LoMo: Local Modality Substitution for Deeper Vision-Language Fusion

Feng Han[1,2] **Zhixiong Zhang**[2,3] **Zheming Liang**[2,4] **Yibin Wang**[1,2] **Jiaqi Wang**[2,5,*]

[1]Fudan University [2]Shanghai Innovation Institute [3]Shanghai Jiao Tong University
[4]University of Science and Technology of China [5]JD.COM

https://maplebb.github.io/LoMo

Abstract

Vision-Language Models (VLMs) have achieved substantial progress across a wide range of understanding and reasoning tasks, driven by large-scale image-text training aimed at multimodal fusion. Ideally, replacing a textual question with its rendered-image counterpart should leave model performance essentially unaffected. In practice, however, such modality substitution induces dramatic performance degradation. We trace this "carrier sensitivity" issue to an inherent bias in current training corpora. Across prevalent datasets such as image captioning, VQA, OCR, and web-sourced interleaved data, text and images are typically organized into distinct and asymmetric roles, with text serving as linguistic queries and images as visual references. Such data bias leads VLMs to exhibit distinct preferences for information acquisition across different modalities. Consequently, VLMs fail to align representations of semantically equivalent content across textual and visual carriers, making model reasoning fragile under modality substitution. To address this, we propose **Lo**cal **Mo**dality Substitution (**LoMo**), a lightweight, architecture-agnostic data curation paradigm designed to provide supervision for cross-modal representational invariance between semantically equivalent text and image carriers. LoMo achieves this by reformulating single-modality prompts into seamlessly interleaved multimodal sequences. It dynamically selects target text spans and recasts them as rendered images, thereby preserving the same semantics across "text, visual, text" carriers. Extensive experiments across 13 diverse multimodal benchmarks demonstrate that LoMo significantly improves overall multimodal reasoning and yields deeper cross-modal fusion. Specifically, it delivers consistent gains across foundational models, improving over standard SFT by 2.68 points on LLaVA-OneVision-1.5-8B and 2.82 points on Qwen3.5-9B.

1 Introduction

Vision-Language Models (VLMs) have demonstrated strong generalization across diverse visual-language understanding tasks. Driven by rich image-text corpora and large-scale training aimed at multimodal fusion, state-of-the-art VLMs [1, 2, 12, 42, 27] exhibit powerful capabilities in tasks such as visual question answering, image captioning, document understanding, and visual grounding [26, 19, 30]. Ideally, replacing the text of a multimodal query with its rendered-image counterpart should keep model performance largely stable. In practice, however, such modality substitution causes mainstream VLMs to suffer consistent and significant performance drops across multiple benchmarks, as shown in Figure 1(a). This exposes a severe **carrier sensitivity** problem.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

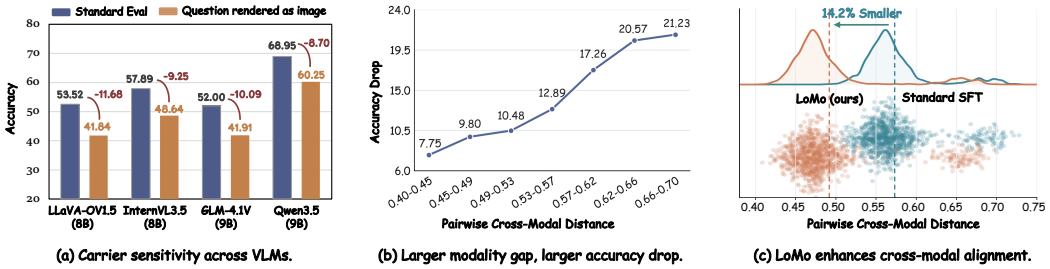


Figure 1: **Current Vision-Language Models exhibit carrier sensitivity driven by an underlying modality gap. (a) Carrier sensitivity across VLMs.** Simply shifting identical semantic content from a text format to a visual format (rendering standard questions as images) causes consistent and significant accuracy drops across state-of-the-art models. **(b) The physical manifestation of the modality gap.** By measuring the pairwise cross-modal distance between the original text and its rendered-image counterpart, we observe a strict monotonic trend, where greater representational distance between the two carriers corresponds to more severe accuracy degradation. Samples are grouped into equal-width bins of this distance, with bin edges rounded to two decimals for display. **(c) LoMo enhances cross-modal alignment.** Our method shifts the cross-modal distance distribution markedly toward smaller values, reducing the average distance by **14.2%** compared to Standard SFT and yielding tighter cross-carrier alignment.

Although current VLMs process images and text jointly, their reasoning remains highly dependent on the modality carrier through which semantic content is presented. Merely switching identical semantics from a text carrier to a visual carrier can markedly degrade performance.

To trace this degradation to its source, we extract the hidden states of text inputs and their rendered-image counterparts, and measure their pairwise cosine distances. Grouping samples by this distance reveals a strict monotonic trend, where the average accuracy drop grows from 7.75% in the closest bin to 21.23% in the farthest (Figure 1(b)). This result indicates that the performance degradation is closely associated with a cross-carrier modality gap between semantically equivalent textual and visual inputs. We identify an inherent bias in current multimodal training corpora as a key contributing factor to this gap. Across prevalent datasets such as image captioning, VQA, OCR, and web-sourced interleaved data, text and images are typically organized into distinct and asymmetric roles. Text often serves as linguistic instructions or queries, while images mainly provide visual references or evidence. Such data bias leads VLMs to exhibit distinct preferences for information acquisition across different modalities. Consequently, VLMs fail to align representations of semantically equivalent content across textual and visual carriers.

Motivated by this, we propose **LoMo**, a lightweight and architecture-agnostic data curation paradigm designed to provide supervision for cross-modal representational invariance through local modality substitution, as shown in Figure 2. LoMo reformulates single-modality prompts into seamlessly interleaved multimodal sequences while preserving the original supervision target. In this way, the standard Supervised Fine-Tuning (SFT) objective is transformed into an implicit cross-carrier alignment signal that encourages the model to associate interleaved image-text inputs with their pure-text semantic counterparts. Specifically, LoMo consists of three sequential stages. **(1) Structure-Aware Span Localization** segments a text-only instance based on its semantic structure to identify target content for visualization. **(2) Visual Rendering** recasts the selected span into a rendered visual carrier and embeds it between the surrounding text tokens, forming a "text $\rightarrow$ visual $\rightarrow$ text" sequence that promotes context-level fusion across modalities. **(3) Perceptual Distortion** applies real-world degradations to the visual carrier, ensuring that the learned fusion remains robust under perceptually challenging conditions. Crucially, LoMo is compatible with any multimodal training pipeline, requires no architectural modifications, introduces zero inference overhead, and demands no additional annotations.

Comprehensive experiments show that LoMo strengthens cross-modal fusion and delivers consistent gains across a wide spectrum of multimodal tasks. At the feature level, LoMo reduces the pairwise cross-modal distance by 14.2% compared to the standard SFT model, indicating tighter cross-carrier alignment, as shown in Figure 1(c). Moreover, on 13 benchmarks spanning mathematical reasoning, VQA, OCR, document understanding, and visual perception, LoMo improves over the standard multimodal SFT baseline by 2.68 points on LLaVA-OneVision-1.5-8B and 2.82 points on Qwen3.5-

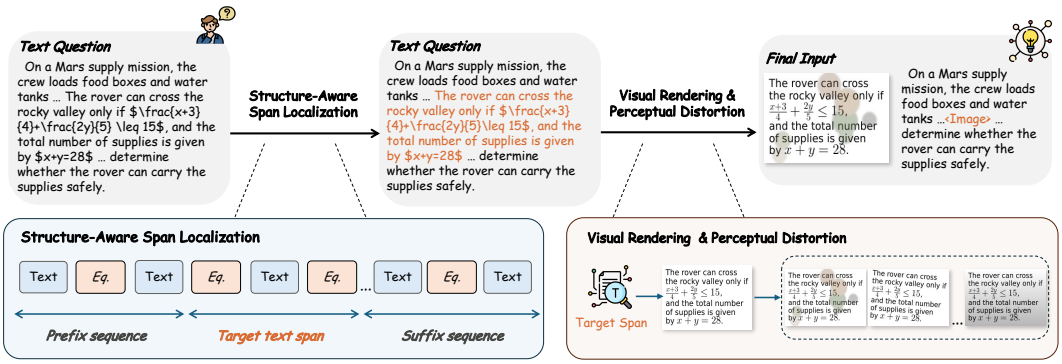


Figure 2: **Overview of LoMo.** LoMo reformulates a text-only instance into a text–image interleaved sequence through three stages. **Structure-Aware Span Localization** chunks the input in a formula-aware manner and selects a semantically coherent middle span as the target for visualization. **Visual Rendering** converts the target span into an image via content-aware routing between LaTeX and standard text renderers. The image is then perturbed by **Perceptual Distortion** and substituted back into the original position, forming a "text → visual carrier → text" instance.

9B, yielding stable improvements across backbones, as shown in Figure 3. We further evaluate our method across data scales, where LoMo yields improvements in both downstream accuracy and representation-alignment metrics. Complementary analyses on the Modality Integration Rate [13] further confirm that LoMo substantially enhances cross-modal fusion.

Our contributions are three-fold. 1) We systematically diagnose the **carrier sensitivity** problem in VLMs, revealing that it is closely associated with a cross-carrier modality gap linked to the distinct and asymmetric roles of text and images in standard training corpora. 2) We propose LoMo, a data-centric paradigm that performs local modality substitution to provide supervision for cross-modal representational invariance without architectural modifications or inference overhead. 3) We extensively validate LoMo on 13 multimodal benchmarks, demonstrating consistent accuracy improvements alongside improved cross-carrier representational consistency, with average gains of 2.68 and 2.82 on LLaVA-OneVision-1.5-8B and Qwen3.5-9B, respectively.

2 Related Work

Vision-Language Models. Vision-language models (VLMs) extend LLMs to jointly process visual and textual inputs, typically by aligning a pretrained vision encoder with an LLM backbone. Architecturally, LLaVA [24, 23] established the simple ViT–MLP–LLM template, which has been scaled by InternVL [4] and refined through systematic exploration of vision encoders and connectors [38, 39]. On the training side, recent open-source families have improved data curation and post-training: LLaVA-OneVision-1.5 [1] restructures the SFT corpus, Mantis [15] reformats interleaved multi-image instructions, and Insight-V [9] introduces long-chain visual reasoning data. Qwen3-VL [2], InternVL3.5 [42], and GLM-4.1V-Thinking [12] further push performance via larger backbones and reinforcement learning. Despite their architectural and training-side advances, these recipes consistently treat text and images as modality-specific inputs, with text serving as instructions and images as visual scenes.

Text-as-Pixels Modeling. In parallel, another line of work [46, 41, 16, 5, 44] has explored modeling text in pixel form rather than as discrete tokens. Early efforts in OCR-free document understanding, such as Pix2Struct [17], learn to parse rendered text through screenshot pretraining. Latent Compression Learning [48] pushes this further by training vision encoders directly on web-scale image–text documents through a compression objective. More recently, Glyph [5] renders long documents into compact images to extend the effective context window of VLMs, and DeepSeek-OCR [44] formalizes this idea as *contexts optical compression*, achieving high decoding accuracy at $10\times$ token compression. A recent study [21] further shows that even off-the-shelf VLMs can read rendered text inputs with roughly half the decoder tokens at little accuracy cost. VIM [20] and SimpleOCR [32] render entire instructions or questions as images to improve reading of rendered text, and See-Tok [47] replaces subword tokenization with visual reading. These methods treat text-as-pixels as an efficiency-driven *substitute* for text-as-tokens, aiming at OCR-style decoding or context compression.

In contrast, our method treats text-as-pixels as a *complement* to text-as-tokens within a single training instance, inducing an implicit cross-modal alignment supervision between the two carriers.

Modality Gap and Cross-Modal Alignment. Aligning visual and textual representations remains a long-standing challenge for multimodal models. The *modality gap* [22] was first identified in CLIP-style models, where image and text embeddings occupy disjoint regions of the shared space. Subsequent analysis [35] traces this phenomenon to information imbalance between images and captions, and shows that closing the gap can improve downstream performance. Within decoder-based VLMs, the visual embedding space inherited from CLIP has been shown to carry systematic blind spots that propagate into MLLMs [39], and the Modality Integration Rate (MIR) [13] reveals that a measurable text–vision distribution gap persists in the shallow LLM layers even after large-scale instruction tuning. VISTA-Bench [25] further documents this gap with a paired text–pixel benchmark, REST [40] links internal representational alignment to cross-modal answer consistency, and a recent behavioral analysis [36] disentangles OCR failure from reasoning collapse. The same misalignment also drives multimodal hallucinations, motivating decoding-time fixes such as VCD [18] and preference-optimization methods such as HA-DPO [37]. These remedies operate at the decoding, or objective level. In contrast, our method addresses the same gap from the data side, reformulating text-only instances into text→visual→text interleaved sequences so that cross-carrier alignment becomes a task-level requirement during standard SFT, with no architectural change and no inference overhead.

3 Methodology

3.1 Overview and Formulation

Overview. As discussed in Section 1, current multimodal training paradigms lack explicit supervision for cross-modal representational invariance, leaving VLMs vulnerable to carrier sensitivity. To address this limitation, we propose **LoMo**, a data curation paradigm that provides an implicit cross-modal alignment supervision signal through local modality substitution. As illustrated in Figure 2, LoMo dynamically recasts a selected text span into a visual carrier through three successive stages. In **Structure-Aware Span Localization**, the input text is segmented into three parts, with the middle span identified as the target content. In **Visual Rendering**, the selected span is converted into images through a content-aware rendering pipeline. Finally, **Perceptual Distortion** applies semantics-preserving degradations to the rendered image, which is then substituted back into the position of the selected span, yielding a text–image interleaved instance. This reformulation is architecture-agnostic and compatible with any multimodal training pipeline, requiring no architectural changes, no additional annotations, and no inference overheads.

Formulation. Let (x, a) denote an original text-only instance, where x is the question and a is the ground-truth answer. LoMo transforms x through three successive operators, Structure-Aware Span Localization $\mathcal{S}(\cdot)$, Visual Rendering $\mathcal{R}(\cdot)$, and Perceptual Distortion $\mathcal{A}(\cdot)$. Formally,

$$(x_{\text{pre}},\, x_{\text{mid}},\, x_{\text{suf}}) = \mathcal{S}(x), \qquad I' = \mathcal{A}\big(\mathcal{R}(x_{\text{mid}})\big), \tag{1}$$

which together produce the final mapping

$$\mathcal{T}(x) \triangleq (x_{\text{pre}},\, I',\, x_{\text{suf}}), \qquad (x, a) \;\longrightarrow\; \big((x_{\text{pre}},\, I',\, x_{\text{suf}}),\, a\big) \tag{2}$$

The resulting instance forms a "text → visual → text" skeleton, requiring the model to jointly comprehend the surrounding textual context and the embedded visual carrier in order to recover the full semantics and predict a.

3.2 Implementation of LoMo

The carrier-substitution operator $\mathcal{T}(\cdot)$ is realized through three successive stages, jointly transforming a text-only instance (x, a) into a text-image interleaved instance $(\mathcal{T}(x), a)$ while preserving the supervision target.

Structure-Aware Span Localization ($\mathcal{S}$) identifies a semantically coherent target span x_{mid} for substitution. We first estimate the input length by sentence count. Short instances are taken entirely as x_{mid} to fully preserve their semantic context, while long instances undergo a lightweight formula-aware chunking step that treats explicit mathematical expressions and common LaTeX commands as

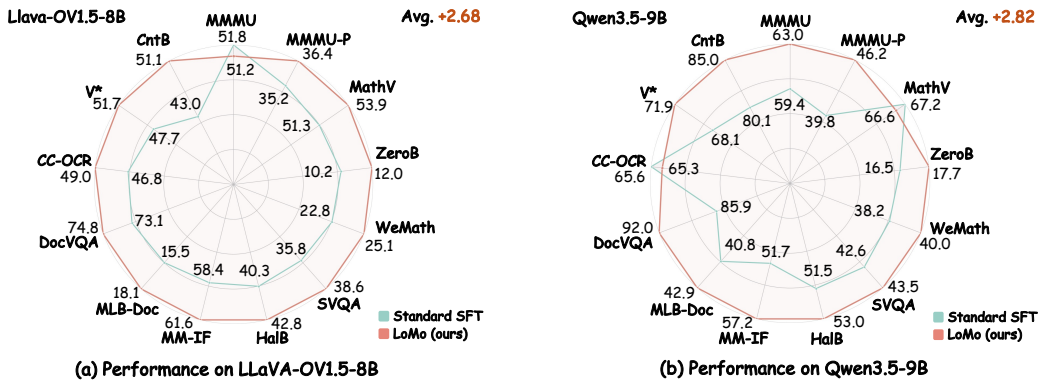


Figure 3: LoMo yields consistent improvements over Standard SFT across two backbones (**(a)** +2.68 on LLaVA-OV1.5-8B; **(b)** +2.82 on Qwen3.5-9B over 13 benchmarks).

atomic, indivisible units. After chunking, the text x is represented as an interleaved sequence of text and formula blocks,

$$x \mapsto ((t_1, l_1), (m_1, l_2), \ldots, (t_n, l_{2n-1})), \tag{3}$$

where t and m denote text and formula blocks respectively, and l records the length of each block in characters. Guided by this representation, we extract the middle one-third of the sequence at block-level granularity as x_{mid}, ensuring that truncation boundaries never fall within an equation. The surrounding text x_{pre} and x_{suf} are retained, forming a "text $\rightarrow$ visual $\rightarrow$ text" skeleton that compels the model to fuse both carriers in order to recover the full semantics and predict a.

Visual Rendering ($\mathcal{R}$) converts x_{mid} into a rendered image through a content-aware routing pipeline that adapts to the properties of each span. Spans containing mathematical expressions are routed to a LaTeX-based renderer, which yields substantially more reliable formula typesetting than general-purpose text rendering, while spans without mathematical content are routed to a standard text-rendering pipeline. To safeguard throughput at scale, the renderer is wrapped in a fallback mechanism that automatically re-routes any LaTeX failure to the text renderer rather than discarding the instance. A mild margin-trimming step further removes large empty regions while preserving all rendered content, keeping image sizes bounded without altering their semantics.

Perceptual Distortion ($\mathcal{A}$) further perturbs each rendered image with semantics-preserving degradations, simulating the distortions document images commonly undergo during real-world capture and ensuring that the learned cross-carrier alignment is anchored to the underlying semantics. We define four sets of operations that jointly cover the range of perceptual noise observed in practical scenarios. *Rotate* applies a large-angle or small-angle rotation to simulate orientation variations and slight tilt during capture. *Blur* applies Gaussian, box, or motion blur to simulate camera shake. *Shadow-or-stain* overlays edge shadows or surface stains to replicate uneven illumination and physical contamination, and *Wave* induces local geometric deformations typical of folded paper or scanning artifacts. The final augmented image I' is obtained by sampling one operation or by leaving the image unchanged.

3.3 Implicit Cross-Modal Alignment Supervision of LoMo

We examine how the local modality substitution of LoMo in Section 3.1 reshapes the supervision signal of standard SFT. Standard SFT optimizes f_θ on each text-only instance (x, a) through the negative log-likelihood

$$\mathcal{L}_{\text{SFT}}(\theta; x, a) = -\log p_\theta(a \mid x). \tag{4}$$

We apply the same loss to the substituted carrier $\mathcal{T}(x)$, so the only change lies in the carrier through which the instance is presented. We show below that the effect of this change follows from the two carriers sharing a common supervision target.

Let $p_x = p_\theta(\cdot \mid x)$ and $p_T = p_\theta(\cdot \mid \mathcal{T}(x))$ denote the predictive distributions over answers under the two carriers. The supervision target is determined by the semantic content of the instance rather than by the carrier through which it is presented. Since $\mathcal{T}$ preserves this content, the two carriers share the same target distribution, $q(a \mid x) = q(a \mid \mathcal{T}(x)) \equiv q_x(a)$.

With rewrite ratio r, each text-only instance is presented in its original form with probability $1-r$ and in its substituted form with probability r. Taking the expectation over this sampling, the training loss is

$$\mathcal{R}_r = \mathbb{E}_x\big[(1-r)\,H(q_x, p_x) + r\,H(q_x, p_T)\big], \tag{5}$$

where $H(q,p) = -\sum_a q(a)\log p(a)$. Since $H(q_x, p) = H(q_x) + \mathrm{KL}(q_x\|p)$ and $H(q_x)$ does not depend on θ, minimizing Eq. 5 is equivalent to minimizing

$$\Delta\mathcal{R}_r = \mathbb{E}_x\big[(1-r)\,\mathrm{KL}(q_x\|p_x) + r\,\mathrm{KL}(q_x\|p_T)\big]. \tag{6}$$

Equation 5 is the standard SFT cross-entropy applied under carrier resampling, and Eq. 6 is its KL form. The two differ only by the constant $\mathbb{E}_x H(q_x)$ and share the same minimizer over θ. Equation 6 shows that both p_x and p_T are fitted to the same fixed target q_x.

Let $\mathrm{TV}(p,q) = \frac{1}{2}\sum_a |p(a) - q(a)|$ denote the total variation distance. Applying the triangle inequality through the shared target q_x, followed by Pinsker's and Jensen's inequalities, gives

$$\mathbb{E}_x\big[\mathrm{TV}(p_x, p_T)\big] \le \sqrt{2\,\Delta\mathcal{R}_{1/2}} \tag{7}$$

at $r = 1/2$. The disagreement between the two carriers is thus bounded by the training loss itself. Since the two KL terms inside this bound are exactly the terms that Eq. 6 minimizes, the predictive distributions under the two carriers converge as the training loss decreases. Appendix B gives the complete derivation.

We verify Eq. 7 on 1,000 held-out GSM8K examples. At convergence LoMo reduces the token-averaged total variation distance between the two carriers by 47.03% on LLaVA-OV1.5-8B and 71.53% on Qwen3.5-9B relative to Standard SFT, and the reduction is monotone over training. The measurement protocol and the full results are reported in Appendix B.

4 Experiments

4.1 Experimental Setup

Models and training data. We examine LoMo on two open-source VLM backbones with substantially different architectures: **LLaVA-OneVision1.5-8B-Base** [1] and **Qwen3.5-9B-Base** [2]. The training data is randomly sampled from the official LLaVA-OneVision1.5 SFT corpus [1], comprising two million multimodal instruction examples and two million text-only instruction examples. The Standard SFT baseline directly fine-tunes on this pool. LoMo shares the same data pool, optimizer, learning-rate schedule, and number of training steps. The only difference is that a fraction of the text-only examples is reformatted into interleaved text-visual sequences. By construction, the two methods are matched in data scale, compute, and hyperparameters. Training hyperparameters and other implementation details of LoMo are provided in the Appendix D.

Evaluation benchmarks. We report results on 13 multimodal benchmarks spanning six categories. On general reasoning, we evaluate MMMU [50] and MMMU-Pro [51]. Math reasoning is covered by MathVista [28], ZeroBench [34], and WeMath [33]. We assess factuality with SimpleVQA [6] and HallusionBench [11], and measure instruction following with MM-IFEval [8]. Document and OCR understanding is probed via MMLongBench-Doc [29], DocVQA [30], and CC-OCR [49]. Finally, V^* [45] and CountBench [31] target visual perception. All evaluations are conducted with EvalScope under identical prompting and decoding configurations.

Evaluation protocols. We evaluate every benchmark under two protocols. *Standard Evaluation* feeds the original (image, text question) pair to the model, matching standard practice. *Rendered Evaluation* renders the entire text question as a single image, which replaces the original text and is fed to the model together with the original image. The linguistic content is identical across the two protocols and only the input modality differs.

Cross-modal alignment metrics. Beyond accuracy, we adopt two intrinsic metrics to probe the model's internal cross-modal alignment. *(i) Modality Integration Rate (MIR)* [13] quantifies the distributional gap between visual and textual tokens inside the VLM. Specifically, at each decoder layer the hidden states of visual and textual tokens are extracted and viewed as samples from two high-dimensional distributions, whose discrepancy is measured by the Fréchet Distance (FID). The per-layer FID computation follows the original paper. Since different backbones differ in the number

Table 1: Main results across 13 multimodal benchmarks under two evaluation protocols. **Standard Evaluation** feeds the original multimodal inputs (image + text question) to the model. **Rendered Evaluation** renders the entire text question as a single image. Δ denotes the absolute change of LoMo over Standard SFT; $\uparrow$ and $\downarrow$ indicate gains and drops. Benchmark abbreviations: MMMU-P: MMMU-Pro, MathV: MathVista, ZeroB: ZeroBench, SVQA: SimpleVQA, HalB: HallusionBench, MM-IF: MM-IFEval, MLB-Doc: MMLongBench-Doc, CntB: CountBench.

Model	Setting	Reasoning		Math			Factuality		Instruct.	Document & OCR			Visual Percept.		Avg.
		MMMU	MMMU-P	MathV	ZeroB	WeMath	SVQA	HalB	MM-IF	MLB-Doc	DocVQA	CC-OCR	V*	CntB	
Standard Evaluation															
LLaVA-OV1.5-8B	Standard SFT	51.78	35.24	51.30	10.18	22.76	35.51	40.35	58.40	15.49	73.05	46.76	47.71	42.97	40.88
	+ LoMo	**51.22**	**36.36**	**53.90**	**11.98**	**25.14**	**38.62**	**42.82**	**61.61**	**18.06**	**74.77**	**48.97**	**51.70**	**51.12**	**43.56**
	Δ	$\downarrow$0.56	$\uparrow$1.12	$\uparrow$2.60	$\uparrow$1.80	$\uparrow$2.38	$\uparrow$3.11	$\uparrow$2.47	$\uparrow$3.21	$\uparrow$2.57	$\uparrow$1.72	$\uparrow$2.21	$\uparrow$3.99	$\uparrow$8.15	$\uparrow$2.68
Qwen3.5-9B	Standard SFT	59.44	39.83	67.20	16.54	38.19	42.57	51.53	51.74	40.79	85.89	65.61	68.13	80.08	54.43
	+ LoMo	**63.00**	**46.18**	**66.60**	**17.66**	**40.00**	**43.51**	**52.99**	**57.23**	**42.90**	**91.99**	**65.31**	**71.86**	**85.01**	**57.25**
	Δ	$\uparrow$3.56	$\uparrow$6.35	$\downarrow$0.60	$\uparrow$1.12	$\uparrow$1.81	$\uparrow$0.94	$\uparrow$1.46	$\uparrow$5.49	$\uparrow$2.11	$\uparrow$6.10	$\downarrow$0.30	$\uparrow$3.73	$\uparrow$4.93	$\uparrow$2.82
Rendered Evaluation															
LLaVA-OV1.5-8B	Standard SFT	21.22	16.01	18.10	3.59	0.95	22.42	41.36	28.83	3.94	15.38	14.49	8.12	3.67	15.24
	+ LoMo	**35.56**	**27.59**	**39.50**	**7.63**	**11.33**	**31.06**	**61.65**	**48.83**	**6.69**	**58.89**	**39.89**	**36.13**	**38.49**	**34.10**
	Δ	$\uparrow$14.34	$\uparrow$11.58	$\uparrow$21.40	$\uparrow$4.04	$\uparrow$10.38	$\uparrow$8.64	$\uparrow$20.29	$\uparrow$20.00	$\uparrow$2.75	$\uparrow$43.51	$\uparrow$25.40	$\uparrow$28.01	$\uparrow$34.82	$\uparrow$18.86
Qwen3.5-9B	Standard SFT	49.52	33.06	56.90	15.94	23.43	39.95	43.37	41.25	28.87	47.24	49.60	61.58	71.66	43.26
	+ LoMo	**62.48**	**45.29**	**65.50**	**16.62**	**39.72**	**43.26**	**47.05**	**54.32**	**36.57**	**91.73**	**66.14**	**64.73**	**83.98**	**55.18**
	Δ	$\uparrow$12.96	$\uparrow$12.23	$\uparrow$8.60	$\uparrow$0.68	$\uparrow$16.29	$\uparrow$3.31	$\uparrow$3.68	$\uparrow$13.07	$\uparrow$7.70	$\uparrow$44.49	$\uparrow$16.54	$\uparrow$3.15	$\uparrow$12.32	$\uparrow$11.92

of decoder layers, we report the layer-wise mean of FID as MIR. A lower MIR indicates a smaller distributional gap between textual and visual representations, reflecting tighter cross-modal integration. *(ii) Pairwise Cross-Modal Distance* is a sample-level alignment metric. For each evaluation sample, we compute the mean hidden states of its text tokens and the corresponding rendered-image tokens at the output of the first VLM self-attention layer, denoted $\bar{h}_{\text{text}}$ and $\bar{h}_{\text{img}}$, and define their cosine distance as:

$$d = 1 - \cos(\bar{h}_{\text{text}}, \bar{h}_{\text{img}}). \tag{8}$$

We average d over the evaluation set; a lower value indicates that paired text and rendered image lie closer in representation space.

4.2 Main Results

Table 1 reports performance under both evaluation protocols across all 13 benchmarks.

Standard Evaluation. The upper block of Table 1 reports Standard results on both backbones. LoMo consistently outperforms Standard SFT, with average gains of +2.68 on LLaVA-OneVision-8B and +2.82 on Qwen3.5-9B. The improvements are most pronounced on instruction following (MM-IFEval: +3.21 / +5.49) and visual perception (CountBench: +8.15 / +4.93; V*: +3.99 / +3.73), with consistent gains on factuality (SimpleVQA: +3.11 / +0.94; HallusionBench: +2.47 / +1.46), document & OCR (DocVQA: +1.72 / +6.10; MMLongBench-Doc: +2.57 / +2.11), and math reasoning (WeMath: +2.38 / +1.81). Overall, LoMo improves performance in 23 out of 26 comparisons, with only three marginal regressions, demonstrating its robust effectiveness across evaluation categories.

Rendered Evaluation. The lower block of Table 1 shows that the gap between LoMo and Standard SFT widens substantially when text is delivered through pixels. Average gains rise to **+18.86** on LLaVA-OneVision-8B and **+11.92** on Qwen3.5-9B, which is roughly $7\times$ and $4\times$ the corresponding Standard gains. The most dramatic improvements appear on document & OCR (DocVQA: +43.51 / +44.49; CC-OCR: +25.40 / +16.54) and visual perception (CountBench: +34.82 / +12.32; V*: +28.01 / +3.15). On Qwen3.5-9B, the Standard$\rightarrow$Rendered drop is compressed from **11.17** points under Standard SFT to just **2.07** under LoMo. In other words, models trained with Standard SFT collapse when the same linguistic content is delivered as pixels, whereas LoMo-trained models retain near-Standard performance.

We attribute the consistent gains under both protocols to LoMo's interleaved training format. By embedding rendered images between textual prefix and suffix, the model is repeatedly required to bridge text and pixels within a single sample. This fine-grained cross-modal fusion manifests as stronger visual perception and instruction following under Standard Evaluation, while also enabling more robust comprehension of pixel-rendered text under Rendered Evaluation. The key reason is that LoMo explicitly establishes correspondence between *text-as-tokens* and *text-as-pixels*, encouraging their representations to become more aligned. In contrast, Standard SFT largely preserves the

Table 2: Component ablation of LoMo on LLaVA-OV1.5-8B. *Full-Text Rendering* renders the entire input as images without our Structure-Aware Span Localization or Perceptual Distortion. PD: Perceptual Distortion. Benchmark abbreviations follow Table 1. Best average is **bold**.

Method	Reasoning		Math			Factuality		Instruct.	Document & OCR			Visual Percept.		Avg.
	MMMU	MMMU-P	MathV	ZeroB	WeMath	SVQA	HalB	MM-IF	MLB-Doc	DocVQA	CC-OCR	V*	CntB	
Standard SFT	51.78	35.24	51.30	10.18	22.76	35.51	40.35	58.40	15.49	73.05	46.76	47.71	42.97	40.88
+ Full-Text Rendering	50.56	35.53	55.90	11.68	22.76	35.60	39.45	59.51	16.87	71.80	47.35	47.32	52.55	42.07
+ LoMo w/o PD	51.00	35.38	55.40	11.98	25.33	36.59	45.17	60.00	16.59	73.77	48.15	50.65	50.31	43.10
+ LoMo	51.22	36.36	53.90	11.98	25.14	38.62	42.82	61.61	18.06	74.77	48.97	51.70	51.12	**43.56**

Table 3: **Quantitative results of different rewrite ratios on LLaVA-OV1.5-8B.** The rewrite ratio controls the fraction of text-only training samples reformatted into text–image interleaved sequences with LoMo. Δ denotes the average improvement over the Standard SFT baseline. Best average is **bold**.

Setting	Rewrite Ratio	Reasoning		Math			Factuality		Instruct.	Document & OCR			Visual Percept.		Avg.	Δ
		MMMU	MMMU-P	MathV	ZeroB	WeMath	SVQA	HalB	MM-IF	MLB-Doc	DocVQA	CC-OCR	V*	CntB		
Standard SFT	0%	51.78	35.24	51.30	10.18	22.76	35.51	40.35	58.40	15.49	73.05	46.76	47.71	42.97	40.88	–
LoMo	25%	50.44	37.19	51.60	12.13	24.67	37.33	42.84	62.58	17.14	74.52	49.84	51.18	46.24	42.90	+2.02
LoMo	50% (ours)	51.22	36.36	53.90	11.98	25.14	38.62	42.82	61.61	18.06	74.77	48.97	51.70	51.12	**43.56**	+2.68
LoMo	75%	50.56	36.13	52.40	13.62	27.14	37.23	44.25	61.22	17.51	74.74	49.19	49.08	49.03	43.24	+2.36
LoMo	100%	50.67	36.32	53.80	11.23	26.38	37.63	41.37	61.07	17.14	74.77	48.23	50.07	46.18	42.68	+1.80

functional separation between the two carriers, making the model sensitive to carrier substitution. Results on two additional benchmarks, VISTA-Bench and BLINK, are provided in Appendix C.

Cross-modal Representation Analysis. We further evaluate cross-modal representation alignment using two complementary metrics: MIR captures *set-level* alignment between the visual and textual token populations, while Paired Cross-Modal Distance captures *pair-level* alignment. Both metrics start at the same values at initialization, but diverge sharply after training. At the 4M scale (Fig. 4), LoMo reduces MIR by an additional 0.122 over Standard SFT, indicating a stronger global distributional alignment. The pair-level result is more striking: Standard SFT *increases* the paired distance from 0.52 to 0.57, suggesting that conventional SFT pushes paired text and rendered-image representations apart at the sample level, whereas LoMo reduces it to 0.49.

This divergence highlights the distinction between set-level and pair-level alignment. In Standard SFT, text and images usually play complementary roles: text specifies the query, while images provide visual content. Thus, the model can optimize the training objective without explicitly aligning semantically equivalent text and rendered-image inputs. LoMo changes this training signal by splitting the required semantic cues between textual context and the rendered span. As a result, predicting the answer requires cross-carrier integration between *text-as-tokens* and *text-as-pixels*, turning alignment into a task-level requirement. This explains why LoMo improves both global distributional alignment and paired-sample alignment.

4.3 Ablations

Component Ablation. Table 2 compares three variants against Standard SFT: **Full-Text Rendering**, which renders the entire question as images without Structure-Aware Span Localization or Perceptual Distortion; **LoMo without Perceptual Distortion (LoMo w/o PD)**, which retains Structure-Aware Span Localization but skips Perceptual Distortion; and **LoMo**, which denotes the full pipeline. Naively rendering the full question yields only a +1.19 average gain, indicating that simple exposure to rendered text is insufficient when the rendered image is treated as an isolated visual input. In contrast, **LoMo w/o PD** already improves the average score by +2.22 over Standard SFT, showing that the Structure-Aware Span Localization is the dominant contributor. Meanwhile, adding Perceptual Distortion further raises the gain to +2.68, with the largest improvements appearing on visual perception and document understanding tasks, confirming the benefit of exposing the model to realistic visual carrier variations.

Data Scale Ablation. To examine whether LoMo's advantage holds across data scales, we conduct experiments at four training scales (1M, 2M, 3M, 4M) and track three quantities along the same axis: average accuracy (Fig. 4a), set-level alignment measured by MIR (Fig. 4b), and pair-level alignment measured by Paired Cross-Modal Distance (Fig. 4c). LoMo's accuracy gain over Standard SFT

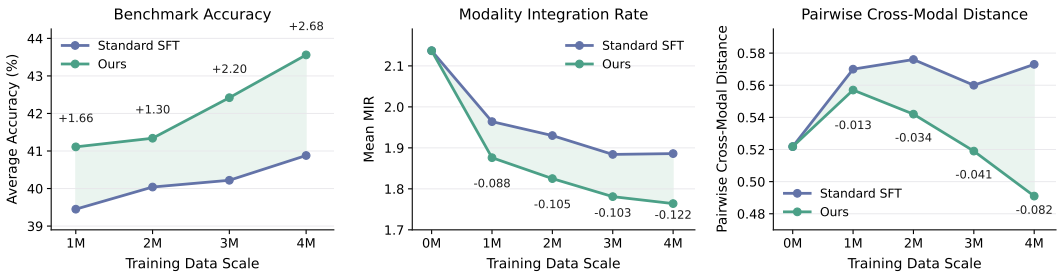


Figure 4: **LoMo consistently outperforms standard SFT across data scales on three metrics. (a)** Average accuracy on 13 multimodal benchmarks (higher is better). **(b)** Mean MIR [13], measuring text-visual representation alignment (lower is better). **(c)** Pairwise cross-modal distance, defined as $1 - \cos(\bar{h}_{\text{text}}, \bar{h}_{\text{img}})$ between hidden states of text and its rendered image (lower is better).

Table 4: **Quantitative results of different rendering positions on LLaVA-OV1.5-8B.** The position controls where the rendered image is inserted into the original text sequence. *Prefix* renders the first one-third of the prompt, yielding an image–text structure. *Middle* (ours) renders the central one-third, yielding a text–image–text structure. *Suffix* renders the last one-third. *Multi-Span* renders two short spans, yielding a text–image–text–image–text structure. Δ denotes the average improvement over the Standard SFT baseline. Best average is **bold**.

Setting	Position	Reasoning		Math			Factuality		Instruct.	Document & OCR			Visual Percept.		Avg.	Δ
		MMMU	MMMU-P	MathV	ZeroB	WeMath	SVQA	HalB	MM-IF	MLB-Doc	DocVQA	CC-OCR	V*	CntB		
Standard SFT	–	51.78	35.24	51.30	10.18	22.76	35.51	40.35	58.40	15.49	73.05	46.76	47.71	42.97	40.88	–
LoMo	Prefix	52.89	36.49	53.50	11.83	24.00	35.85	40.75	59.06	15.77	74.60	47.98	51.70	47.21	42.44	+1.56
LoMo	Middle (ours)	51.22	36.36	53.90	11.98	25.14	38.62	42.82	61.61	18.06	74.77	48.97	51.70	51.12	**43.56**	+2.68
LoMo	Suffix	50.78	36.71	52.10	10.63	26.10	37.58	40.20	60.64	15.12	75.58	49.59	49.93	45.29	42.33	+1.45
LoMo	Multi-Span	50.89	37.22	55.00	10.48	24.57	37.48	41.49	62.01	16.32	75.30	49.72	50.52	43.29	42.64	+1.76

widens from +1.66 at 1M to +2.68 at 4M, and is accompanied by progressively stronger alignment on both metrics: at 4M, LoMo's MIR is 0.122 lower than Standard SFT, and its paired distance drops to 0.49, whereas Standard SFT's paired distance instead *increases* from 0.52 to 0.57 over training. Across all four scales, LoMo improves both accuracy and cross-modal alignment over Standard SFT, indicating that its benefits are not specific to a particular training budget. The initial rise of LoMo's paired distance in Fig. 4c reflects two competing forces. Standard multimodal supervision first pushes the paired representations apart, and LoMo's cross-carrier consistency signal accumulates with training and gradually pulls them back together.

Ablation on Rewrite Ratio. Quantitative comparisons are conducted on five rewrite ratios (0%, 25%, 50%, 75%, 100%) on LLaVA-OV1.5-8B with other configurations fixed. Table 3 reports results across the 13 multimodal benchmarks. All non-zero rewrite ratios yield substantial improvements over Standard SFT, confirming that LoMo consistently benefits multimodal fusion. Besides, the average accuracy first rises with the rewrite ratio, reaches 43.56 at 50%, and then declines to 42.68 at 100%. This trend suggests that the gain from cross-carrier supervision saturates at a moderate ratio, after which further rewriting yields diminishing returns.

Ablation on Rendering Position. We compare four rendering positions on LLaVA-OV1.5-8B with all other configurations held fixed. **Prefix** and **Suffix** render the first or last one-third of the prompt, producing a single-image structure with text on one side. **Middle** renders the central one-third, producing a text–image–text structure that places the rendered span between two textual contexts. **Multi-Span** divides the prompt into multiple short segments and renders alternating segments, producing a text–image–text–image–text structure with multiple interleaving boundaries. Table 4 reports results across the 13 multimodal benchmarks. Among these positions, Middle attains the highest average accuracy of 43.56. The advantage of Middle over Prefix and Suffix indicates that placing the rendered image between two textual segments enforces stronger cross-carrier integration. Meanwhile, the result that Middle outperforms Multi-Span suggests that a single rendered span provides a more focused cross-carrier supervision signal than distributing renderings across multiple positions in the prompt.

Controlled Comparison: Beyond More Image-Bearing Samples. LoMo converts a subset of originally text-only samples into image-bearing samples by rendering selected text spans, thereby increasing the effective number of image-bearing samples for training. To examine whether LoMo's

Table 5: Controlled comparison of LoMo under different image-bearing to text-only sample ratios on LLaVA-OV1.5-8B. The original setting uses LoMo's natural 3:1 ratio after rewriting, while the matched setting controls the effective ratio back to 1:1 to match Standard SFT for a fair comparison. Δ denotes the average improvement over Standard SFT. Best average is **bold**.

Setting	Image:Text Ratio	Reasoning		Math			Factuality		Instruct.	Document & OCR			Visual Percept.		Avg.	Δ
		MMMU	MMMU-P	MathV	ZeroB	WeMath	SVQA	HalB	MM-IF	MLB-Doc	DocVQA	CC-OCR	V^*	CntB		
Standard SFT	1:1	51.78	35.24	51.30	10.18	22.76	35.51	40.35	58.40	15.49	73.05	46.76	47.71	42.97	40.88	–
LoMo	3:1 (Original)	51.22	36.36	53.90	11.98	25.14	38.62	42.82	61.61	18.06	74.77	48.97	51.70	51.12	**43.56**	+2.68
LoMo	1:1 (Matched)	51.44	36.24	51.40	12.57	24.86	38.27	43.70	62.10	18.88	74.02	48.37	51.57	49.97	43.33	+2.45

gains simply come from this increased multimodal exposure, we resample the training data pool so that Standard SFT and LoMo share the same 1:1 ratio of image-bearing to text-only samples. Table 5 shows that LoMo still outperforms Standard SFT by 2.45 points on average under this matched setting. This indicates that LoMo's gains are driven by its interleaved cross-carrier formulation rather than by simply exposing the model to more image-bearing samples.

5 Conclusion

In this work, we identify a carrier sensitivity phenomenon in current VLMs, where rendering identical textual content as a visual input causes consistent and substantial performance drops, with the magnitude tightly correlated to the cross-modal representational distance between the two carriers. We identify the asymmetric roles of text and images in standard training corpora as a key contributing factor to this gap, and propose **LoMo**, a lightweight data curation paradigm that augments standard SFT with implicit cross-modal alignment supervision through local modality substitution. We hope LoMo offers a simple data-side recipe for bridging the modality gap and inspires further exploration of treating text and visuals as truly interchangeable semantic carriers.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by JD.COM and Shanghai Innovation Institute.

Competing interests. The authors declare no competing interests.

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

[4] 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*, pages 24185–24198, 2024.

[5] Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, et al. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*, 2025.

[6] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646, 2025.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[8] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1099–1109, 2025.

[9] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072, 2025.

[10] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[11] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

[12] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[13] Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models with modality integration rate. *arXiv preprint arXiv:2410.07167*, 2024.

[14] Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pages 58791–58831, 2025.

[15] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[16] Ilker Kesen, Jonas F Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. Multilingual pretraining for pixel language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29582–29599, 2025.

[17] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[18] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[19] 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*, 2023.

[20] Xiujun Li, Yujie Lu, Zhe Gan, Jianfeng Gao, William Yang Wang, and Yejin Choi. Text as images: Can multimodal large language models follow printed instructions in pixels? *arXiv preprint arXiv:2311.17647*, 2023.

[21] Yanhong Li, Zixuan Lan, and Jiawei Zhou. Text or pixels? it takes half: On the token efficiency of visual text inputs in multimodal llms. *arXiv preprint arXiv:2510.18279*, 2025.

[22] 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.

[23] 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*, pages 26296–26306, 2024.

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

[25] Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. Vista-bench: Do vision-language models really understand visualized text as well as pure text? *arXiv preprint arXiv:2602.04802*, 2026.

[26] 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*, pages 216–233. Springer, 2024.

[27] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[28] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[29] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

[30] 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*, pages 2200–2209, 2021.

[31] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3170–3180, 2023.

[32] Yibo Peng, Peng Xia, Ding Zhong, Kaide Zeng, Siwei Han, Yiyang Zhou, Jiaqi Liu, Ruiyi Zhang, and Huaxiu Yao. Simpleocr: Rendering visual questions to teach mllms to read. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 10697–10710, 2026.

[33] Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070, 2025.

[34] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.

[35] 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. *arXiv preprint arXiv:2404.07983*, 2024.

[36] Kaiser Sun, Xiaochuang Yuan, Hongjun Liu, Chen Zhao, Cheng Zhang, Mark Dredze, and Fan Bai. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms. *arXiv preprint arXiv:2603.09095*, 2026.

[37] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

[38] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[39] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[40] Angela van Sprang, Laurens Samson, Ana Lucic, Erman Acar, Sennay Ghebreab, and Yuki M Asano. Same content, different answers: Cross-modal inconsistency in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8790, 2026.

[41] Alex Jinpeng Wang, Linjie Li, Yiqi Lin, Min Li, Lijuan Wang, and Mike Zheng Shou. Leveraging visual tokens for extended text contexts in multi-modal learning. *Advances in Neural Information Processing Systems*, 37:14325–14348, 2024.

[42] 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.

[43] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

[44] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

[45] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

[46] Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. Vision-centric token compression in large language model. *arXiv preprint arXiv:2502.00791*, 2025.

[47] Ling Xing, Rui Yan, Alex Jinpeng Wang, Zechao Li, and Jinhui Tang. See the text: From tokenization to visual reading. *arXiv preprint arXiv:2510.18840*, 2025.

[48] Chenyu Yang, Xizhou Zhu, Jinguo Zhu, Weijie Su, Junjie Wang, Xuan Dong, Wenhai Wang, Lewei Lu, Bin Li, Jie Zhou, et al. Vision model pre-training on interleaved image-text data via latent compression learning. *Advances in Neural Information Processing Systems*, 37:23912–23938, 2024.

[49] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, et al. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21744–21754, 2025.

[50] 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*, pages 9556–9567, 2024.

[51] 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)*, pages 15134–15186, 2025.

[52] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

A Pure-Text Capability Analysis

To verify the effect of LoMo on pure-text capabilities, we evaluate Standard SFT and LoMo on five pure-text benchmarks: MMLU-Pro [43], GSM8K [7], HumanEval [3], LiveCodeBench V6 [14], and IFEval [52]. As shown in Table 6, LoMo matches or slightly exceeds Standard SFT on every benchmark, with average gains of $+0.28$ and $+0.58$ on the two backbones. On Qwen3.5-9B, the pure-text IFEval gain ($+2.59$) is in the same direction as the multimodal MM-IFEval gain ($+5.49$, Table 1). These results show that LoMo improves multimodal performance without compromising pure-text capabilities, achieving small average gains on both backbones.

Table 6: Pure-text capability sanity check on LLaVA-OV1.5-8B and Qwen3.5-9B. We evaluate both Standard SFT and LoMo on five widely used pure-text benchmarks covering general knowledge, mathematical reasoning, code generation, and instruction following. Δ denotes the absolute change of LoMo over Standard SFT; $\uparrow$ and $\downarrow$ indicate gains and drops. Benchmark abbreviations: MMLU-P: MMLU-Pro, LCB-V6: LiveCodeBench V6.

Model	Setting	Reasoning	Coding		Math	Instruct.	Avg.
		MMLU-P	HumanEval	LCB-V6	GSM8K	IFEval	
LLaVA-OV1.5-8B	Standard SFT	62.58	82.62	29.19	92.87	75.42	68.54
	+ LoMo	**62.27**	**82.93**	**29.91**	**93.40**	**75.60**	**68.82**
	Δ	$\downarrow$0.31	$\uparrow$0.31	$\uparrow$0.72	$\uparrow$0.53	$\uparrow$0.18	$\uparrow$0.28
Qwen3.5-9B	Standard SFT	70.74	71.34	33.29	95.38	72.27	68.60
	+ LoMo	**71.16**	**71.95**	**33.43**	**94.54**	**74.86**	**69.19**
	Δ	$\uparrow$0.42	$\uparrow$0.61	$\uparrow$0.14	$\downarrow$0.84	$\uparrow$2.59	$\uparrow$0.58

Algorithm 1: The pipeline of LoMo.

Input: Text-only instance (x, a)
Output: Interleaved instance (U, a)
// Stage 1: Span localization
if $\mathrm{SentenceCount}(x) \leq 3$ **then**
  $(x_{\mathrm{pre}}, x_{\mathrm{mid}}, x_{\mathrm{suf}}) \leftarrow (\emptyset, x, \emptyset)$
else
  $\mathcal{C} \leftarrow \mathrm{FormulaAwareChunk}(x)$
  $(x_{\mathrm{pre}}, x_{\mathrm{mid}}, x_{\mathrm{suf}}) \leftarrow \mathrm{ExtractMiddle}(\mathcal{C})$
// Stage 2: Visual rendering
if $\mathrm{ContainsMath}(x_{\mathrm{mid}})$ **then**
  $I \leftarrow \mathrm{LaTeXRender}(x_{\mathrm{mid}})$
  if $I = \mathrm{None}$ **then** $I \leftarrow \mathrm{TextRender}(x_{\mathrm{mid}})$
else
  $I \leftarrow \mathrm{TextRender}(x_{\mathrm{mid}})$
$I \leftarrow \mathrm{TrimMargin}(I)$
// Stage 3: Perceptual distortion
$\mathcal{O} \leftarrow \{\mathrm{Clean}, \mathrm{Rotate}, \mathrm{Blur}, \mathrm{ShadowOrStain}, \mathrm{Wave}\}$
$I' \leftarrow \mathrm{SubOps}(o)(I)$, where $o \sim \mathrm{Categorical}(\mathcal{O})$
return $((x_{\mathrm{pre}}, I', x_{\mathrm{suf}}), a)$

B Derivation of the Cross-Carrier Consistency Bound

This appendix gives the complete derivation of Eq. 7, together with the measurement protocol for the empirical verification reported in Section 3.3.

B.1 Notation and Assumption

Let x be a text-only instance, a its supervision target, and $\mathcal{T}(x)$ the substituted interleaved carrier defined in Section 3.1. We write $p_x = p_\theta(\cdot \mid x)$ and $p_T = p_\theta(\cdot \mid \mathcal{T}(x))$ for the model's predictive distributions over answers under the two carriers. The supervision target is determined by the semantic content of the instance rather than by its carrier. We assume that $\mathcal{T}$ preserves the semantics of x,

under which the target distribution is carrier-independent,

$$q(a \mid x) = q(a \mid \mathcal{T}(x)) \equiv q_x(a). \tag{9}$$

When the supervision target is a single reference answer, as in the SFT corpus, q_x is the corresponding one-hot distribution. All statements below hold pointwise in x.

B.2 Training Objective

LoMo adds no term to the SFT loss. For each text-only instance a Bernoulli variable $C \sim \mathrm{Bernoulli}(r)$ selects the carrier, and the cross-entropy loss is applied to whichever carrier is presented. Taking the expectation over C and over the text-only instances gives

$$\mathcal{R}_r = \mathbb{E}_x\big[(1-r)\,H(q_x, p_x) + r\,H(q_x, p_T)\big], \qquad H(q,p) = -\sum_a q(a)\log p_\theta(a). \tag{10}$$

Using the standard identity

$$H(q_x, p) = H(q_x) + \mathrm{KL}(q_x \| p) \tag{11}$$

and noting that $H(q_x)$ is a constant with respect to θ, minimizing Eq. 10 is equivalent to minimizing

$$\Delta\mathcal{R}_r = \mathbb{E}_x\big[(1-r)\,\mathrm{KL}(q_x \| p_x) + r\,\mathrm{KL}(q_x \| p_T)\big]. \tag{12}$$

Equation 12 differs from the standard SFT objective only by the constant $\mathbb{E}_x H(q_x)$. Its KL form shows that both p_x and p_T are fitted to the same fixed target q_x.

B.3 Consistency Bound

Let $\mathrm{TV}(p,q) = \frac{1}{2}\sum_a |p(a) - q(a)|$ denote the total variation distance. The triangle inequality gives

$$\mathrm{TV}(p_x, p_T) \le \mathrm{TV}(p_x, q_x) + \mathrm{TV}(q_x, p_T). \tag{13}$$

Pinsker's inequality states that $\mathrm{TV}(q,p) \le \sqrt{\frac{1}{2}\mathrm{KL}(q\|p)}$, so

$$\mathrm{TV}(p_x, p_T) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(q_x \| p_x)} + \sqrt{\tfrac{1}{2}\mathrm{KL}(q_x \| p_T)}. \tag{14}$$

Since $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$ for $a, b \ge 0$, the right-hand side is at most

$$\sqrt{\mathrm{KL}(q_x \| p_x) + \mathrm{KL}(q_x \| p_T)}. \tag{15}$$

Taking expectations over x and applying Jensen's inequality to the concave square root yields

$$\mathbb{E}_x\big[\mathrm{TV}(p_x, p_T)\big] \le \sqrt{\mathbb{E}_x\big[\mathrm{KL}(q_x \| p_x) + \mathrm{KL}(q_x \| p_T)\big]}. \tag{16}$$

At $r = 1/2$, Eq. 12 gives $\mathbb{E}_x[\mathrm{KL}(q_x\|p_x) + \mathrm{KL}(q_x\|p_T)] = 2\,\Delta\mathcal{R}_{1/2}$, and substitution into Eq. 16 gives the bound of Eq. 7,

$$\mathbb{E}_x\big[\mathrm{TV}(p_x, p_T)\big] \le \sqrt{2\,\Delta\mathcal{R}_{1/2}}. \tag{17}$$

Since $\Delta\mathcal{R}_{1/2}$ decreases together with the training loss, Eq. 17 shows that $\mathbb{E}_x[\mathrm{TV}(p_x, p_T)]$ is driven down as training proceeds, pulling the predictive distributions under the two carriers toward agreement.

B.4 Measurement Protocol

We verify Eq. 17 on 1,000 held-out GSM8K examples. For sample i with reference answer $a_i = (a_{i,1}, \ldots, a_{i,L_i})$, teacher forcing gives both carriers the same prefix $a_{i,<t}$. We report the token-averaged cross-carrier KL

$$\overline{D} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L_i}\sum_{t=1}^{L_i}\mathrm{KL}\big(p_\theta(\cdot \mid x_i, a_{i,<t}) \,\big\|\, p_\theta(\cdot \mid \mathcal{T}(x_i), a_{i,<t})\big) \tag{18}$$

with $N = 1{,}000$, together with the token-averaged total variation distance between the same pair of distributions.

As shown in Table 7, both metrics decrease monotonically over training and stay well below Standard SFT at every checkpoint. At convergence LoMo reduces the token-averaged total variation distance by 47.03% on LLaVA-OV1.5-8B and 71.53% on Qwen3.5-9B relative to Standard SFT, and $\overline{D}$ by 70.36% and 93.18% respectively, confirming that the predictive consistency implied by Eq. 17 is realized in practice.

Table 7: $\overline{D}$ (Eq. 18) and token-averaged $\mathrm{TV}(p_x, p_T)$ between the textual and rendered carriers on 1,000 held-out GSM8K examples, at checkpoints spanning training. The final column reports LoMo's relative reduction over Standard SFT at convergence. Lower is better.

Backbone	Metric	20%	40%	60%	80%	100%	Red.
LLaVA-OV1.5-8B	$\overline{D}$ (SFT)	0.0593	0.0532	0.0510	0.0479	0.0469	–
	$\overline{D}$ (LoMo)	0.0160	0.0144	0.0143	0.0140	**0.0139**	**70.36%**
	TV (SFT)	0.0514	0.0492	0.0489	0.0477	0.0472	–
	TV (LoMo)	0.0267	0.0257	0.0255	0.0253	**0.0250**	**47.03%**
Qwen3.5-9B	$\overline{D}$ (SFT)	0.1970	0.1391	0.1010	0.0964	0.0946	–
	$\overline{D}$ (LoMo)	0.0077	0.0072	0.0072	0.0066	**0.0065**	**93.18%**
	TV (SFT)	0.0721	0.0682	0.0677	0.0645	0.0597	–
	TV (LoMo)	0.0193	0.0186	0.0182	0.0173	**0.0170**	**71.53%**

C Additional Benchmark Results

We report results on two additional benchmarks, comparing Standard SFT and LoMo on both backbones under the same training setup as Table 1. VISTA-Bench [25] presents identical content in tokenized-text (Text) and visualized-text (VT) forms, and BLINK [10] targets vision-centric perception. As shown in Table 8, LoMo improves accuracy under both carriers on VISTA-Bench, narrows the Text-to-VT gap on both backbones, and yields consistent gains on BLINK.

Table 8: Results on VISTA-Bench and BLINK for Standard SFT and LoMo on both backbones. Text and VT denote the tokenized-text and visualized-text forms of VISTA-Bench, and Gap denotes the Text-to-VT accuracy drop.

Backbone	Setting	VISTA Text ↑	VISTA VT ↑	Gap ↓	BLINK ↑
LLaVA-OV1.5-8B	Standard SFT	47.26	31.73	15.53	40.17
	+ LoMo	**49.16**	**38.33**	**10.83**	**42.39**
Qwen3.5-9B	Standard SFT	56.98	50.58	6.40	55.70
	+ LoMo	**58.76**	**54.60**	**4.16**	**60.86**

D Implementation Details

D.1 Compute Resources

Image rendering for the LoMo data construction pipeline was performed on two CPU servers with 128 cores each, taking approximately 20 hours in total. All model training and evaluation experiments were conducted on a single node equipped with 8 NVIDIA H200 GPUs.

D.2 Training Data Construction

Our training corpus combines the original LLaVA-OneVision 1.5 instruct data with our LoMo-augmented rendered-text data. Specifically, we sample 2M multimodal and 2M text-only instances, with 50% of the latter rendered via LoMo. For a fair comparison, the standard SFT baseline is trained on the same 4M instances without any LoMo rendering. The rendered-text data is constructed by rendering text question prompts from the pure-text corpus into images. Rendering is performed with the Python-based text renderer with LaTeX support enabled. The plain text uses font size 20 with line height 22, while mathematical expressions are typeset in common Latin Modern Math at size 26. Rendered images are retained at their native resolution based on the length of text. The resulting rendered images are inserted back into multimodal SFT examples in place of the corresponding text spans.

D.3 Training Setup

We fine-tune LLaVA-OneVision-1.5-8B-Base and Qwen3.5-9B-Base using LLaMA-Factory under a standard supervised fine-tuning regime. We adopt a maximum sequence length of $32{,}768$ tokens, and a maximum image resolution of $2{,}560{,}000$ pixels. Training employs FlashAttention 2, bf16 precision and DeepSpeed ZeRO Stage 1, with liger kernels and sequence packing for throughput. The learning rate follows a cosine schedule from $4\mathrm{e}{-5}$ to $1\mathrm{e}{-6}$ with a warmup ratio of 0.002. We use a per-device batch size of 1 with 4 gradient accumulation steps. The language model and the multimodal projector are updated, while the vision tower remains frozen throughout training. We optimize with AdamW (β_1=0.9, β_2=0.99, weight decay 0.01) for one epoch.

D.4 Standard Evaluation Protocol

We evaluate all models with EvalScope under its standard protocol. The process of model prediction is deterministic with temperature 0 and a maximum context length of 32K tokens.

D.5 Rendered Evaluation Protocol

Under the rendered evaluation protocol, each textual question is rendered into an image using a python-based renderer and font configuration with font size 20 and line height 22, and is then used in place of the original text question. All other settings remain identical to the standard evaluation protocol.

E Limitations

While LoMo delivers consistent improvements, several aspects remain beyond the scope of this work. We apply LoMo only during the SFT stage, leaving its integration with pre-training or RL-based post-training to future exploration. Our span localization adopts a block-level middle-span heuristic; exploring more fine-grained strategies such as difficulty-aware or curriculum-style selection may yield further gains. Finally, due to compute constraints, we validate LoMo on two backbones at the 8B–9B scale, and verifying its behavior on substantially larger models is left to future work.

F Broader Impacts

LoMo strengthens cross-modal alignment in Vision-Language Models without modifying model architectures or introducing inference overhead. By treating text and images as semantically interchangeable carriers, it improves model robustness in real-world scenarios where information appears in mixed visual-textual forms, benefiting applications such as document understanding, assistive reading tools, and educational platforms that integrate handwritten or printed materials. As LoMo builds on pre-trained VLMs, it inherits the biases and failure modes of these foundation models, and we encourage practitioners to follow established VLM safety practices when deploying LoMo-trained models in high-stakes domains.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The main claims in the abstract and introduction (Section 1) are consistent with the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The cross-carrier consistency bound in Section 3.3 states its assumption (the two carriers share the same supervision target), and the complete derivation is provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The LoMo pipeline is described in Section 3, and the models, training data, benchmarks, evaluation protocols, and alignment metrics are described in Section 4. Full training hyperparameters, data-processing details, and evaluation configurations are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide reproduction details in the supplemental material and will release the code and LoMo data-construction pipeline upon publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training data, models, benchmarks, evaluation protocols, and metrics are described in Section 4.1, with full hyperparameters in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or significance tests due to compute cost, but evaluate across two backbones, 13 benchmarks, and multiple ablations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics and does not involve private data or human-subject experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new high-risk dataset, model, or deployed system.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing models, datasets, benchmarks, and tools used in this paper are open-source assets; we cite their original sources and follow their stated licenses and terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code and LoMo data-construction pipeline will be documented upon publication.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper studies and fine-tunes LLM-based VLMs as the core experimental subjects.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.