities (e.g., audio, video) and investigating its potential to be adapted to wider variety of tasks.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anthropic. Introducing claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, June 2024. Accessed: 2025-11-14.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pp. 8469–8488. PMLR, 2023.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Fabian Gröger, Shuo Wen, Huyen Le, and Maria Brbić. With limited data for multimodal alignment, let the structure guide you. *arXiv preprint arXiv:2506.16895*, 2025.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. *arXiv preprint arXiv:2510.08492*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *arXiv preprint arXiv:2505.17316*, 2025.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multimodal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023b.

Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Yuan Li, Zhaopeng Tu, and Xiaolong Li. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023c.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple word2vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5030–5047, 2024.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

OpenAI. Gpt-4v(ision) system card. `https://openai.com/index/gpt-4v-system-card/`, September 2023. Accessed: 2025-11-14.

OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, May 2024. Accessed: 2025-11-14.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3967–3976, 2019.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Yifu Qiu, Paul-Ambroise Duquenne, and Holger Schwenk. Unified vision-language modeling via concept space alignment. *arXiv preprint arXiv:2603.01096*, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Alexander J Smola and Bernhard Schölkopf. *Learning with kernels*, volume 4. GMD-Forschungszentrum Informationstechnik Berlin, Germany, 1998.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024b.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, April 2024. Accessed: YYYY-MM-DD.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *The Twelfth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

In the Appendix, we provide the following:

- Formal definition of mutual-KNN alignment metric in Appendix A
- Details of training configurations in Appendix B
- Full layer-wise alignment evaluation in Appendix C
- Full results on the impact of RSA loss on data efficiency in Appendix D
- Results on RSA loss improves cross-modal alignment in Appendix E
- Ablation on distance metrics in Appendix F
- Ablation on strength of RSA loss in Appendix G
- Impacts of multimodal traininig on cross-modal alignment in Appendix H
- Qualitative examples of VLMs trained with RSA loss in Appendix I
- Additional discussion on limitations in Appendix J

A MUTUAL K-NEAREST-NEIGHBOR ALIGNMENT METRIC

Prior works have proposed alignment metrics to measure the alignment between two latent spaces, such as CKA (Kornblith et al., 2019), Unbiased CKA (Song et al., 2012), and SVCCA (Raghu et al., 2017). Following Huh et al. (2024), we use mutual K-Nearest-Neighbor alignment metric (mutual-KNN) to measure the alignment between the vision backbone and language backbone of a vision-language model (VLM).

We focus on vision-language models (VLM) consisting of a vision encoder $\mathcal{F}$, a large language model (LLM) $\mathcal{G}$, and a projection layer $\mathcal{P}$. For a batch of image-caption pairs $\{I_i, T_i\}_{i=1}^N$, each image will be processed into image tokens and each caption contains multiple language tokens. The encoders then produce **image and caption representations** by averaging the image and caption token representations:

$$z_i^{(v)} = \mathcal{F}(I_i) \in \mathbb{R}^{d_\mathcal{F}} \tag{7}$$

$$z_i^{(t)} = \mathcal{F}(T_i) \in \mathbb{R}^{d_\mathcal{T}}, \tag{8}$$

where $d_\mathcal{F}$ and $d_\mathcal{T}$ are hidden dimensions of the vision backbone and language backbone. Thus, for this batch of image-caption pairs, we obtain $\boldsymbol{Z}^{(v)} = \{z_i^{(v)}\}_{i=1}^N$ and $\boldsymbol{Z}^{(t)} = \{z_i^{(t)}\}_{i=1}^N$. For each $(z_i^{(v)}, z_i^{(t)})$ pair, we compute the respective nearest neighbor sets $\boldsymbol{S}(z_i^{(v)})$ and $\boldsymbol{S}(z_i^{(t)})$. We compute mutual-KNN by measuring the average intersection between these two sets via

$$\text{mutual-KNN} = \frac{1}{k}|\boldsymbol{S}(z_i^{(v)}) \cap \boldsymbol{S}(z_i^{(t)})|, \tag{9}$$

where $|\cdot|$ is the size of the intersection.

B TRAINING CONFIGURATIONS

To facilitate reproducibility of our work, Table B.1 summarizes the training configurations and setups used in our training pipeline. We incorporate our proposed Representational Space Alignment (RSA) loss as an auxiliary loss into Stage-1 and Stage-1.5.

C FULL LAYER-WISE ALIGNMENT EVALUATION

Figure C.1 shows the layer-wise mutual-KNN scores across all layers of vision backbone and language backbone of LLaVA-OneVision (Li et al.), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025) on MSCOCO (Chen et al., 2015), Flickr30k (Plummer et al., 2015), and Wikipedia caption dataset (WIT) (Srinivasan et al., 2021).

		Stage-1 Language-Image Alignment	Stage-1.5 High-Quality Knowledge Learning	Stage-2 Visual Instruction Tuning
Vision	**Resolution**	384	384×{2×2, 1×{2,3}, {2,3}×1}	384×{{1×1}, ⋯, {6×6}}
	#Tokens	729	Max 729×5	Max 729×10
Data	**Dataset**	LLaVA-ReCap-LCS	High-Quality Knowledge Data	LLaVA-OneVision-Single-Image
	#Samples	558K	4M	3.2M
Model	**Trainable**	Vision Backbone $\mathcal{V}$ & Projector $\mathcal{P}$	Full Model	Full Model
	1.5B LLM	635.4M	2.2B	2.2B
	7.6B LLM	960.0M	8.0B	8.0B
Training	**RSA Loss**	✔	✔	✘
	λ_{RSA}	0.01	0.01	-
	Batch Size	256	256	256
	LR: Vision $\mathcal{V}$	2×10^{-6}	2×10^{-6}	2×10^{-6}
	LR: {Projector $\mathcal{P}$, LLM $\mathcal{V}$}	2×10^{-6}	1×10^{-5}	1×10^{-5}
	Epoch	1	1	1

Table B.1: Detailed configuration for each training stage of the LLaVA-OneVision model with Representational Space Alignment (RSA) loss. The table outlines the progression of vision parameters, dataset characteristics, model specifications, and training hyperparameters across different stages. We apply RSA loss as an auxiliary loss at Stage-1 and Stage-1.5.

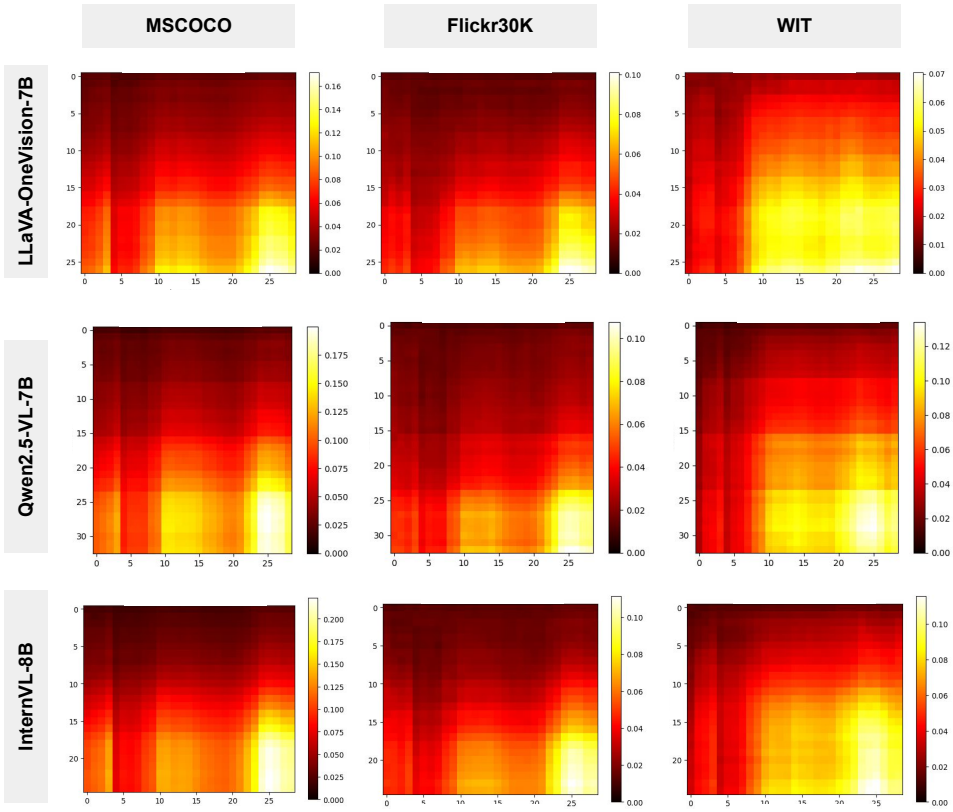

Figure C.1: Layer-wise mutual-kNN alignment between vision and language backbones of LLaVA-OneVision (Li et al.), Qwen2.5-VL (Bai et al., 2025), and InternVL3 (Zhu et al., 2025). X-axes refer to layers of language backbone. Y-axes refer to layers of vision backbone. Brighter regions indicate higher mutual-KNN scores. Alignment gradually increases with layer depth and peaks near the top layers of both modalities, suggesting that higher layers encode more abstract and semantically aligned features.

Among all datasets, the alignment score increases gradually with layer depth and peaks near the final layers of both modalities. We attribute the stronger alignment in the top layers to the fact that both the vision and language models encode higher-level semantic concepts in the final layers. Interestingly, we also observe that for the same dataset, different VLMs exhibit highly similar layer-wise alignment patterns. This consistency likely arises because their vision backbone and language backbone share comparable pretraining objectives.

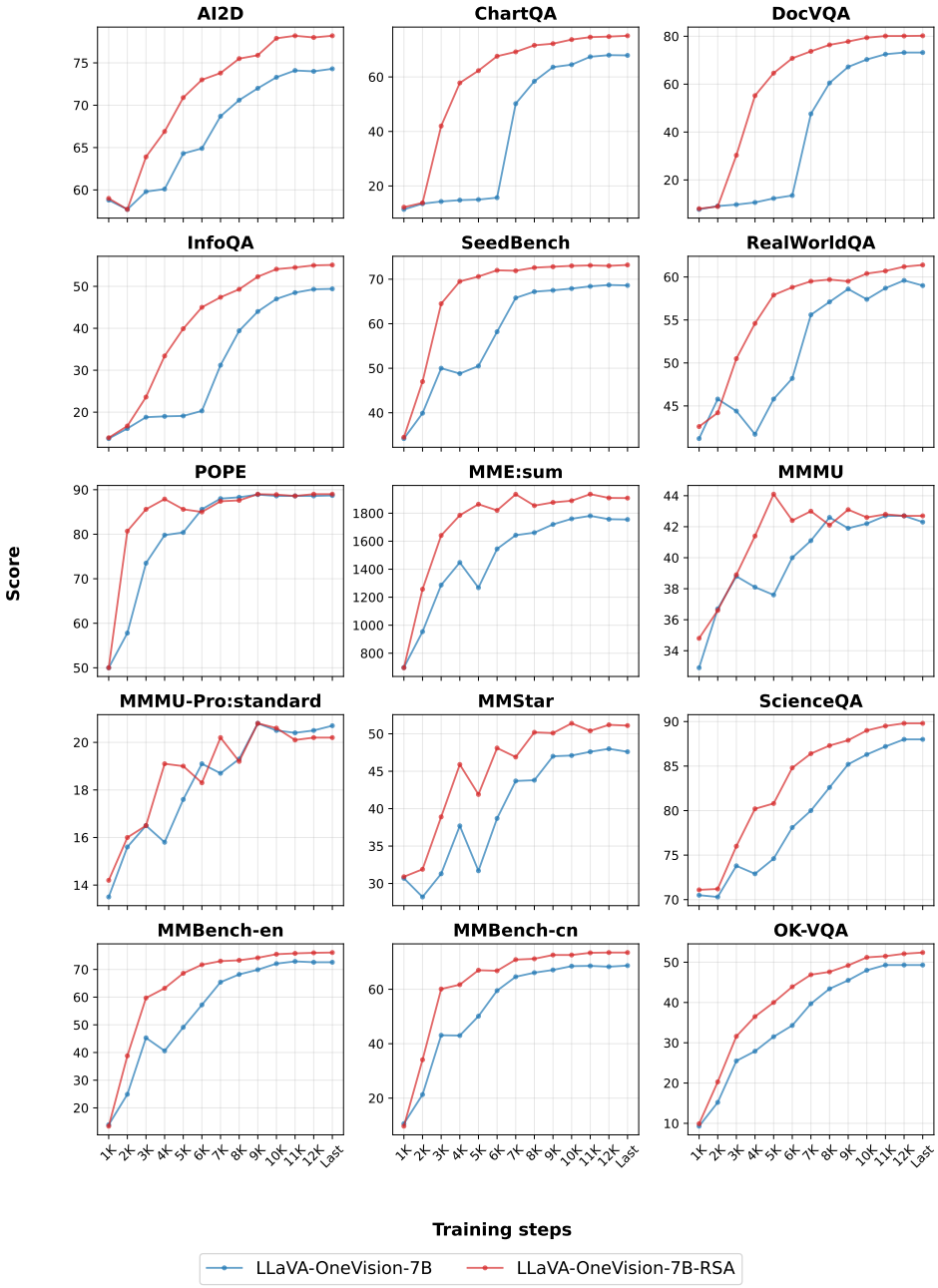

Figure D.1: RSA improves data efficiency for VLMs. We remove the intermediate Stage 1.5, thus removing substantial amount of training data (4M) and also removing compute-intensive training of language backbone. Across a diverse set of downstream tasks, the model trained with RSA loss learns significantly faster and reaches stronger performance.

D Full Results on the Impact of RSA Loss on Data Efficiency

Figure D.1 shows the per-task performance of LLaVA-Onevision-7B (with or without RSA loss) over the Stage-2 training. For almost all the tasks, RSA-trained model not only outperforms the baseline, but also reaches consistently faster convergence, particularly in the early learning phase, indicating that a well-aligned representational structure between the unimodal backbones of a VLM provides an effective supervision signal.

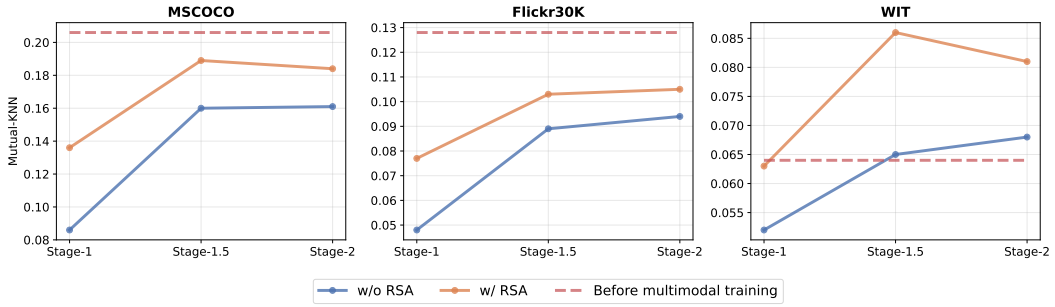

Figure E.1: RSA loss improves cross-modal alignment for VLMs across multimodal training stages. We report mutual-KNN scores between the vision and language backbones after each stage of multimodal training (Stage-1, Stage-1.5, and Stage-2), comparing models trained with and without RSA loss. The red dashed line indicates the alignment before any multimodal training. Compared to model trained without RSA loss, our model trained with RSA loss consistently yields higher alignment across all datasets and training stages.

E RSA Loss Improves Cross-Modal Alignment for VLMs

To understand how RSA loss can impact alignment between the unimodal backbones of a VLM throughout multimodal training, we compute mutual-KNN after each training stage of LLaVA-OneVision-7B with or without RSA loss.

Figure E.1 shows the mutual-KNN scores of LLaVA-OneVision-7B after Stage-1/-1.5/-2 of multimodal training pipeline. Across MSCOCO, Flickr30K, and WIT, our model trained with RSA loss consistently achieve higher cross-modal alignment than the model without RSA at every stage. Though RSA loss is only applied at Stage-1 and -1.5, the cross-modal alignment keeps stable after Stage-2, suggesting that the alignment established in early training stages can be preserved through subsequent instruction tuning. Interesting, we observe a substantial drop on the alignment before any multimodal training and after Stage-1. This suggests that multimodal training rapidly reshape the vision latent space for multimodal tasks such that the cross-modal alignment is compromised in early training stages and improved in later stages.

F Ablation on Distance Metric

We conduct an ablation study on the choice of distance metric for RSA loss computation. We explore Euclidean distance and cosine distance, where Euclidean distance is considered suboptimal due to the high-dimensionality of the vision and language latent space, and cosine distance is a better choice. To save compute resources, instead of following the full training pipeline for this ablation, we conduct further finetuning on the off-the-shelf LLaVA-OneVision-7B with Stage-1 data (Section 5.1 with $\lambda_{RSA} = 0.01$ for Euclidean-based RSA and $\lambda_{RSA} = 1.0$ for cosine-based RSA. Ablation on the strength λ_{RSA} of RSA can be found in Appendix G.

Table F.1 shows the mutual-KNN scores before and after further finetuning LLaVA-OneVision-7B. We observe that further finetuning with RSA loss bring better alignment between the vision and language backbones, demonstrating the effectiveness of RSA loss on cross-modal alignment. Surprisingly, Euclidean-based RSA benefits the alignment between vision and language backbones more than the cosine-based one. This determines our choice of distance metric for RSA loss computation.

G Ablation on Strengh of RSA Loss

We conduct an ablation study on the strength λ_{RSA} of RSA loss. For Euclidean-based RSA, we consider $\lambda_{RSA} \in \{1.0, 0.1, 0.01\}$, given that we find the values of Euclidean-based RSA loss is usually very large empirically. For cosine-based RSA, we try $\lambda_{RSA} \in \{1.0, 10.0, 100.0\}$, given that we observe the values of Euclidean-based RSA loss is usually very small. We run the same further finetuning setup as in Appendix F on LLaVA-OneVision-7B on Stage-1 data.

		MSCOCO	Flickr30K	WIT
	Random	0.010		
	Vision-only	0.530	0.440	0.510
	Language-only	0.830	0.810	0.800
	Off-the-shelf	0.174	0.107	0.078
Euclidean Distance	w/o RSA	0.171	0.100	0.071
	w/ RSA	0.203	0.120	0.094
	Δ	+0.032	+0.020	+0.023
Cosine Distance	w/o RSA	0.171	0.100	0.071
	w/ RSA	0.185	0.109	0.013
	Δ	+0.015	+0.009	-0.058

Table F.1: Ablation study on distance metrics used for RSA loss. We compare Euclidean and cosine distance when computing the RSA loss by further finetuning the off-the-shelf LLaVA-OneVision-7B on Stage-1 data. For both metrics, RSA consistently improves mutual-KNN alignment across MSCOCO, Flickr30K, and WIT, confirming the effectiveness of the RSA objective. Interestingly, Euclidean-based RSA yields larger gains than cosine-based RSA despite the high dimensionality of the latent spaces, motivating our choice of Euclidean distance as the distance metric for RSA.

		MSCOCO	Flickr30K	WIT
	w/o RSA	0.171	0.100	0.071
Euclidean Distance	1.0	0.120	0.071	0.055
	0.1	0.164	0.096	0.072
	0.01	**0.203**	**0.120**	**0.094**
Cosine Distance	1	0.185	0.109	0.013
	10	0.180	0.107	0.015
	100	0.165	0.097	0.015

Table G.1: Ablation on the strength λ_{RSA} of the RSA loss. We further finetune off-the-shelf LLaVA-OneVision-7B on Stage-1 data to evaluate how different loss weights affect cross-modal alignment. Euclidean-based RSA with $\lambda_{RSA} = 0.01$ achieves the highest alignment, which determines our use of this setup for our main experiments.

Table G.1 shows the mutual-KNN scores with different λ_{RSA} for Euclidean- and cosine-based RSA. We observe that λ_{RSA}=0.01 for Euclidean-based reaches the best alignment between vision and language backbones across all setups, motivating us to use this setup for our main experiments. We also find that when Euclidean-based RSA is paired with larger λ_{RSA}, the alignment will decrease, indicating that alignment-wise overfitting happens during the further finetuning on Stage-1 data. More aggressive sweep of λ_{RSA} can be done in order to look for a better combination between the distance metric and λ_{RSA}.

H Impact of Multimodal Training on Cross-Modal Alignment

It remains unclear how multimodal training (i.e., Stage-1/-1.5/-2 in Table B.1 would impact the alignment between the unimodal backbones in VLMs. Therefore, we evaluate the alignment between unimodal backbones of LLaVA-OneVision (Li et al.) and InternVL3 (Zhu et al., 2025) before and after multimodal training. For the alignment after multimodal training, we directly measure the alignment on the off-the-shelf models. For the alignment before multimodal training, we measure the alignment between the vision backbone and language backbone before any multimodal training.

Table G.2 shows the mutual-KNN scores before and after multimodal training. For both LLaVA-OneVision-7B and InternVL3-8B, the alignment between unimodal backbones before and after multimodal training changes only marginally, and even decreases on some datasets. This indicates that multimodal training primarily teaches the models to perform well on the training tasks, rather than aligning the internal representations of the vision and language backbones. Compared to LLaVA-OneVision-7B, InternVL3-8B reaches better alignment after the multimodal training. This can be

		MSCOCO	Flickr30k	WIT
Random	-	0.010		
Vision-only	-	0.530	0.440	0.510
Language-only	-	0.830	0.810	0.800
LLaVA-OneVision-7B	Before	0.206	0.128	0.064
	After	0.174	0.107	0.078
	Δ	-0.032	-0.021	+0.014
InternVL3-8B	Before	0.223	0.119	0.078
	After	0.218	0.123	0.119
	Δ	-0.005	+0.004	+0.041

Table G.2: Alignment between unimodal backbones before and after multimodal training. The low alignment differences before and after multimodal training indicate that large-scale multimodal training does not substantially improve the alignment between unimodal backbones, and in some cases pushes the modalities further apart (e.g., MSCOCO). The substantial gap relative to Vision-only and Language-only highlights that existing multimodal training pipelines do not resolve the modality gap yet, motivating the need for explicit alignment objectives such as RSA.

caused by the fact that InternVL3-8B optimizes all parameters throughout its multimodal training. However, there remains a substantial gap between cross-modal alignment and Vision-/Language-only. These findings validate that existing multimodal training pipelines do not resolve the modality gap, and thus motivate the need for explicit alignment objectives.

I Qualitative Examples of VLMs

Figure I.1 shows evaluation examples of LLaVA-OneVision-7B trained with or without RSA loss on OK-VQA (Marino et al., 2019). In general, model trained with RSA loss can understand and reason over the image and instruction better than the model trained without RSA loss, indicating that cross-modal alignment can benefit visual perception, understanding and reasoning capabilities of VLMs.

J Limitations

While our work demonstrates that aligning representational space structures across vision backbone and language backbone of VLMs improves both cross-modal alignment and downstream task performance, several limitations remain. First, an ablation on which layers of the vision and language backbones are most suitable for RSA has not been conducted yet. We only apply RSA to the final layers, but aligning earlier or multiple layers may yield better effects. Second, our exploration of distance metrics is limited to Euclidean and cosine distance; other relational measures (e.g., Wasserstein, Manhattan distances) may provide stronger alignment signals. Third, all experiments are conducted on 1.5B- or 7B-scale models, and it remains unclear how RSA scales to even larger VLMs. We leave a more exhaustive study of layer choices, distance metrics, and model scales to future work.

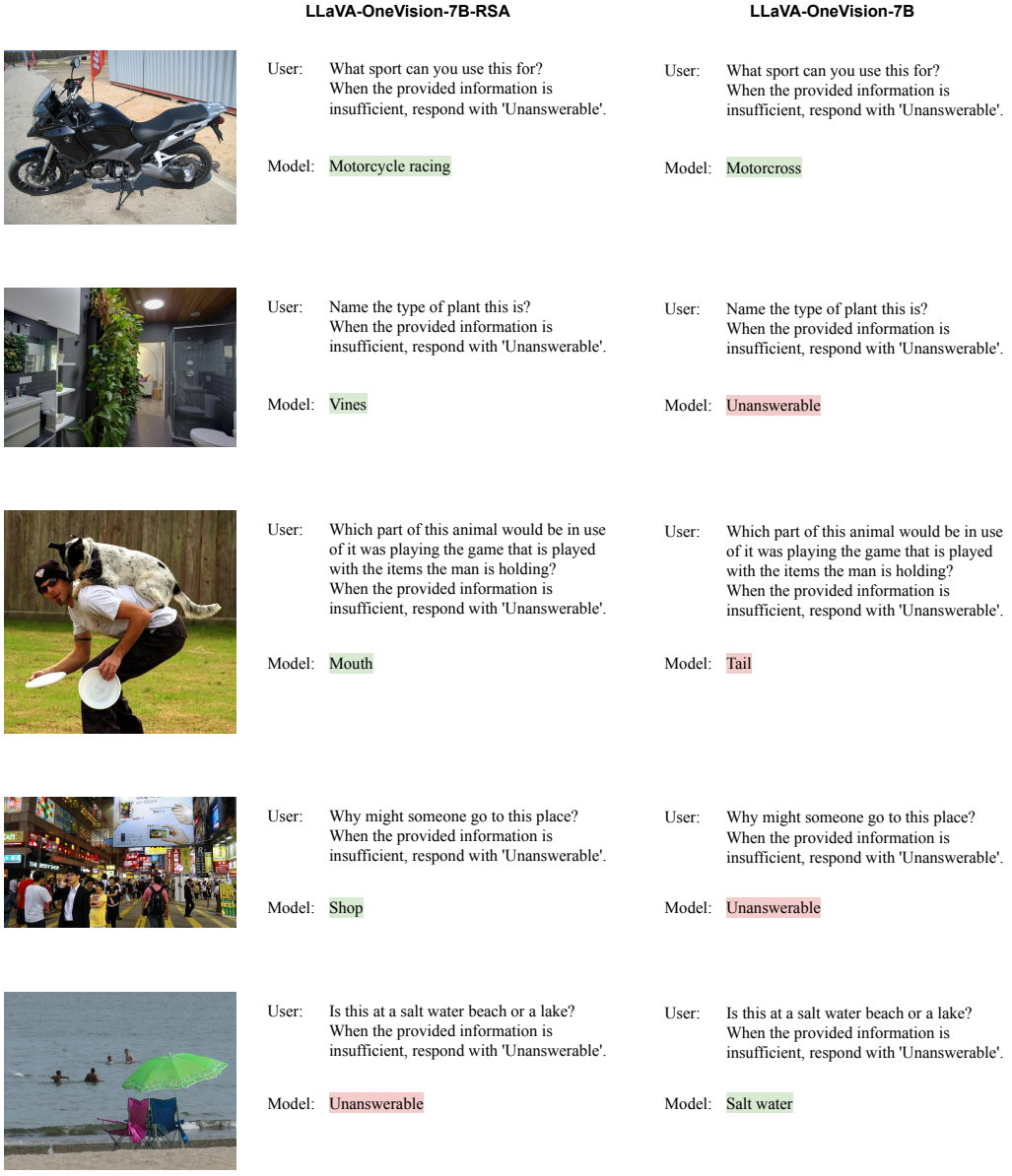

Figure I.1: Evaluation examples on OK-VQA on LLaVA-OneVision-7B trained with or without RSA loss.