Unleashing the Intrinsic Visual Representation Capability of Multimodal Large Language Models

Supplementary Material

A. Additional Preliminaries

Masked Image Modeling. Masked Image Modeling (MIM) is a self-supervised learning paradigm wherein a model learns to reconstruct representations of masked image regions. Following iBOT [60], an input image $\boldsymbol{I}$ is partitioned into N patches and encoded via a vision encoder $\mathcal{G}_\xi$. A binary mask $\mathcal{M} \in \{0,1\}^N$ indicates which patches are masked. iBOT employs a teacher-student framework where the student predicts the teacher's representations at masked positions. The objective minimizes the cross-entropy between student predictions and teacher targets:

$$\mathcal{L}_{\text{MIM}} = -\sum_{i \in \mathcal{P}_\mathcal{M}} \text{softmax}(\hat{\boldsymbol{z}}_i/\tau_{\text{tea.}}) \cdot \log \text{softmax}(\tilde{\boldsymbol{z}}_i/\tau_{\text{stu.}}), \tag{1}$$

where $\mathcal{P}_\mathcal{M} = \{i \in \{1,\dots,N\} \mid \mathcal{M}_i = 1\}$ denotes the masked position indices, and $\tau_{\text{tea.}}, \tau_{\text{stu.}}$ are temperature parameters controlling the softmax sharpness for teacher and student distributions, respectively. We adopt $\tau_{\text{tea.}} = 0.04$ and $\tau_{\text{stu.}} = 0.1$ by default. This formulation enables the student to learn discriminative semantic representations through reconstruction of masked region features.

CKNNA. *Centered Kernel Nearest-Neighbor Alignment* (CKNNA) [20] is a metric for measuring feature alignment between models, derived as a relaxed variant of *Centered Kernel Alignment* (CKA) [23]. Given two feature sets, CKA quantifies their global similarity through kernel matrices as:

$$\text{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\text{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\text{HSIC}(\mathbf{K}, \mathbf{K})\text{HSIC}(\mathbf{L}, \mathbf{L})}}, \tag{2}$$

where $\mathbf{K}$ and $\mathbf{L}$ denote kernel matrices computed from the feature sets, and $\text{HSIC}(\cdot,\cdot)$ represents the Hilbert-Schmidt Independence Criterion measuring feature dependence. The kernel matrices are defined as $\mathbf{K}_{ij} = \kappa(\boldsymbol{k}_i, \boldsymbol{k}_j)$ and $\mathbf{L}_{ij} = \kappa(\boldsymbol{l}_i, \boldsymbol{l}_j)$, where $\kappa(\cdot,\cdot)$ is the kernel function and $\boldsymbol{k}_i$, $\boldsymbol{l}_i$ are feature vectors. Using the inner product kernel, HSIC is formulated as:

$$\text{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(N-1)^2}\Big(\sum_{i=1}^{N}\sum_{j=1}^{N}\big(\langle \boldsymbol{k}_i, \boldsymbol{k}_j\rangle - \mathbb{E}[\langle \boldsymbol{k}_i, \boldsymbol{k}_j\rangle]\big) \big(\langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle - \mathbb{E}[\langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle]\big)\Big). \tag{3}$$

CKNNA refines CKA by restricting alignment to nearest neighbors, replacing $\text{HSIC}(\cdot,\cdot)$ with $\text{HSIC}_{\text{kNN}}(\cdot,\cdot)$:

$$\text{HSIC}_{\text{kNN}}(\mathbf{K}, \mathbf{L}) = \frac{1}{(N-1)^2}\Big(\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{I}(i,j)\big(\langle \boldsymbol{k}_i, \boldsymbol{k}_j\rangle - \mathbb{E}[\langle \boldsymbol{k}_i, \boldsymbol{k}_j\rangle]\big) \big(\langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle - \mathbb{E}[\langle \boldsymbol{l}_i, \boldsymbol{l}_j\rangle]\big)\Big), \tag{4}$$

where $\mathbb{I}(i,j)$ is the nearest-neighbor indicator:

$$\mathbb{I}(i,j) = \begin{cases} 1 & \text{if } i \neq j, \boldsymbol{k}_i \in \text{kNN}(\boldsymbol{l}_j; k), \boldsymbol{l}_j \in \text{kNN}(\boldsymbol{k}_i; k), \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

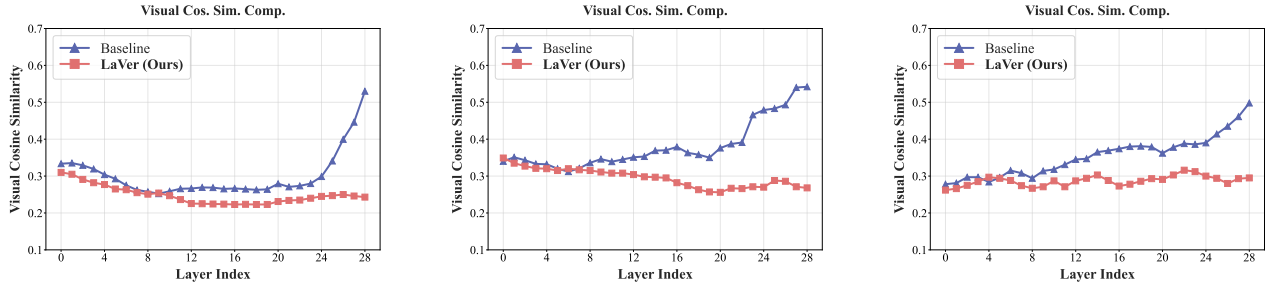
with $\text{kNN}(\boldsymbol{x}; k)$ denoting the k-nearest neighbors of $\boldsymbol{x}$. We set $k = 10$ by default.

B. Additional Discussion on Visual Feature Homogenization

MLLMs exhibit *modality imbalance* [6, 56, 58], systematically biasing toward textual information over visual inputs [21, 26, 34, 40, 56, 58], with more allocated attention scores and predictions predominantly grounded in text modality [40, 53].

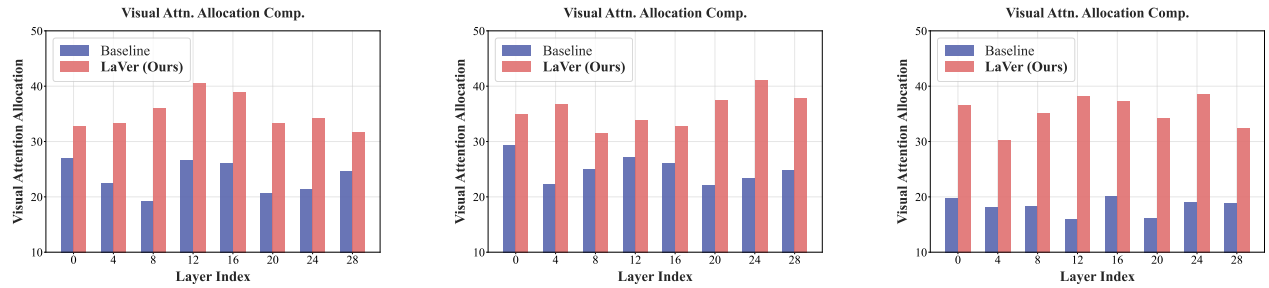
To further validate the *progressive visual feature homogenization* phenomenon illustrated in Fig. 2 of the main text, we conduct comprehensive empirical analyses across multiple vision encoders and evaluation metrics. This phenomenon reveals a critical insight: MLLMs progressively discard rich visual information throughout their layers, potentially retaining only those visual features directly relevant to textual reasoning tasks. We posit that this mechanism fundamentally stems from the *next-text-token-prediction* objective, which inherently emphasizes the text modality and provides only indirect, implicit supervisory signals for developing intrinsic visual modeling capabilities. While this text-centric training paradigm proves effective for tasks demanding sophisticated language generation competence, it is suboptimal for training MLLMs that should seamlessly integrate multimodal information without disproportionately favoring any particular modality [10, 48]. The limited performance of MLLMs on dense visual understanding tasks serves as compelling evidence of this text-modality bias.

We provide extensive empirical validation of the *progressive visual feature homogenization* phenomenon in Fig. 1, Fig. 2, and Fig. 3, examining diverse vision encoders including Qwen2.5-7B-Instruct [41] paired with SigLIP 2 [49], CLIP [42], and DINOv2 [39], demonstrating that



(a) Visual cosine similarity for SigLIP 2. (b) Visual cosine similarity for CLIP. (c) Visual cosine similarity for DINOv2.

Figure 1. **Averaged visual cosine similarity across the layers. (a)** shows the visual cosine similarity for SigLIP 2 [49]. **(b)** shows the visual cosine similarity for CLIP [42]. **(c)** shows the visual cosine similarity for DINOv2 [39].



(a) Visual attention allocation for SigLIP 2. (b) Visual attention allocation for CLIP. (c) Visual attention allocation for DINOv2.

Figure 2. **Averaged visual attention allocation across the layers. (a)** shows the visual attention allocation for SigLIP 2 [49]. **(b)** shows the visual attention allocation for CLIP [42]. **(c)** shows the visual attention allocation for DINOv2 [39].

this phenomenon persists across different architectural configurations. Specifically, we employ the CKNNA metric to quantify feature alignment between intermediate visual representations and input visual features, with detailed formulation provided in Sec. A. We construct visual feature sets using images from MMVP [48] and set $k = 10$ as the default neighborhood size.

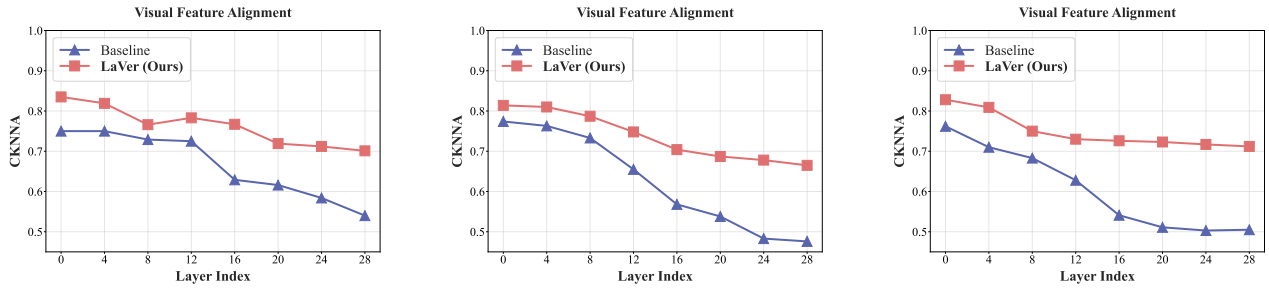
Our results reveal several critical observations. First, visual feature homogenization intensifies in deeper layers, as evidenced by the progressively increasing averaged visual feature-wise cosine similarity shown in Fig. 1. Second, the substantially diminished visual attention allocation illustrated in Fig. 2 demonstrates that models predominantly leverage information from text tokens, leaving abundant visual information severely underutilized. Notably, our empirical findings indicate that this underutilization of visual information persists consistently across all layers. Third, the gradually decreasing CKNNA metric depicted in Fig. 3 indicates progressive misalignment of intermediate visual features from their original representations. In stark contrast to the lower CKNNA similarity exhibited by baseline models, our proposed LaVer consistently maintains substantially higher CKNNA similarity scores. This demonstrates that LaVer effectively enables models to preserve rich visual information from vision encoders and cultivate robust intrinsic visual modeling capabilities.

The comprehensive evaluation across diverse benchmarks conclusively demonstrates that by introducing explicit vision-centric supervisory signals, models' multimodal capabilities can be significantly enhanced, particularly on tasks demanding dense visual information comprehension and fine-grained visual understanding.

C. Additional Discussion on Visual Feature Inconsistency

As demonstrated in Sec. 4.2, applying the Masked Image Modeling (MIM) objective in isolation paradoxically leads to increased visual cosine similarity, signaling severe visual information degradation rather than enhancement. Our analysis of training dynamics reveals a critical pattern: while visual cosine similarity initially decreases during early training iterations, it rapidly escalates in subsequent phases, ultimately surpassing baseline levels. This counterintuitive behavior arises because the MIM objective, while encouraging the model to reconstruct its own visual embeddings, fails to explicitly constrain the model from generating homogeneous visual features, i.e., a degenerate solution that minimizes MIM loss at the expense of preserving meaningful visual distinctions.

This phenomenon bears striking resemblance to observations reported in prior work [7, 39, 43], where models progressively generate visual features exhibiting high co-



(a) CKNNA for SigLIP 2. (b) CKNNA for CLIP. (c) CKNNA for DINOv2.

Figure 3. **CKNNA metrics across the layers.** **(a)** displays the CKNNA metric for SigLIP 2 [49]. **(b)** displays the CKNNA metric for CLIP [42]. **(c)** displays the CKNNA metric for DINOv2 [39].

Table 1. **Hyperparameters of training stages.**

Configuration	Stage 1	Stage 2	Stage 3
Trainable parameters	Connector $\mathcal{H}_\phi$	Connector $\mathcal{H}_\phi$, LLM $\mathcal{F}_\theta$, Vision Head $\mathcal{V}_\psi$	Connector $\mathcal{H}_\phi$, LLM $\mathcal{F}_\theta$
Frozen parameters	Vision Encoder $\mathcal{G}_\xi$, LLM $\mathcal{F}_\theta$	Vision Encoder $\mathcal{G}_\xi$, Teacher LLM $\mathcal{F}_{\hat{\theta}}$, Teacher Vision Head $\mathcal{V}_{\hat{\psi}}$	Vision Encoder $\mathcal{G}_\xi$
Global batch size	128	128	128
Batch size per GPU	4	2	4
Accumulation steps	2	4	2
Max sequence length	2048	2048	2048
DeepSpeed Zero Stage	2	2	2
Learning rate	2.0×10^{-3}	2.0×10^{-5}	1.0×10^{-5}
Learning rate schedule	Cosine	Cosine	Cosine
Warmup ratio	0.05	0.05	0.05
Weight decay	0	0	0
Training steps	4360	6250	6250
Data scale	558K	800K	800K
Optimizer	AdamW	AdamW	AdamW
β_1, β_2	0.9, 0.999	0.9, 0.999	0.9, 0.999
Precision	bf16	bf16	bf16

Table 2. **Hyperparameters of LaVer.**

Configuration	LaVer
Visual hidden dimension	8192
Loss coefficient ω_{MIM}	1.0
Loss coefficient ω_{CGA}	1.0
Teacher temperature $\tau_{\text{tea.}}$	0.04
Student temperature $\tau_{\text{stu.}}$	0.1
Masking ratio	0.05
Masking schedule	Cosine
EMA decay rate	0.95
EMA update steps	100
EMA schedule	Cosine

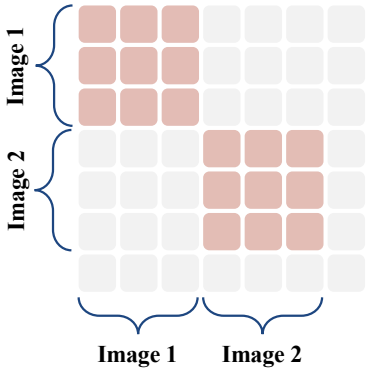
Table 3. **Hyperparameters of ReasonSeg.**

Configuration	ReasonSeg
Trainable parameters	Connector $\mathcal{H}_\phi$, LLM $\mathcal{F}_\theta$
Frozen parameters	Vision Encoder $\mathcal{G}_\xi$
Global batch size	128
Batch size per GPU	4
Accumulation steps	2
Max sequence length	2048
DeepSpeed Zero Stage	2
Learning rate	2.0×10^{-4}
Learning rate schedule	Cosine
Warmup ratio	0.01
Weight decay	0.01
Training steps	30000
Data scale	3.8M
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Precision	bf16

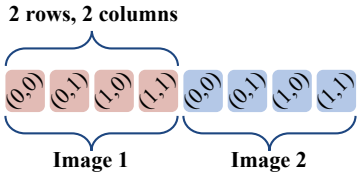
sine similarity with global semantic tokens while discarding fine-grained local structural information. Given the conceptual alignment between our observations under isolated MIM training and these established findings, we adopt the terminology *visual feature inconsistency* to characterize this pathology, wherein visual patches exhibit spuriously high cosine similarity despite encoding fundamentally distinct local visual information. While this phenomenon was attributed to the global semantic contrastive loss in the DINO series [43], our empirical findings reveal that local MIM objectives can independently induce the same degenerative

Table 4. **Computational cost comparison of stage 2.** LaVer introduces modest computational overhead compared to the baseline, with training time increases of 13-16% and memory consumption increases of 14-26% across different vision encoders. Despite these additional costs, the substantial performance improvements demonstrated in our experiments justify this overhead, making LaVer a practical and effective approach for enhancing MLLM capabilities.

Vision encoder	Method	Trainable Parameters	Training Speed	Training Time	Total GFLOPs	Memory Consumption
SigLIP 2	Baseline	7.63B	4.87 s/iter	8 h 27 min	2.46×10^{10}	55.24 GB
	LaVer	7.65B	5.51 s/iter	9 h 34 min	3.04×10^{10}	69.82 GB
CLIP	Baseline	7.63B	4.23 s/iter	7 h 42 min	2.24×10^{10}	53.24 GB
	LaVer	7.65B	4.67 s/iter	8 h 28 min	2.80×10^{10}	66.79 GB
DINOv2	Baseline	7.63B	3.98 s/iter	6 h 55 min	1.85×10^{10}	51.24 GB
	LaVer	7.65B	4.45 s/iter	7 h 44 min	2.19×10^{10}	58.24 GB



(a) Diagonally blocked full attention.



(b) Blocked 2D-RoPE.

Figure 4. **Illustration of packed visual sequences. (a)** illustrates diagonally blocked full attention for packed visual sequence. **(b)** illustrates blocked 2D-RoPE for packed visual sequence.

behavior, highlighting a new failure mode.

To address this challenge, the *Gram-Anchoring* mechanism was proposed by [43] to explicitly enforce preservation of spatial structural information while learning discriminative local embeddings for each patch. However, *Gram-Anchoring* exhibits a fundamental limitation: it symmetrically penalizes deviations in both directions, thereby inadvertently discouraging the emergence of discriminative visual features. Specifically, when the model attempts to produce more distinctive representations characterized by lower feature-wise cosine similarity, it incurs penalties equivalent to those for generating overly homogeneous features. While this symmetric regularization proves benign for vision-only models [43], it becomes problematic in the context of MLLMs, which inherently suffer from modality imbalance. The pre-existing bias toward textual modality creates a perverse incentive: the model can exploit this imbalance by generating nearly identical visual features to trivially minimize the MIM objective, effectively circumventing genuine visual understanding.

To overcome this limitation, we propose *Clipped Gram-Anchoring*, an asymmetric regularization strategy that selectively penalizes the model only when it tends to produce visual features with excessively high cosine similarity. By imposing penalties exclusively on the homogenization direction while permitting the model to freely explore more discriminative feature spaces, this regularizer effectively prevents visual feature inconsistency. This design aligns with the fundamental requirement for MLLMs, i.e., maintaining rich, discriminative visual representations that can meaningfully contribute to multimodal reasoning, rather than collapsing into degenerate solutions that superficially satisfy training objectives while sacrificing genuine visual understanding capabilities.

Visual Feature Homogenization v.s. Visual Feature Inconsistency. To elucidate the fundamental distinctions between these two phenomena, we provide a justification of their divergent characteristics. While both *visual feature homogenization* and *visual feature inconsistency* manifest the same statistical signature, namely, the generation of visual features exhibiting elevated feature-wise cosine similarity, their underlying causal mechanisms and downstream implications differ substantially.

Visual feature homogenization emerges as a direct consequence of the modality imbalance inherent in MLLMs' training paradigm. Under the text-centric next-token-prediction objective, models systematically exploit textual tokens to generate responses while marginalizing the majority of available visual information. This preferential reliance on textual modality precipitates a progressive degradation of visual representations, manifesting as increased homogeneity among visual features across layers. Critically, this phenomenon reflects a fundamental loss of discriminative visual information rather than a mere representational artifact.

In contrast, *visual feature inconsistency* originates from the explicit application of the MIM objective in isolation.

Table 5. **Scalability of LaVer on model parameters.** LaVer demonstrates strong model scaling properties, consistently improving performance across different parameter sizes (1.5B, 3B, and 7B) with both SigLIP 2 and CLIP vision encoders.

Benchmark	SigLIP2									CLIP								
	1.5B			3B			7B			1.5B			3B			7B		
	Baseline	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}
GQA	45.98	**48.51**	↑2.53	48.99	**53.02**	↑4.03	55.03	**56.78**	↑1.75	46.98	**48.25**	↑1.27	**50.00**	48.99	↓1.01	51.51	**54.77**	↑3.26
MMB^{EN}	61.89	**65.51**	↑3.62	70.79	**74.23**	↑3.44	73.97	**75.60**	↑1.63	54.33	**60.22**	↑5.89	62.80	**63.23**	↑0.43	68.64	**69.93**	↑1.29
SEED^{I}	58.65	**63.04**	↑4.39	64.45	**66.10**	↑1.65	67.57	**68.62**	↑1.05	56.18	**62.22**	↑6.04	62.63	**62.67**	↑0.04	64.36	**65.20**	↑0.84
MME	1261.35	**1285.18**	↑23.83	1384.78	**1445.33**	↑60.55	1510.73	**1512.50**	↑1.77	1175.37	**1213.01**	↑37.64	1260.48	**1379.26**	↑118.78	1289.62	**1474.65**	↑185.03
RWQA	**52.68**	52.29	↓0.39	56.08	**58.56**	↑2.48	53.86	**59.35**	↑5.49	50.85	**52.16**	↑1.31	**56.73**	55.29	↓1.44	54.25	**56.47**	↑2.22
MMMU	40.11	**41.11**	↑1.00	41.11	**42.00**	↑0.89	44.78	**46.33**	↑1.55	38.33	**39.44**	↑1.11	**42.44**	40.67	↓1.77	**44.56**	**44.56**	↑0.00
MM*	38.02	**40.90**	↑2.88	46.80	**50.07**	↑3.27	49.06	**52.01**	↑2.95	37.42	**40.83**	↑3.41	41.97	**44.05**	↑2.08	43.17	**45.45**	↑2.28
OCRB	237	**258**	↑21	353	**397**	↑44	536	**639**	↑103	136	**162**	↑26	226	**265**	↑39	306	**365**	↑59
TVQA	42.42	**44.96**	↑2.54	51.33	**55.78**	↑4.45	62.06	**63.93**	↑1.87	31.70	**34.65**	↑2.95	40.85	**48.49**	↑7.64	43.86	**49.93**	↑6.07
CQA	27.84	**31.52**	↑3.68	40.00	**42.32**	↑2.32	43.52	**50.24**	↑6.72	18.40	**19.04**	↑0.64	26.00	**30.52**	↑4.52	27.36	**39.36**	↑12.00
AI2D	59.94	**62.56**	↑2.62	72.05	**75.39**	↑3.34	73.74	**75.55**	↑1.81	59.39	**60.46**	↑1.07	64.73	**65.54**	↑0.81	66.00	**70.40**	↑4.40
MMVP	60.67	**63.67**	↑3.00	**65.00**	64.00	↓1.00	69.00	**70.33**	↑1.33	52.00	**54.67**	↑2.67	55.00	**59.00**	↑4.00	60.00	**64.00**	↑4.00
CV-B^{2D}	52.69	**53.13**	↑0.44	58.69	**62.80**	↑4.11	67.87	**69.82**	↑1.95	51.95	**60.50**	↑8.55	57.93	**61.17**	↑3.24	64.39	**65.58**	↑1.19
SQA	72.78	**76.45**	↑3.67	83.75	**85.09**	↑1.34	86.51	**89.09**	↑2.58	69.36	**76.00**	↑6.64	72.73	**73.18**	↑0.45	81.61	**83.30**	↑1.69
MathV	35.90	**43.10**	↑7.20	46.20	**50.80**	↑4.60	52.20	**55.60**	↑3.40	34.50	**41.30**	↑6.80	41.60	**43.30**	↑1.70	45.40	**48.90**	↑3.50
Hallu	50.26	**52.37**	↑2.11	53.99	**54.68**	↑0.69	56.05	**58.04**	↑1.99	46.37	**48.37**	↑2.00	**53.89**	51.84	↓2.05	53.42	**55.95**	↑2.53
POPE	86.87	**88.67**	↑1.80	**86.03**	85.53	↓0.50	90.90	**91.23**	↑0.33	86.03	**88.60**	↑2.57	87.17	**90.70**	↑3.53	**90.50**	90.30	↓0.20
Average	46.32	**48.74**	↑2.42	52.13	**54.20**	↑2.07	55.72	**57.87**	↑2.15	43.20	**46.32**	↑3.12	48.07	**49.38**	↑1.31	50.58	**53.24**	↑2.66

Table 6. **Scalability of LaVer on datasets.** LaVer demonstrates strong data scaling properties, consistently improving performance across different training dataset sizes.

Benchmark	SigLIP 2									CLIP								
	800K			2M			4M			800K			2M			4M		
	Baseline	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}	**Baseline**	**LaVer**	Δ_{Baseline}
GQA	55.03	**56.78**	↑1.75	53.52	**58.29**	↑4.77	57.54	**58.79**	↑1.25	51.51	**54.77**	↑3.26	53.27	**55.53**	↑2.26	51.01	**53.52**	↑2.51
MMB^{EN}	73.97	**75.60**	↑1.63	74.83	**75.69**	↑0.86	73.37	**76.20**	↑2.83	68.64	**69.93**	↑1.29	68.81	**73.02**	↑4.21	69.50	**76.12**	↑6.62
SEED^{I}	67.57	**68.62**	↑1.05	67.11	**68.15**	↑1.04	67.61	**68.02**	↑0.41	64.36	**65.20**	↑0.84	63.66	**65.70**	↑2.04	64.58	**66.97**	↑2.39
MME	1510.73	**1512.50**	↑1.77	1549.66	**1589.66**	↑40.00	1574.33	**1635.16**	↑60.83	1289.62	**1474.65**	↑185.03	1312.79	**1390.06**	↑77.27	1392.78	**1461.37**	↑68.59
RWQA	53.86	**59.35**	↑5.49	60.65	**63.53**	↑2.88	59.61	**63.92**	↑4.31	54.25	**56.47**	↑2.22	54.63	**58.04**	↑3.41	53.86	**59.22**	↑5.36
MMMU	44.78	**46.33**	↑1.55	43.56	**45.56**	↑2.00	42.89	**45.79**	↑2.90	**44.56**	**44.56**	↑0.00	43.11	**44.44**	↑1.33	41.00	**42.44**	↑1.44
MM*	49.06	**52.01**	↑2.95	50.67	**50.74**	↑0.07	51.54	**52.54**	↑1.00	43.17	**45.45**	↑2.28	46.79	**48.85**	↑2.06	44.85	**50.61**	↑5.76
OCRB	536	**639**	↑103	637	**678**	↑41	665	**684**	↑19	306	**365**	↑59	355	**387**	↑32	351	**400**	↑49
TVQA	62.06	**63.93**	↑1.87	64.93	**68.65**	↑3.72	67.71	**69.11**	↑1.40	43.86	**49.93**	↑6.07	45.65	**54.26**	↑8.61	52.39	**57.73**	↑5.34
CQA	43.52	**50.24**	↑6.72	58.88	**64.48**	↑5.60	61.44	**64.88**	↑3.44	27.36	**39.36**	↑12.00	38.24	**45.20**	↑6.96	42.84	**51.04**	↑8.20
AI2D	73.74	**75.55**	↑1.81	73.44	**75.22**	↑1.78	73.15	**75.06**	↑1.91	66.00	**70.40**	↑4.40	66.65	**70.98**	↑4.33	67.97	**73.39**	↑5.42
MMVP	69.00	**70.33**	↑1.33	67.00	**71.33**	↑4.33	70.67	**72.33**	↑1.66	60.00	**64.00**	↑4.00	62.67	**66.33**	↑3.66	64.00	**69.00**	↑5.00
CV-B^{2D}	67.87	**69.82**	↑1.95	71.70	**71.77**	↑0.07	72.11	**73.35**	↑1.24	64.39	**65.58**	↑1.19	61.27	**64.67**	↑3.40	62.94	**67.66**	↑4.72
SQA	86.51	**89.09**	↑2.58	81.95	**86.47**	↑4.52	84.28	**87.01**	↑2.73	81.61	**83.30**	↑1.69	83.89	**84.83**	↑0.94	84.79	**90.88**	↑6.09
MathV	52.20	**55.60**	↑3.40	53.10	**56.70**	↑3.60	53.50	**56.70**	↑3.20	45.40	**48.90**	↑3.50	47.00	**51.70**	↑4.70	49.90	**56.70**	↑6.80
Hallu	56.05	**58.04**	↑1.99	60.15	**61.41**	↑1.26	58.36	**60.99**	↑2.63	53.42	**55.95**	↑2.53	56.20	**57.41**	↑1.21	**57.52**	56.89	↓0.63
POPE	90.90	**91.23**	↑0.33	91.33	**91.47**	↑0.14	**92.23**	91.93	↓0.30	**90.50**	90.30	↓0.20	88.47	**90.97**	↑2.50	**90.13**	89.07	↓1.06
Average	55.72	**57.87**	↑2.15	57.30	**59.46**	↑2.16	58.08	**59.88**	↑1.80	50.58	**53.24**	↑2.66	51.84	**54.88**	↑3.04	52.84	**56.60**	↑3.77

While MIM provides direct supervisory signals that guide the evolution of visual embeddings, it simultaneously introduces an exploitable optimization shortcut: the model can trivially minimize MIM loss by generating nearly identical visual features, thereby achieving low reconstruction error without preserving meaningful visual distinctions. Consequently, the isolated MIM objective exacerbates rather than ameliorates the situation, failing to provide balanced supervisory signals across modalities and inadvertently reinforcing the collapse toward homogeneous representations.

Our proposed LaVer framework addresses both pathologies synergistically. By integrating the Clipped Gram-Anchoring mechanism with the MIM objective, LaVer effectively prevents visual feature inconsistency through asymmetric regularization that selectively penalizes excessive homogenization while permitting discriminative feature learning. Simultaneously, by introducing explicit vision-centric supervisory signals, LaVer mitigates the underlying modality imbalance issue, enabling models to maintain rich, discriminative visual representations throughout their layers. This dual-pronged approach culminates in substantially enhanced performance across diverse multimodal benchmarks, particularly on tasks demanding dense visual understanding.

D. Implementation Details

Hyperparameters. We summarize the hyperparameters for each training stage in Table 1. Following common practice in LLaVA-OneVision 1.5 [2], we adopt standard configurations for our three-stage training pipeline. Note that we do not use the exact datasets from LLaVA-OneVision 1.5, as they were not fully available at the time of our experiments. LaVer is applied exclusively to Stage 2, where visual knowledge is injected into the model. The specific hyperparameters of LaVer are detailed in Table 2. We observe that LaVer's performance is robust to most hyperparameter choices, requiring minimal tuning. Through comprehensive ablation studies on masking strategies and EMA up-

Table 7. **Ablation study on masking strategies.** LaVer demonstrates strong robustness across different masking strategies.

Benchmark	SigLIP 2									CLIP								
	Baseline	Cosine				Constant				Baseline	Cosine				Constant			
		0.05	0.1	0.2	0.3	0.0002	0.01	0.05	0.1		0.05	0.1	0.2	0.3	0.0002	0.01	0.05	0.1
GQA	55.03	**56.78**	56.53	56.26	54.87	54.83	55.96	55.22	56.11	51.51	**54.77**	54.71	54.52	52.76	52.26	53.02	52.26	51.01
MMB^{EN}	73.97	75.60	**75.65**	75.04	73.99	74.89	73.74	74.46	73.80	68.64	69.93	69.24	70.62	71.39	65.38	70.10	**72.77**	70.27
SEED^{I}	67.57	**68.62**	68.50	68.49	67.48	67.82	68.42	68.34	67.96	64.36	**65.20**	65.02	64.27	64.99	64.67	64.21	64.81	65.03
MME	1510.73	1512.50	**1546.12**	1515.43	1534.31	1504.69	1496.10	1532.54	1495.53	1289.62	**1474.65**	1470.28	1407.86	1404.61	1421.08	1381.88	1399.67	1441.57
RWQA	53.86	**59.35**	58.77	58.62	55.51	58.33	55.18	54.55	56.17	54.25	**56.47**	55.59	55.56	54.12	52.42	54.51	55.56	52.94
MMMU	44.78	**46.33**	46.31	45.96	44.63	44.85	45.77	44.57	45.95	44.56	44.56	**45.44**	44.11	43.22	41.33	44.89	43.89	43.89
MM*	49.06	52.01	**52.04**	51.44	50.12	51.31	48.97	50.71	50.95	43.17	45.45	45.31	45.25	**47.39**	42.30	45.11	45.72	46.12
OCRB	536	**639**	638	625	582	619	606	607	597	306	**365**	354	345	301	292	302	300	297
TVQA	62.06	63.93	**64.11**	63.98	62.07	61.79	63.59	62.37	63.92	43.86	**49.93**	49.28	48.55	48.78	49.03	48.07	49.05	47.47
CQA	43.52	**50.24**	46.20	48.49	48.94	46.66	46.52	46.21	48.48	27.36	**39.36**	36.32	34.00	34.24	33.52	33.20	33.36	33.04
AI2D	73.74	75.55	**75.86**	74.68	73.61	74.37	73.65	74.35	73.70	66.00	**70.40**	69.46	68.62	69.59	65.38	70.14	69.95	69.24
MMVP	69.00	**70.33**	**70.33**	69.33	69.00	70.00	69.67	69.33	68.33	60.00	**64.00**	63.67	63.67	62.33	59.33	63.67	61.33	61.33
CV-B^{2D}	67.87	69.82	68.96	**70.06**	67.67	68.87	68.68	67.97	68.40	64.39	**65.58**	65.42	64.26	62.17	60.36	62.38	61.20	62.10
SQA	86.51	**89.09**	86.11	88.78	88.17	87.09	86.62	86.65	87.45	81.61	**83.30**	83.29	82.15	82.90	80.47	82.40	82.60	82.30
MathV	52.20	55.60	52.16	**55.82**	54.32	52.38	51.98	54.26	54.16	45.40	**48.90**	46.20	45.90	47.70	46.10	46.90	47.10	46.80
Hallu	56.05	58.04	58.29	**58.33**	56.27	57.72	56.89	57.45	56.61	53.42	**55.95**	**55.95**	55.10	51.00	54.26	54.00	54.89	55.63
POPE	90.90	91.23	**91.24**	90.73	90.51	90.55	90.58	90.49	90.57	90.50	90.30	**90.93**	90.10	90.43	89.43	89.26	90.00	89.27
Average	55.72	**57.87**	57.20	57.49	56.38	56.63	56.32	56.36	56.69	50.58	**53.24**	52.75	52.21	51.99	50.42	51.93	52.08	51.61

Table 8. **Ablation study on EMA teacher strategies.**

Method	SigLIP 2											CLIP										
	Baseline	LaVer										Baseline	LaVer									
EMA Stra.	-	Con.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	-	Con.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.	Cos.
EMA Frq.	-	100	100	1	50	100	200	100	100	100	100	-	100	100	1	50	100	200	100	100	100	100
EMA Ratio	-	0.95	0.95	0.95	0.95	0.95	0.95	0.9	0.95	0.99	0.999	-	0.95	0.95	0.95	0.95	0.95	0.95	0.9	0.95	0.99	0.999
GQA	55.03	55.63	56.78	53.65	55.22	56.78	**57.19**	56.44	56.78	55.90	56.07	51.51	51.72	54.77	50.91	51.39	54.77	**55.03**	54.09	54.77	54.00	53.97
MMB^{EN}	73.97	**77.39**	75.60	70.44	73.70	75.60	73.43	75.21	75.60	75.12	75.70	68.64	70.99	69.93	65.54	70.56	69.93	70.58	71.19	69.93	**71.22**	69.57
SEED^{I}	67.57	67.82	68.62	64.39	69.43	68.62	69.49	**69.57**	68.62	69.16	69.36	64.36	66.71	65.20	61.54	64.43	65.20	65.96	63.81	65.20	66.75	**66.92**
MME	1510.73	1546.52	1512.50	1436.02	**1574.55**	1512.50	1521.53	1544.95	1512.50	1501.03	1507.67	1289.62	1393.95	1474.65	1361.29	1315.01	1474.65	1366.78	1328.63	1474.65	1462.85	**1481.98**
RWQA	53.86	55.56	**59.35**	55.31	53.34	**59.35**	56.42	55.90	**59.35**	56.46	55.91	54.25	55.10	56.47	53.34	54.79	56.47	54.43	56.01	56.47	**56.70**	55.35
MMMU	44.78	45.48	**46.33**	43.21	45.85	**46.33**	45.18	44.81	**46.33**	44.69	44.57	44.56	44.96	44.56	42.33	44.24	44.56	45.43	**45.86**	44.56	44.89	45.02
MM*	49.06	50.49	**52.01**	49.32	48.84	**52.01**	50.35	49.36	**52.01**	51.74	51.93	43.17	44.80	45.45	42.18	45.17	45.45	43.57	45.79	45.45	46.23	**46.50**
OCRB	536	**651**	639	537	633	639	637	577	639	649	565	306	351	**365**	298	355	**365**	357	**365**	**365**	337	362
TVQA	62.06	**65.79**	63.93	59.38	62.33	63.93	65.27	62.29	63.93	63.97	65.08	43.86	45.39	**49.93**	46.66	47.42	**49.93**	45.40	46.76	**49.93**	46.04	48.64
CQA	43.52	47.00	**50.24**	44.98	45.54	**50.24**	49.01	47.87	**50.24**	49.66	47.48	27.36	**40.53**	39.36	30.08	36.81	39.36	33.66	37.27	39.36	37.22	40.24
AI2D	73.74	75.68	75.55	70.87	75.36	75.55	**75.76**	74.53	75.55	75.01	73.38	66.00	65.59	**70.40**	63.35	65.96	**70.40**	67.99	68.59	**70.40**	69.86	69.72
MMVP	69.00	**72.67**	70.33	66.00	71.00	70.33	72.00	71.33	70.33	72.33	70.67	60.00	62.00	64.00	59.67	**64.67**	64.00	63.00	62.33	64.00	62.00	63.33
CV-B^{2D}	67.87	69.08	69.82	66.10	**70.44**	69.82	70.05	67.29	69.82	69.30	69.58	64.39	67.42	65.58	61.70	65.54	65.58	67.09	65.45	65.58	**67.52**	66.21
SQA	86.51	89.59	89.09	82.90	88.33	89.09	87.24	**90.41**	89.09	86.34	86.86	81.61	85.17	83.30	77.91	84.23	83.30	**85.23**	82.95	83.30	82.39	82.25
MathV	52.20	**55.73**	55.60	52.46	55.06	55.60	54.23	53.39	55.60	55.68	52.74	45.40	**50.01**	48.90	45.75	49.20	48.90	47.51	46.63	48.90	47.86	48.47
Hallu	56.05	56.72	58.04	54.29	57.15	58.04	56.55	**59.46**	58.04	59.13	56.13	53.42	**57.28**	55.95	52.25	55.32	55.95	56.37	53.10	55.95	54.98	56.32
POPE	90.90	90.44	91.23	86.37	90.21	91.23	91.64	91.61	91.23	90.07	**91.97**	90.50	90.96	90.30	85.97	91.77	90.30	**92.56**	90.03	90.30	91.09	91.68
Average	55.72	57.43	**57.87**	54.17	56.65	**57.87**	57.36	57.10	**57.87**	57.40	56.98	50.58	52.92	53.24	49.41	52.49	53.24	52.63	52.40	53.24	52.92	**53.25**

dating strategies, we demonstrate LaVer's robustness to hyperparameter variations. Our findings indicate that a small masking ratio with cosine scheduling is sufficient for effective learning. For EMA updates, we find that the strategy is insensitive to the update schedule as long as the update frequency is not excessively high; overly frequent updates can cause the model to exploit the MIM loss by inadvertently propagating visual feature inconsistencies into the teacher model. For the vision head architecture, we employ a lightweight 3-layer MLP with 8192 hidden dimensions, parallel to the language head. This design introduces only a negligible number of additional trainable parameters. Following the established practice in [60], we set the teacher temperature $\tau_{\text{tea.}} = 0.04$ and student temperature $\tau_{\text{stu.}} = 0.1$, which ensures stable convergence throughout training. For fair comparison with ROSS [51], we adopt their 2-stage training protocol with identical configurations as specified in [51]. All experiments are conducted on 16 NVIDIA A100 GPUs with 80GB memory each.

Datasets. Our training pipeline employs off-the-shelf datasets across three stages. For Stage 1, we adopt the LLaVA-558K dataset [31] for vision-language alignment. For Stage 2, we randomly sample 800K image-text pairs from FineVision 23M [52], maintaining the original dataset's proportions to preserve its diverse visual knowledge sources. Due to computational constraints, we do not utilize the complete dataset; however, we validate LaVer's data scaling properties by sampling up to 4M pairs from FineVision. For Stage 3, we randomly sample 800K instruction-tuning pairs from LLaVA-OneVision 4M [28], again preserving the original proportions. While potential overlap may exist between Stage 2 and Stage 3 data, we retain all samples as this configuration has proven empirically effective. For fair comparison with ROSS [51], we adopt their 2-stage protocol using LLaVA-558K [31] for Stage 1 and Cambrian-737K [47] for Stage 2.

Vision encoders. We evaluate LaVer across diverse vision encoders to demonstrate its broad applicability. **SigLIP 2** [49] employs pairwise sigmoid loss instead of softmax for enhanced image-text alignment and multilingual capabilities. For our main experiments, we adopt SigLIP 2-ViT-SO400M/14@384, which encodes 384×384 images into 729 vision tokens. For comparison with ROSS [51], we use SigLIP-ViT-SO400M/14@384 [55] to

Table 9. **Ablation study on spatial awareness.**

$\mathcal{G}_\xi$	LaVer	Mixed Attn.	2D-RoPE	GQA	MMB^{EN}	SEED^{I}	MME	RWQA	MMMU	MM*	OCRB	TVQA	CQA	AI2D	MMVP	CV-B^{2D}	SQA	MathV	Hallu	POPE	Avg.
SigLIP 2	✗	✗	✗	55.03	73.97	67.57	1510.73	53.86	44.78	49.06	536	62.06	43.52	73.74	69.00	67.87	86.51	52.20	56.05	90.90	55.72
	✗	✓	✗	54.33	74.98	68.31	1505.55	55.17	45.32	49.54	**641**	61.40	45.38	75.94	**71.00**	**70.62**	88.62	**56.05**	55.83	**91.57**	56.78
	✗	✗	✓	54.87	73.69	68.45	1509.39	52.90	43.99	48.89	541	61.72	43.98	73.97	68.33	69.14	86.41	51.46	56.17	89.56	55.57
	✗	✓	✓	54.07	74.02	66.82	1481.78	58.01	44.73	51.11	605	62.13	46.71	**76.18**	69.00	66.81	87.81	53.20	57.30	90.20	56.43
	✓	✓	✓	**56.78**	**75.60**	**68.62**	**1512.50**	**59.35**	**46.33**	**52.01**	639	**63.93**	**50.24**	75.55	70.33	69.82	**89.09**	55.60	**58.04**	91.23	**57.87**
CLIP	✗	✗	✗	51.51	68.64	64.36	1289.62	54.25	**44.56**	43.17	306	43.86	27.36	66.00	60.00	64.39	81.61	45.40	53.42	90.50	50.58
	✗	✓	✗	51.71	69.00	63.84	1335.01	53.71	41.72	44.14	310	48.69	34.99	69.93	62.00	63.38	80.77	45.82	53.36	89.82	51.40
	✗	✗	✓	51.54	68.98	64.98	1302.20	54.04	42.02	42.96	310	44.33	27.44	65.37	61.67	64.69	82.33	45.37	53.46	90.05	50.59
	✗	✓	✓	52.76	**69.99**	65.15	1327.62	53.59	42.89	44.78	313	48.77	35.76	69.75	63.00	64.46	82.20	45.50	53.73	**90.66**	51.99
	✓	✓	✓	**54.77**	69.93	**65.20**	**1474.65**	**56.47**	**44.56**	**45.45**	**365**	**49.93**	**39.36**	**70.40**	**64.00**	**65.58**	**83.30**	**48.90**	**55.95**	90.30	**53.24**

Table 10. **Ablation study on loss functions.**

$\mathcal{G}_\xi$	$+\mathcal{L}_{\text{MIM}}$	$+\mathcal{L}_{\text{GA}}$	$+\mathcal{L}_{\text{CGA}}$	GQA	MMB^{EN}	SEED^{I}	MME	RWQA	MMMU	MM*	OCRB	TVQA	CQA	AI2D	MMVP	CV-B^{2D}	SQA	MathV	Hallu	POPE	Avg.
SigLIP 2	✗	✗	✗	55.03	73.97	67.57	1510.73	53.86	44.78	49.06	536	62.06	43.52	73.74	69.00	67.87	86.51	52.20	56.05	90.90	55.72
	✓	✗	✗	55.02	71.97	64.59	1500.80	51.38	44.35	47.41	512	60.40	42.20	70.35	66.00	64.22	82.57	52.72	54.11	85.53	53.76
	✓	✓	✗	54.65	73.56	65.84	1456.21	58.97	45.21	**52.09**	617	63.07	49.07	72.26	67.67	68.94	85.93	**55.89**	55.91	89.49	56.46
	✓	✗	✓	**56.78**	**75.60**	**68.62**	**1512.50**	**59.35**	**46.33**	52.01	**639**	**63.93**	**50.24**	**75.55**	**70.33**	**69.82**	**89.09**	55.60	**58.04**	**91.23**	**57.87**
CLIP	✗	✗	✗	51.51	68.64	64.36	1289.62	54.25	**44.56**	43.17	306	43.86	27.36	66.00	60.00	64.39	81.61	45.40	53.42	**90.50**	50.58
	✓	✗	✗	48.48	66.08	63.93	1222.30	52.14	42.42	43.24	290	43.89	26.52	64.28	60.67	61.27	83.16	45.04	54.44	88.77	49.71
	✓	✓	✗	54.72	68.16	63.26	1431.67	53.80	43.78	44.78	359	49.14	37.83	**70.67**	**64.67**	**65.68**	79.36	46.46	53.91	86.98	52.01
	✓	✗	✓	**54.77**	**69.93**	**65.20**	**1474.65**	**56.47**	**44.56**	**45.45**	**365**	**49.93**	**39.36**	70.40	64.00	65.58	**83.30**	**48.90**	**55.95**	90.30	**53.24**

ensure fair evaluation. **CLIP** [42] is trained with contrastive loss for vision-language alignment. We adopt CLIP-ViT-L/14@336, which processes 336×336 images into 576 vision tokens. **DINOv2** [43] leverages self-contrastive and self-distillation learning for visual feature extraction. We use DINOv2-Large/14@224, encoding 224×224 images into 384 vision tokens. **AIMv2** [15] is a native-resolution encoder trained via autoregressive pixel-wise prediction with an auxiliary LLM backbone. We adopt AIMv2-Large/14, which patchifies images into 14×14 patches, with images resized to a maximum of 224×224 pixels. **Qwen-ViT** [3] serves as the native-resolution vision encoder in Qwen2.5-VL with patch size 14. We utilize the encoder from Qwen2.5-VL-7B-Instruct and set the maximum resolution to 512×512 pixels. **Encoder-free architecture.** To validate LaVer's generalizability beyond traditional vision encoders, we adopt an encoder-free architecture [11, 13, 25]. We employ a 3-layer MLP with 3584 intermediate hidden dimensions to project images into visual tokens with patch size 16, supporting up to 512×512 pixels. During Stage 1, only the MLP is trained; in Stages 2 and 3, both the MLP projector and LLM backbone are jointly optimized.

Packed visual sequence. To improve training efficiency while maintaining independent visual reconstruction without interfering with multimodal sequence modeling, we pack vision tokens from multiple images into a single sequence, excluding their corresponding text tokens. Specifically, we construct diagonally blocked bidirectional attention and concatenated 2D-RoPE for the packed visual sequences. Across all vision encoders, we pack vision tokens from 2 images into a single sequence of length 2048 with padding tokens, yielding two separate visual sequences per local batch on each GPU with batch size 4. The attention and positional embedding mechanisms for packed visual sequences are illustrated in Fig. 4.

Evaluation. We conduct comprehensive evaluation using the VLMEvalKit [14] toolbox. Our evaluation suite encompasses a diverse set of benchmarks: GQA [19] for compositional visual reasoning, MMBench (MMB^{EN}) [32] for comprehensive multimodal understanding, SEED-Image (SEED^{I}) [27] for generative comprehension, MME [16] for perception and cognition evaluation, RealWorldQA (RWQA) [12] for real-world visual question answering, MMMU [54] for massive multi-discipline understanding, MMStar (MM*) [8] for challenging multi-modal reasoning, OCR-Bench (OCRB) [33] for text recognition capabilities, TextVQA (TVQA) [44] for reading text in images, ChartQA (CQA) [37] for chart understanding, AI2D [22] for diagram comprehension, CV-Bench-2D (CV-B^{2D}) [47] for vision-centric capabilities, MMVP [48] for visual perception, ScienceQA (SQA) [35] for science question answering with explanations, MathVista (MathV) [36] for mathematical reasoning in visual contexts, HallucinationBench (Hallu) [17] for hallucination detection, and POPE [29] for object hallucination evaluation. We employ the default system prompt and instruction template for each benchmark to ensure fair comparison. For averaged results, we compute the mean value across all benchmarks, with MME and OCR-Bench normalized to the range $[0.0, 1.0]$ to maintain consistent scaling.

For ReasonSeg [24] evaluation, we adopt the fine-tuning protocol established in [46]. Specifically, we initialize models using checkpoints from stage 2 for both the baseline and LaVer configurations. Following [46], we fine-tune the models to acquire sophisticated segmentation reasoning capabilities using a comprehensive dataset mixture comprising semantic segmentation datasets (COCOStuff [4],

Table 11. **Language performance comparison.**

Benchmark	SigLIP 2			CLIP		
	Baseline	LaVer	Δ_{Baseline}	Baseline	LaVer	Δ_{Baseline}
IFEval	70.34	**70.77**	↑0.43	**68.96**	68.24	↓0.72
MMLU	56.52	**57.48**	↑0.96	58.34	**58.67**	↑0.33
BBH	**33.25**	32.96	↓0.29	34.78	**34.93**	↑0.15
Average	53.37	**53.74**	↑0.37	**54.03**	53.95	↓0.08

Mapillary [38], ADE20K [59], nuScenes [5]), referring segmentation datasets (RefCOCO [5], RefCOCO+ [5], RefCOCOg [5]), and the general VQA dataset LLaVA-665K [30], totaling approximately 4M samples. The detailed training configuration is presented in Table 3, with hyperparameters primarily adopted from [46].

Computational overhead analysis. The implementation of LaVer introduces additional computational overhead, primarily stemming from the forward passes required for both student and teacher models to process image tokens, as well as the EMA updates for the teacher model parameters. As detailed in Table 4, we conduct a comprehensive computational cost analysis focusing on stage 2, where LaVer is applied (stages 1 and 3 remain identical to the baseline). Across different vision encoders, LaVer incurs training time increases of 13-16% and memory consumption increases of 14-26% compared to the baseline. Specifically, with SigLIP 2 [49], training time increases from 8h 27min to 9h 34min, while memory consumption rises from 55.24 GB to 69.82 GB per GPU (averaged across 16 GPUs). Similar patterns are observed with CLIP [42] and DINOv2 [39] vision encoders. Despite these additional costs, the substantial performance improvements demonstrated across diverse benchmarks and architectural configurations justify this computational overhead, establishing LaVer as a practical and effective approach for enhancing MLLMs.

E. Visualization details

we provide the details about how the figures are generated in the paper.

Fig. 1. We present a comprehensive performance comparison between LaVer and the baseline across 17 benchmarks, utilizing SigLIP 2 [49] as the vision encoder and Qwen2.5-7B-Instruct [41] as the LLM backbone.

Fig. 2. **(a)** We randomly select an image from MMVP [48] and perform forward passes through both the baseline and LaVer models, extracting hidden states at different layers. Using SigLIP 2 [49] as the vision encoder and Qwen2.5-7B-Instruct [41] as the LLM backbone, we normalize the hidden states and compute feature-wise cosine similarity. The visualization reveals that vision tokens exhibit substantially higher inter-feature cosine similarity in the last layer compared to middle layers, demonstrating progressive visual representation homogenization. **(b-c)** We extract the last-layer hidden states from both baseline and LaVer models and apply t-SNE [50] to project the high-dimensional features into 2D space. The scattered features are color-coded to distinguish vision and text tokens. Compared to the baseline, LaVer's visual features exhibit stronger interaction with textual features, indicating more effective learning of joint multimodal embeddings. **(d)** We process images from MMVP [48] through both models and extract hidden states of vision tokens across all layers, computing the averaged cosine similarity between normalized features. For the baseline, the averaged visual cosine similarity decreases mildly in early and middle layers but increases drastically in deeper layers, indicating rapid visual information loss in the final stages. In contrast, LaVer maintains a consistent decreasing trend throughout all layers, preserving visual discriminability. **(e)** We analyze the proportion of attention allocated to vision tokens across layers, following [9, 25]. Specifically, using images and corresponding queries from MMVP [48], we compute the proportion of attention scores allocated to previous vision tokens for each predicted token. These layer-wise proportions are averaged across all predictions to reveal the overall trend. The comparison demonstrates that LaVer enables the model to allocate significantly more attention to vision tokens, indicating more effective utilization of visual representations.

Fig. 4. **(a)** We extract the hidden states of vision tokens from the last layer and reshape them to recover their spatial structure. Applying PCA [1] for dimensionality reduction to 3 components, we normalize the features to the range $[0, 255]$ and visualize them as a 3-channel RGB image. The visualization demonstrates that LaVer generates highly discriminative visual features while preserving clear spatial structural information. All experiments use images from MMVP [48], SigLIP 2 [49] as the vision encoder, and Qwen2.5-7B-Instruct [41] as the LLM backbone. **(b)** We analyze the training dynamics of the baseline and models trained with different loss configurations. Specifically, using images from MMVP [48], we compute the averaged normalized cosine similarity of last-layer visual features at 1K iteration intervals. While naive application of MIM leads to visual feature inconsistency, LaVer effectively prevents the MIM shortcut, guiding the model toward more discriminative visual representations throughout training.

Fig. 6. We visualize the attention scores of the last predicted token on images from MMVP [48], alongside the corresponding PCA visualization of last-layer visual features. The results demonstrate that LaVer enhances the model's ability to focus on spatially relevant regions that correspond to the generated text tokens, indicating improved visual-textual alignment.

Table 12. **Performance comparison with different LLM backbone.** LaVer achieves superior performance on Vicuna-7B-v1.5 [57], demonstrating its effectiveness and generalizability across diverse architectural configurations.

Benchmark	SigLIP 2			CLIP			DINOv2		
	Baseline	LaVer	Δ_{Baseline}	Baseline	LaVer	Δ_{Baseline}	Baseline	LaVer	Δ_{Baseline}
GQA	50.00	**53.52**	↑3.52	51.01	**54.02**	↑3.01	49.75	**51.76**	↑2.01
MMB^{EN}	67.83	**68.55**	↑0.72	**66.50**	65.13	↓1.37	53.69	**55.29**	↑1.60
SEED^{I}	66.33	**67.29**	↑0.96	64.58	**65.36**	↑0.78	60.09	**63.77**	↑3.68
MME	1426.52	**1441.46**	↑14.94	1292.78	**1317.35**	↑24.57	1232.67	**1248.07**	↑15.40
RWQA	55.42	**58.43**	↑3.01	53.86	**55.56**	↑1.70	45.75	**48.55**	↑2.80
MMMU	40.89	**42.11**	↑1.22	41.00	**44.33**	↑3.33	40.33	**41.33**	↑1.00
MM*	51.74	**53.28**	↑1.54	44.85	**45.38**	↑0.53	40.82	**44.24**	↑3.42
OCRB	389	**401**	↑12	311	**352**	↑41	306	**349**	↑43
TVQA	65.07	**67.36**	↑2.29	52.39	**52.60**	↑0.21	40.66	**42.25**	↑1.59
CQA	62.48	**64.80**	↑2.32	37.84	**44.56**	↑6.72	28.44	**29.04**	↑0.60
AI2D	70.84	**72.68**	↑1.84	67.97	**69.09**	↑1.12	61.14	**63.09**	↑1.95
MMVP	47.67	**49.00**	↑1.33	41.00	**44.67**	↑3.67	44.33	**46.67**	↑2.34
CV-B^{2D}	54.19	**56.55**	↑2.36	42.94	**44.45**	↑1.51	41.47	**45.49**	↑4.02
SQA	73.89	**75.08**	↑1.19	**70.79**	70.37	↓0.42	62.93	**65.33**	↑2.40
MathV	54.20	**56.80**	↑2.60	43.90	**46.40**	↑2.50	40.80	**43.30**	↑2.50
Hallu	66.89	**67.83**	↑0.94	67.52	**70.25**	↑2.73	60.26	**69.74**	↑9.48
POPE	87.07	**88.07**	↑1.00	90.13	**90.73**	↑0.60	86.90	**87.67**	↑0.77
Average	53.85	**55.43**	↑1.58	49.24	**50.81**	↑1.57	44.60	**46.96**	↑2.37

Table 13. **Performance comparison of stage 2.** LaVer achieves superior performance on stage 2, indicating a direct effect of LaVer on visual representation enhancement.

Benchmark	SigLIP 2			CLIP			DINOv2		
	Baseline	LaVer	Δ_{Baseline}	Baseline	LaVer	Δ_{Baseline}	Baseline	LaVer	Δ_{Baseline}
GQA	52.03	**53.77**	↑1.74	50.25	**53.28**	↑3.03	48.01	**51.26**	↑3.25
MMB^{EN}	70.51	**73.52**	↑3.01	67.04	**67.53**	↑0.49	57.82	**60.77**	↑2.95
SEED^{I}	65.63	**66.92**	↑1.29	58.94	**64.55**	↑5.61	60.74	**61.91**	↑1.17
MME	1387.81	**1434.21**	↑46.40	1215.99	**1313.32**	↑97.33	1169.80	**1224.12**	↑54.32
RWQA	52.43	**58.74**	↑6.31	52.38	**54.76**	↑2.38	48.42	**52.29**	↑3.87
MMMU	42.67	**44.67**	↑2.00	42.44	**42.89**	↑0.45	40.22	**41.67**	↑1.45
MM*	45.11	**51.41**	↑6.30	41.63	**44.25**	↑2.62	39.43	**41.10**	↑1.67
OCRB	304	**407**	↑103	288	**306**	↑18	247	**258**	↑11
TVQA	54.94	**59.99**	↑5.05	60.92	**62.35**	↑1.43	52.44	**54.97**	↑2.53
CQA	51.60	**54.16**	↑2.56	40.40	**46.80**	↑6.40	40.40	**41.20**	↑0.80
AI2D	70.37	**74.94**	↑4.57	69.56	**70.65**	↑1.09	71.14	**72.47**	↑1.33
MMVP	38.33	**43.67**	↑5.34	26.00	**37.00**	↑11.00	26.67	**28.00**	↑1.33
CV-B^{2D}	44.46	**49.35**	↑4.89	40.19	**45.79**	↑5.60	40.68	**42.61**	↑1.93
SQA	70.09	**73.73**	↑3.64	65.49	**69.12**	↑3.63	64.57	**64.60**	↑0.03
MathV	49.10	**55.60**	↑6.50	50.90	**53.90**	↑3.00	42.10	**43.20**	↑1.10
Hallu	56.15	**59.62**	↑3.47	56.38	**58.31**	↑1.93	55.10	**57.05**	↑1.95
POPE	**85.90**	85.67	↓0.23	89.77	**90.83**	↑1.06	88.13	**88.87**	↑0.74
Average	50.01	**53.34**	↑3.33	47.83	**50.76**	↑2.93	45.68	**47.22**	↑1.54

F. Additional Experiment Results

Full results of the parameter scaling property of LaVer. Table 5 presents a comprehensive evaluation of LaVer's scalability across different model parameter sizes, ranging from Qwen2.5-1.5B-Instruct to Qwen2.5-7B-Instruct [41], using both SigLIP 2 [49] and CLIP [42] vision encoders. The results demonstrate that LaVer consistently delivers performance improvements over the baseline across all parameter scales. For SigLIP 2-based models, LaVer achieves average improvements of **+2.42**, **+2.07**, and **+2.15** points for 1.5B, 3B, and 7B parameter configurations, respectively. Similarly, for CLIP-based models, LaVer yields average gains of **+3.12**, **+1.31**, and **+2.66** points across the same parameter scales. Notably, LaVer exhibits particularly strong improvements on challenging benchmarks such as OCR-Bench [33] (up to **+103** points for SigLIP 2-7B) and MME [16] (up to **+185.03** points for CLIP-7B), ChartQA [37] (up to **+12.00** points for CLIP-7B), and MathVista [36] (up to **+7.20** points for SigLIP 2-1.5B). The consistent positive gains across 17 diverse benchmarks and multiple parameter scales substantiate that LaVer possesses robust scaling properties with respect to model parameters, maintaining its effectiveness in mitigating visual representation homogenization regardless of model capacity.

Full results of the data scaling property of LaVer. Table 6 presents a comprehensive analysis of LaVer's scalability across different training dataset sizes, ranging from 800K to 4M samples, using both SigLIP 2 [49] and CLIP [42] vision encoders with Qwen2.5-7B-Instruct [41] as the language model. The results demonstrate that LaVer consistently delivers substantial performance improvements over the baseline across all data scales. For SigLIP 2-based models, LaVer achieves average improvements of **+2.15**, **+2.16**, and **+1.80** points for 800K, 2M,

Table 14. **Compatibility of LaVer with enriched visual inputs.** LaVer consistently improves performance when combined with methods that enrich visual inputs, demonstrating its broad compatibility and effectiveness across diverse visual enhancement strategies.

$\mathcal{G}_\xi$	GQA	MMB^{EN}	$SEED^{I}$	MME	RWQA	MMMU	MM^{*}	OCRB	TVQA	CQA	AI2D	MMVP	$CV\text{-}B^{2D}$	SQA	MathV	Hallu	POPE	Avg.
A-MoF	51.52	70.83	62.53	1312.14	54.77	41.11	41.53	338	59.46	41.04	73.45	41.67	61.34	67.31	51.20	62.47	89.13	51.19
A-MoF + LaVer	**52.26**	**72.02**	**65.66**	**1321.51**	**56.34**	**43.22**	**43.87**	**360**	**61.68**	**42.52**	**78.38**	**44.67**	**62.66**	**68.51**	**54.40**	**62.89**	**89.57**	**52.92**

and 4M training samples, respectively. Similarly, for CLIP-based models, LaVer yields progressively increasing average gains of **+2.66**, **+3.04**, and **+3.77** points across the same data scales, indicating enhanced effectiveness with larger training datasets. Notably, LaVer exhibits particularly strong improvements on challenging benchmarks such as TextVQA [44] (up to **+8.61** points for CLIP-2M). The consistent positive gains across 17 diverse benchmarks and multiple data scales substantiate that LaVer possesses robust scaling properties with respect to training data size, maintaining its effectiveness in mitigating visual representation homogenization regardless of dataset scale. Furthermore, the observation that CLIP-based models show increasing improvements with larger datasets (from **+2.66** to **+3.77**) suggests that LaVer's benefits become more pronounced when trained on more extensive data, highlighting its potential for further performance gains with additional training data.

Full results of the ablation study on masking strategies. Table 7 presents a comprehensive analysis of LaVer's performance under different masking strategies, examining both cosine and constant scheduling approaches across various masking ratios. The results are evaluated on 17 diverse benchmarks using SigLIP 2 [49] and CLIP [42] vision encoders with Qwen2.5-7B-Instruct [41] as the language model. For SigLIP 2-based models, the cosine scheduling strategy with a masking ratio of 0.05 achieves the best average performance of **57.87**, representing a substantial improvement of **+2.15** points over the baseline (55.72). This configuration demonstrates consistent gains across most benchmarks. Comparing scheduling strategies, cosine scheduling generally outperforms constant scheduling across different masking ratios. For instance, at a ratio of 0.05, cosine scheduling achieves **57.87** compared to constant scheduling's **56.32**, demonstrating the effectiveness of gradually varying masking intensity during training. Regarding masking ratio selection, lower ratios (0.05-0.1) consistently yield superior performance compared to higher ratios (0.2-0.3) under cosine scheduling, suggesting that moderate masking preserves sufficient visual information while effectively mitigating representation homogenization. For CLIP-based models, similar trends emerge with cosine scheduling at 0.05 ratio achieving the best average performance of **53.24**, representing a **+2.66** point improvement over the baseline. The results demonstrate LaVer's robustness across different masking configurations, with the cosine scheduling strategy at lower masking ratios consistently delivering optimal performance across both vision encoders and diverse evaluation benchmarks. We hypothesize that the inferior performance observed with high masking ratios stems from insufficient convergence, as conventional MIM-based methods typically require large-scale training to fully develop the model's reconstruction capabilities [7, 39, 43]. Scaling to larger datasets remains an avenue for future investigation.

Full results of the ablation study on EMA updating strategies. Table 8 presents a comprehensive analysis of LaVer's performance under different EMA teacher updating strategies, evaluated across diverse benchmarks using SigLIP 2 [49] and CLIP [42] vision encoders with Qwen2.5-7B-Instruct [41] as the LLM backbone. The results demonstrate LaVer's robustness across various EMA configurations, with cosine scheduling slightly outperforming constant scheduling. Regarding updating frequency, moderate frequencies (50-200 steps) generally yield optimal performance, with 100 steps emerging as the most balanced configuration across both vision encoders. For decay rate selection, a rate of 0.95 demonstrates superior stability and performance compared to both lower (0.9) and higher (0.99, 0.999) rates. The comprehensive evaluation across diverse benchmarks confirms that LaVer's performance remains consistently strong regardless of the specific EMA strategy employed, highlighting the method's robustness and effectiveness in learning intrinsic visual representation capabilities.

Full results of the ablation study on spatial awareness. Table 9 presents a comprehensive analysis of LaVer's spatial awareness mechanisms, evaluated across diverse benchmarks. The results reveal several key insights. First, employing mixed attention alone (full attention for vision tokens) yields moderate improvements over the baseline, demonstrating that enhanced token interactions can benefit visual understanding. Second, applying 2D-RoPE in isolation shows minimal impact, with performance remaining largely comparable to the baseline. Third, combining mixed attention with 2D-RoPE without LaVer produces mixed results, with performance gains that are inconsistent across benchmarks. Most importantly, the full LaVer framework, which integrates all three components achieves the best overall performance. This configuration demonstrates consistent improvements across challenging benchmarks. These results confirm that LaVer's holistic approach to spa-

tial awareness, combining learned visual representations with architectural enhancements, is essential for achieving superior multimodal understanding capabilities.

Full results of the ablation study on loss components. Table 10 provides a comprehensive analysis of different loss function configurations in LaVer, evaluated across diverse benchmarks. The results reveal three critical insights into the design of effective visual representation learning objectives. First, applying masked image modeling ($\mathcal{L}_{\text{MIM}}$) alone significantly degrades performance, with average scores dropping from **55.72** to **53.76** for SigLIP 2 and from **50.58** to **49.71** for CLIP. This deterioration stems from the *visual feature inconsistency* problem, where the model exploits the MIM objective by generating identical visual features, thereby undermining the discriminative capacity essential for downstream tasks. Second, incorporating the global alignment loss ($\mathcal{L}_{\text{GA}}$) alongside $\mathcal{L}_{\text{MIM}}$ partially mitigates this issue, improving average performance to **56.46** for SigLIP 2 and **52.01** for CLIP. However, this configuration still underperforms the full LaVer framework, as the symmetric nature of $\mathcal{L}_{\text{GA}}$ constrains the model's ability to learn sufficiently discriminative representations. Third, our proposed contrastive global alignment loss ($\mathcal{L}_{\text{CGA}}$) effectively addresses both shortcomings, achieving the best overall performance with average scores of **57.87** for SigLIP 2 and **53.24** for CLIP. These results confirm that $\mathcal{L}_{\text{CGA}}$ simultaneously prevents visual feature inconsistency while encouraging discriminative feature learning, thereby enhancing both visual information preservation and utilization across diverse multimodal understanding tasks.

Language capabilities. A potential concern when introducing vision-centric supervisory signals is whether they may compromise the model's language capabilities. To address this, we evaluate the language performance of both the baseline and our LaVer model using SigLIP 2 [49] and CLIP [42] as vision encoders, respectively, as shown in Table 11. Specifically, we adopt three representative benchmarks: IFEval [61], which evaluates the model's capability to strictly follow instructions; MMLU [18], a comprehensive general language benchmark containing diverse tasks across various domains; and BBH [45], which assesses language reasoning and world knowledge capabilities. The results demonstrate that LaVer maintains competitive language performance across both vision encoder configurations. These findings confirm that LaVer successfully preserves language capabilities while simultaneously achieving substantial improvements in visual understanding tasks, thereby validating the effectiveness of our approach in balancing multimodal learning objectives.

Extended comparison with different LLM backbone. To validate the generalizability of LaVer across different architectural configurations, we conduct comprehensive experiments using Vicuna-7B-v1.5 [57] as the LLM backbone, paired with three vision encoders: SigLIP 2 [49], CLIP [42], and DINOv2 [39]. As shown in Table 12, LaVer consistently outperforms the baseline across all three vision encoder configurations, achieving average improvements of **+1.58**, **+1.57**, and **+2.37** points for SigLIP 2, CLIP, and DINOv2, respectively. Notably, LaVer demonstrates substantial gains on benchmarks requiring fine-grained visual understanding, including ChartQA [37] (**+2.32**, **+6.72**, **+0.60**), MMVP [48] (**+1.33**, **+3.67**, **+2.34**), and CV-Bench-2D [47] (**+2.36**, **+1.51**, **+4.02**). These results confirm that LaVer effectively preserves and enhances multimodal capabilities across different LLM architectures, demonstrating its robustness and broad applicability in diverse scenarios.

Performance comparison of stage 2. Following the state-of-the-art open-source framework LLaVA-OneVision [2], we adopt a 3-stage training recipe and exclusively apply LaVer to stage 2, which is responsible for visual knowledge injection. To thoroughly investigate how LaVer facilitates visual knowledge learning, we compare models with and without LaVer during stage 2, using three vision encoders: SigLIP 2 [49], CLIP [42], and DINOv2 [39]. As presented in Table 13, LaVer consistently demonstrates substantial improvements across all configurations, achieving average gains of **+3.33**, **+2.93**, and **+1.54** points for SigLIP 2, CLIP, and DINOv2, respectively. Notably, LaVer exhibits particularly strong performance on benchmarks requiring fine-grained visual understanding: MMVP [48] shows remarkable improvements of **+5.34**, **+11.00** across the three encoders. These results reveal that the performance gains observed in stage 2 are even more pronounced than in the final stage, indicating that LaVer fundamentally enhances the model's capacity to learn representative visual features during the critical visual knowledge injection phase, thereby establishing a strong foundation for superior multimodal performance across diverse downstream tasks.

Compatibility with other methods. LaVer introduces vision-centric supervisory signals by predicting masked vision tokens, a design that naturally ensures compatibility with methods that enrich visual inputs. To validate this compatibility, we integrate LaVer with A-MoF [48], which aggregates visual features from multiple vision encoders to form more informative visual embeddings. Specifically, following [48], we combine visual features from CLIP-ViT-L/14@224 [42] and DINOv2-Large/14@224 [39] with balance coefficients of $0.25, 0.75$, which are empirically validated as optimal hyperparameters in [48]. The identical patch size and resolution ensure that visual features share the same shape and can be directly aggregated. As presented in Table 14, LaVer consistently delivers substantial improvements across all benchmarks when integrated with A-MoF. Overall, LaVer achieves an average improvement of **+1.73** points across all benchmarks, demonstrating that

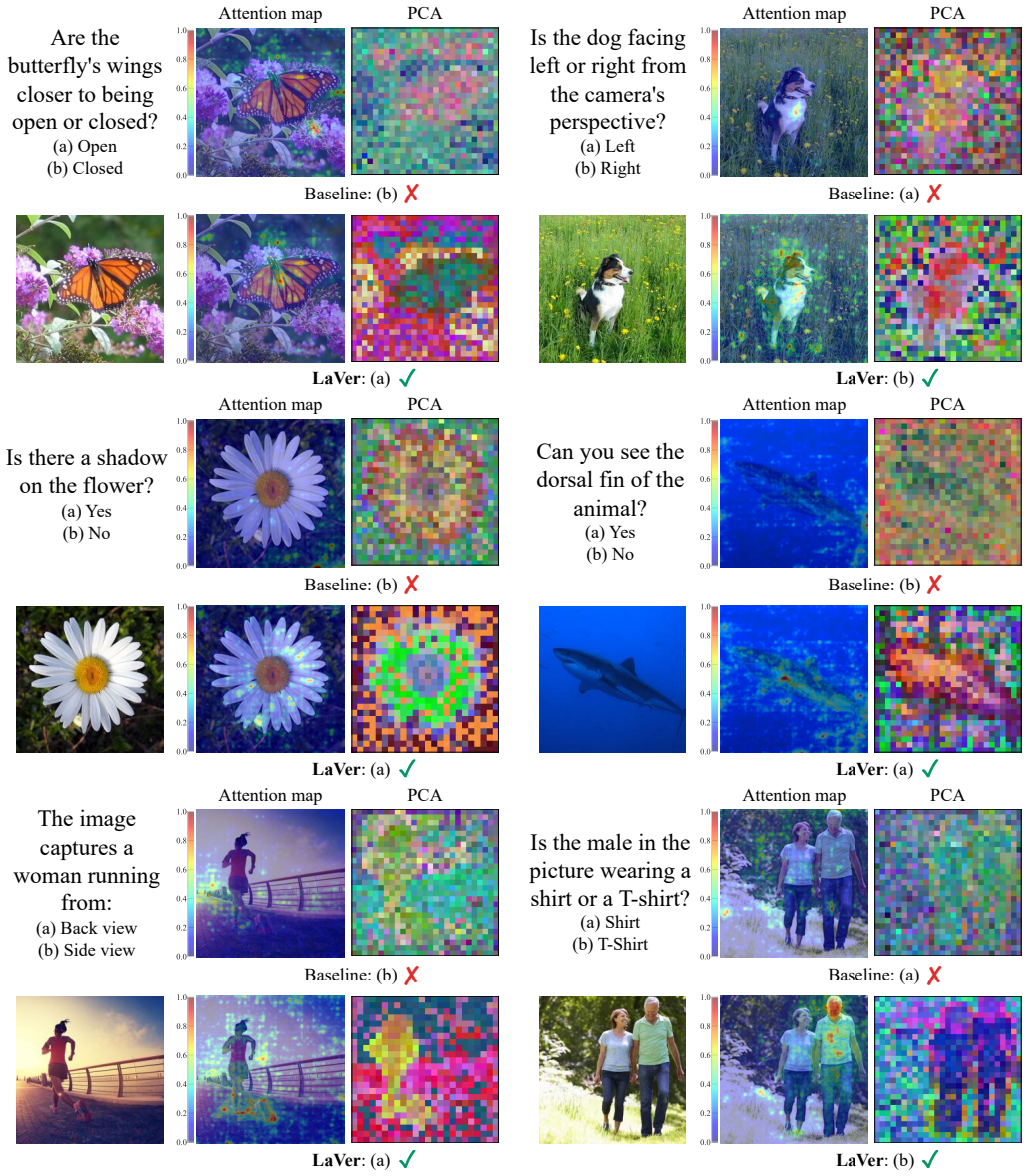


Figure 5. **Qualitative comparisons on MMVP [48].**

it provides additional benefits when integrated with vision-enhanced methods and confirming its broad compatibility with diverse visual enhancement strategies.

G. Qualitative Analysis

We provide additional qualitative comparisons in Fig. 5 on MMVP [48]. As illustrated in Fig. 5, we visualize both the attention scores on vision tokens from the last predicted token and the PCA visualization of visual features extracted from the final layer. The baseline model frequently fails to attend to spatial regions that are semantically relevant to the text query, exhibiting limited attention distributions across the visual input. In contrast, our LaVer consistently allocates substantially higher attention weights to the corresponding regions of interest, thereby enabling more accurate and contextually appropriate responses. Furthermore, LaVer demonstrates the ability to capture diverse visual patterns and generate highly discriminative visual features that encode rich structural information. These qualitative results provide compelling evidence that LaVer effectively mitigates modality imbalance by incorporating visual supervisory signals, enabling the model to integrate visual and textual modalities seamlessly and achieve superior performance on multimodal tasks.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010. 8

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv*, 2509.23661, 2025. 5, 11

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv*, 2502.13923, 2025. 7

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 7

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 8

[6] Weitong Cai, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Mllm as video narrator: Mitigating modality imbalance in video moment retrieval. *PR*, 2025. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 10

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 7

[9] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024. 8

[10] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *EMNLP*, 2024. 1

[11] Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *TMLR*, 2024. 7

[12] X.AI Corp. Grok-1.5 vision preview: Connecting the digital and physical worlds with our first multimodal model, 2024. 7

[13] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In *NeurIPS*, 2024. 7

[14] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACM MM*, 2024. 7

[15] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Costa, Louis Béthune, Zhe Gan, Alexander Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua Susskind, and Alaa El-Nouby. Multimodal autoregressive pre-training of large vision encoders. In *CVPR*, 2025. 7

[16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025. 7, 9

[17] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024. 7

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 11

[19] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 7

[20] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv*, 2405.07987, 2024. 1

[21] Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. In *CVPR*, 2024. 1

[22] Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 7

[23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019. 1

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024. 7

[25] Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *ICCV*, 2025. 7, 8

[26] Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv*, 2410.12787, 2024. 1

[27] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*, 2024. 7

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *TMLR*, 2025. 6

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 7

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv*, 2310.03744, 2023. 8

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 6

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024. 7

[33] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 2024. 7, 9

[34] Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. On the faithfulness of visual thinking: Measurement and enhancement. *arXiv*, 2510.23482, 2025. 1

[35] Pan Lu, Swaroop Mishra, Tony 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. In *NeurIPS*, 2022. 7

[36] 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. In *ICLR*, 2024. 7, 9

[37] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 7, 9, 11

[38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 8

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2, 3, 8, 10, 11

[40] Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *AAAI*, 2025. 1

[41] 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. *arXiv*, 2412.15115, 2025. 1, 8, 9, 10

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 7, 8, 9, 10, 11

[43] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv*, 2508.10104, 2025. 2, 3, 4, 7, 10

[44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 7, 10

[45] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL*, 2023. 11

[46] Hao Tang, Chenwei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. In *NeurIPS*, 2025. 7, 8

[47] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: a fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 6, 7, 11

[48] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs . In *CVPR*, 2024. 1, 2, 7, 8, 11, 12

[49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv*, 2502.14786, 2025. 1, 2, 3, 6, 8, 9, 10, 11

[50] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[51] Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *ICLR*, 2025. 6

[52] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv*, 2510.17269, 2025. 6

[53] Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models. *arXiv*, 2508.10552, 2025. 1

[54] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 7

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 6

[56] YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. Debiasing multimodal large language models via penalization of language priors. In *ACM MM*, 2025. 1

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023. 9, 11

[58] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. Mllms are deeply affected by modality bias. *arXiv*, 2505.18657, 2025. 1

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 8

[60] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022. 1, 6

[61] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv*, 2311.07911, 2023. 11