VDEP: Establishing Equivalence Between Image and Text Token Through Autoregressive Pre-training in MLLMs

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) often underutilize visual information, leading to imbalanced alignment and limited performance. Through theoretical analysis, we reveal that existing alignment objectives risk collapsing into a unimodal, text-only training process. To address this, we propose **V**isual **D**ynamic **E**mbedding-guided **P**retraining (**VDEP**) a hybrid autoregressive framework that supervises image-related hidden states via dynamic embeddings from an MLP appended to the visual encoder. VDEP integrates visual tokens into training without added architectural complexity, reframing alignment as an information recovery task focused on fine-grained visual semantics. Our model-agnostic method consistently outperforms strong baselines across 13 benchmarks, setting a new standard for large-scale vision-language alignment. Code and models are available at https://github.com/anonymous-gpu/VDEP_LLava_1.5.git.

1 Introduction

Large language models (LLMs) such as ChatGPT (Schulman et al., 2022) have revolutionized natural language processing by demonstrating remarkable zero-shot reasoning capabilities across diverse tasks through flexible language instructions. Inspired by this success, multimodal large language models (MLLMs) have rapidly gained traction by unifying vision and language modalities. Among these, LLava has emerged as a dominant architecture due to its simplicity and efficiency, bridging image and text features via a lightweight linear layer—later enhanced to an MLP—enabling competitive multimodal alignment with minimal computational overhead (Li et al., 2022; Zhang et al., 2024b).

However, despite their widespread adoption, such streamlined MLLM architectures exhibit a persistent modality imbalance. Training objectives remain predominantly text-centric, which biases the model towards textual signals and leads to underutilization of visual features. This imbalance manifests in well-documented issues including hallucinations (Wang et al., 2024a, 2023), impaired fine-grained visual understanding (Wu et al., 2024; Lai et al., 2024), and suboptimal performance on vision-language benchmarks (Lin et al., 2024). Our empirical analysis of LLava's layer-wise attention maps (Fig. 1) further corroborates this: the model's attention to image tokens remains nearly static across layers and shows minimal correlation with textual queries, indicating insufficient engagement with visual inputs and limited cross-modal semantic fusion.

From a theoretical standpoint, we rigorously analyze the multimodal alignment objective under the assumption that image-text pairs are strongly correlated. We prove that insufficient preservation of visual semantic content during training can cause the alignment objective to degenerate into a unimodal, text-only training task. In other words, the model effectively ignores visual information, collapsing the multimodal learning process and exacerbating modality imbalance. This insight reveals a fundamental limitation of existing MLLM pretraining objectives, which focus exclusively on reconstructing textual information while leaving visual semantics unreconstructed and unregularized.

To address this critical issue, we propose **V**isual **D**ynamic **E**mbedding-guided **P**retraining (**VDEP**), a principled and effective approach that explicitly incorporates visual semantic retention into the training objective. VDEP introduces a dynamic visual semantic reconstruction task by supervising the LLM's hidden states corresponding to visual inputs with dynamic embeddings generated by the MLP following the visual encoder. This dynamic supervision encourages the model to maintain and reconstruct rich visual semantic information throughout the autoregressive training process, thereby pre-

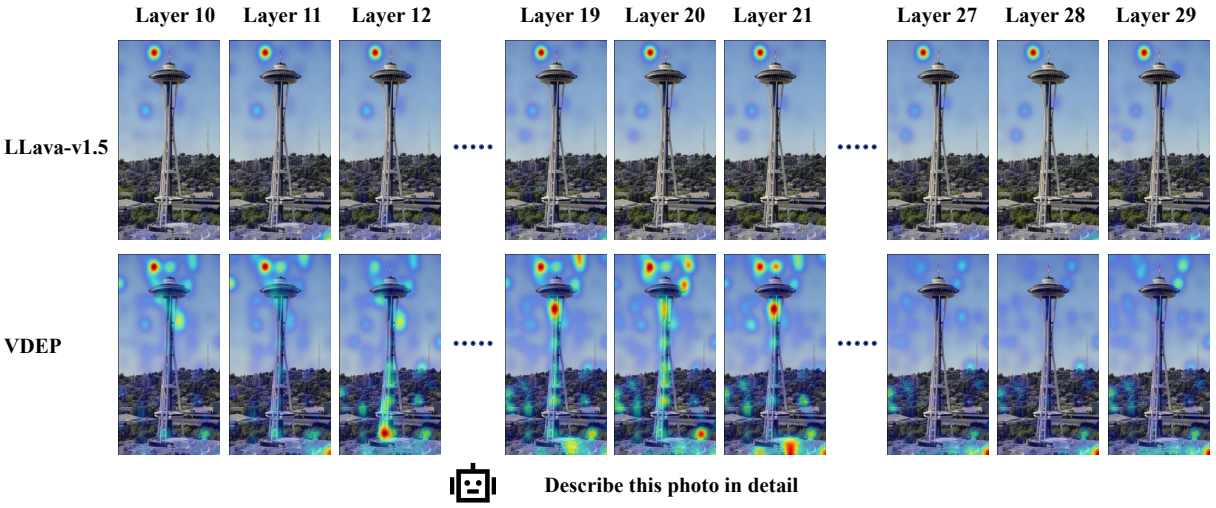

Figure 1: Visualization of layer-wise attention maps for the input image by LLava and VDEP. The example is taken from LLava-Bench (Liu et al., 2024a) with the query "*Describe this photo in detail*". The results demonstrate that VDEP significantly enhances the model's ability to capture critical visual features, particularly excelling at identifying object boundaries.

venting the alignment objective from collapsing into text-only learning.

We provide theoretical guarantees demonstrating that this dynamic visual semantic reconstruction effectively mitigates alignment degeneration, ensuring a more balanced and robust fusion of visual and textual modalities. Importantly, VDEP achieves these benefits without modifying LLava's architecture or requiring additional data, preserving the model's efficiency and adaptability.

As shown in Fig. 1, VDEP substantially enhances the model's sensitivity to critical visual features, such as fine object boundaries and spatial relationships. Our extensive experiments demonstrate that VDEP consistently improves overall VQA performance across different model scales (3B and 7B). Notably, on the challenging MM-Reasoning benchmark (MMStart), VDEP achieves a remarkable 7.7% accuracy gain, underscoring its effectiveness in enhancing visual understanding and generation capabilities. These results comprehensively validate that our method significantly advances the visual grounding and reasoning abilities of MLLMs.

In summary, our contributions are:

- We theoretically analyze the risk of alignment collapse into unimodal text-only training due to insufficient visual semantic preservation, under strong image-text correlation assumptions, thereby providing a theoretical justification for the necessity of incorporating visual semantic training.
- We propose VDEP, a novel training paradigm that integrates dynamic visual semantic reconstruction into LLava's pretraining, effectively preventing alignment degeneration without architectural changes or extra data.
- We conduct extensive experiments across multiple model scales and benchmarks, demonstrating consistent and significant improvements in visual understanding and reasoning, including a 7.7% accuracy boost on MMStart.

2 Related Works

With the rise of multimodal large models, the ViT-MLP-LLM architecture—popularized after LLaVA(Liu et al., 2024b) has become mainstream due to its strong adaptability. It enables LLMs to efficiently gain visual capabilities via lightweight pretraining and fine-tuning, accelerating progress in the field. This section focuses on research related to text bias and modality alignment within this architecture.

2.1 The issue of text bias in MLLM

Prior research has consistently revealed a pronounced bias toward textual information in Multimodal Large Language Models (MLLMs), which undermines effective multimodal integration. For

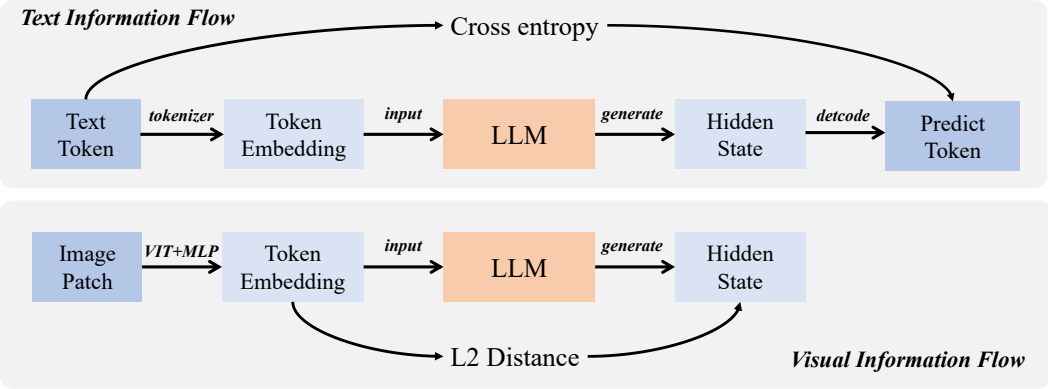

Figure 2: Information flow of VDEP training. (a) Text pre-training: text tokens are embedded and predicted by the LLM with cross-entropy loss. (b) Image pre-training: image patches are encoded into embeddings that guide the LLM hidden states to reconstruct visual information without labels.

example, (Huang et al., 2024) identify that attention mechanisms tend to favor text inputs and propose adjustments to mitigate this imbalance. Similarly, (Parcalabescu and Frank, 2024) find that, across various vision-language models, textual contributions outweigh visual ones in most tasks, although images play a stronger role in explanation generation than in answer prediction. Investigations by (Wang et al., 2024a) and (Zhang et al., 2024b) further expose that hallucinations increase with model depth, linked to an overreliance on textual cues, and suggest attention-based remedies. Extending this line of work, (Leng et al., 2024) systematically analyze hallucinations across language, vision, and audio modalities, attributing them to excessive unimodal priors and spurious cross-modal correlations, and propose the "Curse of Multimodality" benchmark to evaluate these effects. Together, these studies highlight the critical challenge of text dominance in MLLMs and underscore the need for alignment objectives that better preserve multimodal semantics. Although many studies have revealed the phenomenon of this preference, the exploration of the essential reasons for this preference is still insufficient, lacking intuitive and more in-depth theoretical research

2.2 Enhancing modal alignment in MLLM

Robust alignment between visual and textual modalities is crucial for Multimodal Large Language Models. (Zhao et al., 2024) address the common assumption of uniform image-text alignment by grouping pairs based on alignment quality and learning adaptive representations, improving performance across tasks. Data-centric methods like (Yin et al., 2024) and (Chow et al., 2024) enhance alignment by generating accurate synthetic pairs and applying cross-modal contrastive learning. Architecturally, (Tong et al., 2024) introduce a trainable MLP for image generation to boost visual semantics, while (Wang et al., 2024b) use a denoising module to preserve visual features. These works collectively advance modal alignment through adaptive training, improved data, and targeted model design. Although recent advances have demonstrated the effectiveness of incorporating additional vision-related training to enhance alignment in MLLMs, these approaches invariably rely on modifying the original model architecture, introducing extra trainable modules, or improving data quality to indirectly boost alignment. Such dependencies significantly constrain the generalizability and applicability of alignment methods.

3 Background

3.1 Problem statement

Modern multimodal large language models (MLLMs) align vision and language modalities by modeling the joint distribution of image representations $\mathbf{X}_v \in \mathbb{R}^{N\times d}$ and text representations $\mathbf{X}_l \in \mathbb{R}^{M\times d}$. Given an image $\mathcal{I}$ partitioned into patches $\{p_j\}_{j=1}^{N}$, a vision transformer (ViT) with MLP projection produces visual embeddings:

$$\mathbf{X}_v = \text{MLP}(\text{ViT}(\{p_j\}_{j=1}^{N})) \quad (1)$$

The corresponding text sequence $\mathcal{T}$ is tokenized into $\mathbf{X}_l = [x_1, \ldots, x_M]$. During pretraining, the model minimizes cross-entropy loss between predicted text $\hat{\mathbf{X}}_l$ and ground truth:

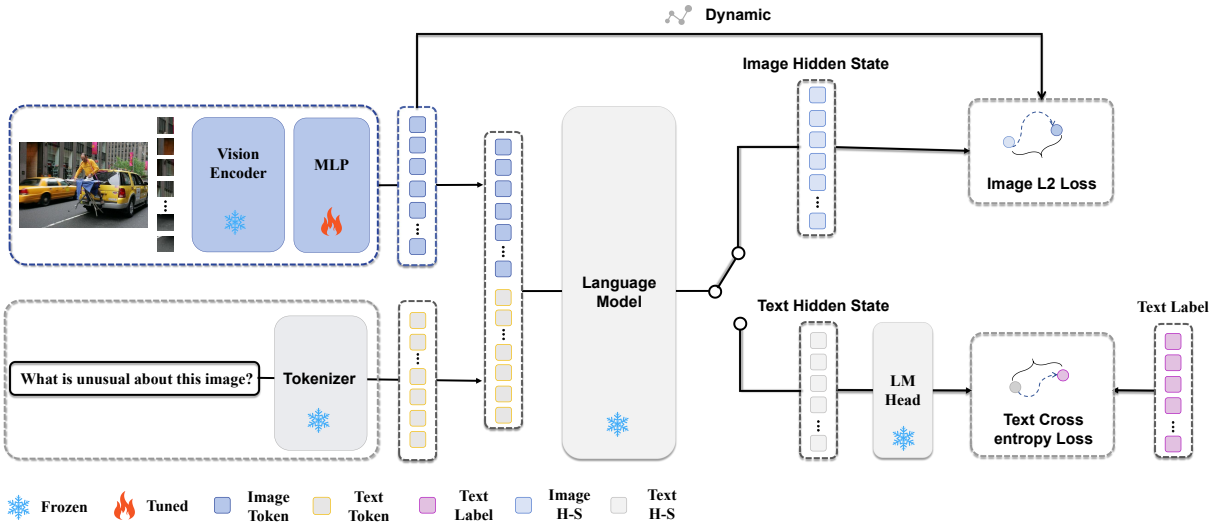

Figure 3: The LLava-VDEP training paradigm incorporates two distinct training modes. The VDEP mode performs supervised learning on image data, while the LLava mode is dedicated to supervised learning on text data. During batch training, a ratio parameter is used to control the proportional occurrence of these two modes within each batch, enabling an effective balance in the learning process.

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{M} \log P(x_t | \mathbf{X}_v, x_{<t}) \tag{2}$$

This parallels standard language model pretraining where text tokens are predicted autoregressively:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{M} \log P(x_t | x_{<t}) \tag{3}$$

3.2 Information-theoretic reformulation

From an information-theoretic perspective, the cross-entropy loss corresponds to conditional entropy minimization:

$$\min \mathcal{L}_{\text{CE}} \equiv \min \mathcal{H}(x_{\geq t} | \mathbf{X}_v, x_{<t}) \tag{4}$$

Let $\mathbf{H}_v \in \mathbb{R}^{N \times d_h}$ and $\mathbf{H}_l \in \mathbb{R}^{M \times d_h}$ denote the hidden states of visual and textual modalities respectively. We express the conditional entropy through mutual information:

$$\begin{aligned} \mathcal{H}(x_{\geq t} | \mathbf{H}_v, \mathbf{H}_l) &= \mathcal{H}(x_{\geq t}) - \mathcal{I}(x_{\geq t}; \mathbf{H}_v, \mathbf{H}_l) \\ \Rightarrow \min \mathcal{L}_{\text{CE}} &\equiv \max \mathcal{I}(x_{\geq t}; \mathbf{H}_v, \mathbf{H}_l) \end{aligned} \tag{5}$$

Applying the chain rule of mutual information to (5):

$$\begin{aligned} &\mathcal{I}(x_{\geq t}; \mathbf{H}_v, \mathbf{H}_l) \\ &= \underbrace{\mathcal{I}(x_{\geq t}; \mathbf{H}_v)}_{\text{Visual Contribution}} + \underbrace{\mathcal{I}(x_{\geq t}; \mathbf{H}_l | \mathbf{H}_v)}_{\text{Textual Flow}} \end{aligned} \tag{6}$$

3.3 Visual semantic retention collapse

We analyze a critical failure mode in multimodal large language models (MLLMs) where the model fails to effectively capture and retain visual semantics during pretraining.

Assumption 1 (No visual information in hidden states).

$$\mathcal{I}(\mathbf{X}_v; \mathbf{H}_v) \to 0$$

i.e., the hidden states $\mathbf{H}_v$ contain no information about the visual input $\mathbf{X}_v$. This scenario can arise because the LLM parameters remain frozen during pretraining; even if a multilayer perceptron (MLP) projects visual features into the LLM embedding space, the LLM may fail to interpret or utilize these visual semantics effectively.

Assumption 2 (Strong image-text correlation).

$$\mathcal{I}(\mathbf{X}_v; \mathbf{X}_l) \gg 0$$

meaning the visual input $\mathbf{X}_v$ and the corresponding textual description $\mathbf{X}_l$ share substantial semantic information. This holds when the MLP successfully projects visual semantics into the textual semantic space, resulting in a strong correlation between image and text modalities.

Under these assumptions, consider the mutual information between the visual hidden state $\mathbf{H}_v$ and the future text tokens $x_{\geq t}$:

$$\mathcal{I}(x_{\geq t}; \mathbf{H}_v) = 0 \tag{7}$$

Substituting Eq. (7) into the mutual information decomposition (cf. Eq. 6), we have:

$$\begin{aligned}&\mathcal{I}(x_{\geq t}; \mathbf{H}_l \mid \mathbf{H}_v) \quad (8)\\ &= \mathcal{H}(x_{\geq t} \mid \mathbf{H}_v) - \mathcal{H}(x_{\geq t} \mid \mathbf{H}_l, \mathbf{H}_v)\\ &= \mathcal{H}(x_{\geq t}) - \mathcal{H}(x_{\geq t} \mid \mathbf{H}_l)\\ &= \mathcal{I}(x_{\geq t}; \mathbf{H}_l). \quad (9)\end{aligned}$$

Consequently, the joint mutual information reduces to:

$$\mathcal{I}(x_{\geq t}; \mathbf{H}_v, \mathbf{H}_l) = \mathcal{I}(x_{\geq t}; \mathbf{H}_l). \quad (10)$$

Interpretation. This analysis uncovers a critical degeneration phenomenon in multimodal alignment: when the visual feature representation $\mathbf{H}_v$ carries no mutual information about the forthcoming tokens, the multimodal pretraining objective effectively reduces to a unimodal language modeling task conditioned solely on $\mathbf{H}_l$. In essence, the model disregards visual inputs entirely, collapsing MLLM pretraining into conventional text-only language modeling.

More importantly, this degeneration extends beyond the extreme case. Under Assumption 2—which generally holds in practical scenarios—insufficient retention of visual semantics causes the alignment target to increasingly approximate that of pure text modeling, revealing an intrinsic bias toward textual information that limits the model's ability to fully leverage multimodal signals. This analysis not only highlights a fundamental limitation of current alignment objectives but also substantiates the necessity of incorporating visual semantic training in MLLM alignment, consistent with observations in related work (see Section 2.2).

As depicted in Fig.2, existing MLLM alignment objectives focus mainly on textual reconstruction without explicitly preserving visual semantics, reinforcing the model's text preference and underscoring the need for new alignment strategies that integrate rich visual representations. While VDEP introduces visual semantics in a concise way to effectively balance the visual and text modalities.

4 The proposed method

4.1 Vision dynamic embedding pretraining

As illustrated in the upper panel of Fig. 2, text tokens benefit from explicit instance-level labels, enabling the use of cross-entropy loss to directly supervise token reconstruction. This forms the primary information flow driving alignment. In contrast, image data lack such granular labels, and consequently, the alignment objective imposes no explicit constraints on reconstructing visual semantics. As analyzed in Section 3.3, this imbalance naturally biases the model towards optimizing text reconstruction, undermining effective multimodal alignment.

To mitigate this degeneration, we introduce Vision Dynamic Embedding Pretraining (VDEP), which supplements the original alignment objective with an explicit reconstruction loss on visual semantic embeddings, as depicted in the lower panel of Fig. 2. Formally, the joint optimization objective is:

$$\mathcal{L}_{Total} = \mathcal{L}_t + \alpha \mathcal{L}_i, \quad (11)$$

where $\mathcal{L}_t$ and $\mathcal{L}_i$ denote the losses for text and image modalities, respectively, and $\alpha \in [0, 1]$ balances their relative contributions. By tuning α, we dynamically regulate the emphasis on visual semantics, promoting robust cross-modal representation learning. This approach not only balances the importance of visual and textual information reconstruction during training from an information flow perspective but also effectively prevents the alignment objective degradation risk caused by insufficient visual supervision, as discussed in Section 3.3.

4.2 Quantifying visual reconstruction target

To quantify the visual reconstruction loss $\mathcal{L}_i$, we adopt the L2 distance between the hidden state embedding $\mathbf{H}_i$ and the target embedding $x_{\geq t}$ as a proxy to estimate the mutual information $\mathcal{I}(\mathbf{H}_i, x_{\geq t})$ between these two embeddings. Intuitively, minimizing this L2 distance encourages the model to capture the shared semantic content, effectively aligning their representations.

This L2-based estimation provides a simple yet effective measure to guide the reconstruction of visual semantics. A theoretical justification and detailed proof of the feasibility of using L2 distance to estimate mutual information are provided in the Appendix D.

4.3 Hybrid multimodal alignment training

In multimodal LLMs, the visual token count often dwarfs that of text tokens, with images represented by hundreds of patches. Naively applying VDEP jointly risks overemphasizing visual features, lead-

METHOD	POPE	SEEDB^{I}	AI2D	MMSTAR	MMTB	OCRB	MMBENCH	
							EN	CN
TinyLLava-3B								
TINYLLAVA	86.58	69.10	60.36	37.19	48.73	337	**67.04**	**42.37**
TINYLLAVA-VDEP (OURS)	**86.98**	**69.35**	**60.85**	**37.65**	**49.08**	**343**	66.70	41.87
CHANGE	+0.40	+0.25	+0.49	+0.56	+0.35	+6.00	-0.26	-0.50
LLava-v1.5-7B								
LLAVA	85.85	66.10	55.63	33.48	48.86	297	64.30	57.62
LLAVA-VDEP (OURS)	**86.20**	**66.70**	**56.57**	**36.06**	48.00	**326**	**66.81**	**58.23**
CHANGE	+0.35	+0.60	+0.94	+2.58	-0.86	+29	+1.84	+0.33

Table 1: Comparison of VDEP (Ours) and LLava on General VLM Evaluation Benchmarks Across Model Sizes

METHOD	VQA^{OK}	GQA	VQA^{V2}	VQA^{T}	RWQA	SQA^{I}
TinyLLava-3B						
TINYLLAVA	57.50	61.20	79.13	51.66	53.33	70.55
TINYLLAVA-VDEP (OURS)	**57.97**	**61.67**	**79.24**	**51.73**	**54.25**	**71.39**
CHANGE	+0.47	+0.47	+-0.09	+0.07	+0.92	+0.84
LLava-v1.5-7B						
LLAVA	53.44	62.00	78.50	46.07	55.82	66.80
LLAVA-VDEP (OURS)	**57.68**	**62.50**	**79.20**	**46.76**	**57.64**	**68.36**
CHANGE	+3.36	+0.50	+0.70	+0.69	+1.86	+1.56

Table 2: Comparison of VDEP (Ours) and LLava on Visual Question Answering Datasets with Various Model Sizes

ing to overfitting on low-level image details and suboptimal alignment.

To address this, we propose a hybrid alignment scheme that decouples visual and textual optimization during pretraining. Specifically, each batch is stochastically split into two subsets, which alternate between optimizing $\mathcal{L}_i$ and $\mathcal{L}_t$. This decoupling stabilizes the textual embedding space, which serves as a reliable semantic anchor for aligning visual representations. By preserving textual distribution integrity and preventing visual dominance, our approach effectively suppresses noise and enhances alignment fidelity.

This hybrid strategy is employed exclusively during pretraining, integrating both VDEP and LLava objectives. The subsequent supervised fine-tuning stage adheres to the original LLava framework, focusing on instruction following. This two-stage design allows pretraining to concentrate on modality alignment, while fine-tuning leverages rich textual supervision to refine multimodal fusion. Empirical results validate the efficacy of our method in substantially improving cross-modal alignment quality.

5 Experiments Setting

In this section, we rigorously evaluate the efficacy of our proposed Visual Dynamic Embedding-guided Pretraining (VDEP) framework across a diverse array of multimodal benchmarks. Our experimental design is meticulously crafted to validate VDEP's ability to enhance multimodal alignment and visual semantic retention without architectural modifications or additional data requirements. We benchmark against strong baselines, including TinyLLava and LLava-v1.5, spanning multiple model scales to demonstrate the generality and robustness of our approach.

Datasets. The pre-training and fine-tuning datasets used in this work are identical to those utilized in LLava-v1.5. For pre-training, we use a subset of the LAION/CC/SBU dataset filtered for balanced concept coverage and enriched with BLIP-generated captions. For instruction tuning, we use a combination of COCO(Lin et al., 2014), GQA(Hudson and Manning, 2019), OCR-VQA(Mishra et al., 2019), TextVQA(Singh et al., 2019), and VisualGenome(Krishna et al., 2017) datasets. The details of the datasets are in the Appendix.

Tasks and evaluation. We conduct extensive evaluations on a broad spectrum of visual question answering (VQA) benchmarks, encompassing OK-VQA (Marino et al., 2019), GQA (Hudson and Manning, 2019),VQAV2(Goyal et al., 2017), TextVQA (Singh et al., 2019), RealWorldQA (x.ai, 2024), and ScienceQA (Lu et al., 2022). To further assess general multimodal understanding, we evaluate on comprehensive benchmarks such as MM-

Bench (Liu et al., 2025), POPE (Li et al., 2023b), SEED (Li et al., 2023a), MMStar(Chen et al., 2024), AI2D (Kembhavi et al., 2016), MMTB (Ying et al., 2024), and OCR-VQA (Mishra et al., 2019). MME (Fu et al., 2024) is used to evaluating model of granular perception We utilize the `lmms-eval` framework (Zhang et al., 2024a; Li et al., 2024), which integrates multiple benchmark evaluation protocols, ensuring standardized and reproducible performance comparisons.

Models. To comprehensively verify the effectiveness of the VDEP, we employ base models of different parameter scales. Specifically, we utilize TinyLLava (Zhou et al., 2024) and LLava-v1.5 (Liu et al., 2024a) as our base models, whose model sizes are 3B and 7B, respectively. A series of carefully designed experiments are conducted to evaluate their performance. We used SigLip on Tinyllava to verify the capability of our method under MLLMS with different Settings.

Baseline and implementation. To facilitate a fair comparison, we double the input data during pretraining, ensuring both LLava and VDEP receive equivalent training exposure. We introduce a novel special token `<auto_image>` to seamlessly switch between autoregressive image embedding training (VDEP mode) and conventional LLava training. This hybrid training strategy dynamically alternates between the two modes, stabilizing convergence and preventing catastrophic forgetting.

5.1 Empirical results and analysis

Visual question answering performance. Table 2 presents a detailed comparison of VDEP and baseline LLava across six challenging VQA datasets. Our method consistently outperforms the baseline across all datasets and model scales. Notably, LLava-VDEP achieves a substantial +3.36 points(relative 6.28% gain) absolute improvement on OK-VQA and a +1.86 points (3.33% gain) on RealWorldQA with the 7B model, underscoring VDEP's effectiveness in enhancing external knowledge integration and spatial reasoning. Improvements on GQA and ScienceQA further demonstrate enhanced compositional and domain-specific reasoning capabilities. These gains validate that explicitly incorporating visual semantic reconstruction into the training objective significantly bolsters the model's multimodal understanding.

General multimodal benchmark performance. As demonstrated in Table 1 and Table 3, VDEP consistently delivers notable performance improvements across a diverse set of general multimodal tasks. In particular, VDEP achieves substantial gains on MMStar, with llava7b-VDEP improving by 2.58 points, corresponding to a 7.7% relative increase. Given that MMStar primarily evaluates vision-centric capabilities, this underscores VDEP's effectiveness in enhancing visual understanding.

On MME, which encompasses 14 subtasks spanning perception and cognition, TinyLLava-VDEP boosts the overall score by nearly 45 points (+5.9%), while LLava-VDEP attains a gain of +6.93 points. Similarly, on SEED-Bench, a benchmark designed to assess comprehensive visual comprehension, VDEP-augmented models exhibit superior performance in visual reasoning and information integration. Although minor fluctuations are observed in certain subtasks for instance, a slight decrease in counting accuracy for LLava-VDEP the overall trend strongly favors VDEP, indicating enhanced robustness and generalization.

It is worth noting that Our 7B model drops on MMTBench while the 3B model declines on MMBench, mainly because MMBench relies more on text, as prior MMStar(Chen et al., 2024) studies show. Despite this, the 7B model still improves on MMBench, revealing a trade-off between visual and text modal. The two models show opposite trends due to differences in scale and ViT architecture. Although perfectly balancing modalities remains challenging and causes slight bias, VDEP consistently boosts overall benchmark performance, demonstrating improved multimodal alignment. The minor drop on one benchmark further suggests that VDEP effectively mitigates modality imbalance rather than fully resolving it.

Hallucination mitigation and visual attention. VDEP achieves a consistent improvement of 0.4 points in F1 score on the hallucination benchmark POPE for both the 3B and 7B models, demonstrating that our method further strengthens the preservation of visual-semantic fidelity in MLLMs. Complementing this quantitative gain, the qualitative analysis of layer-wise attention maps (Fig. 1 and Appendix Fig. 4) shows that VDEP significantly enhances the model's sensitivity to salient visual cues, such as object boundaries and spatial relationships. This improved visual grounding aligns with the observed reduction in hallucination rates on POPE, providing strong evidence that VDEP's autoregressive latent space alignment effectively addresses modality imbalance and mitigates hallu-

METHOD	PERCEPTION	COMMONSENSE QA (REASONING)	COARSE-GRAINED PERCEPTION TASKS				TOTAL SCORES
			EXISTENCE	COUNT	POSITION	COLOR	
TinyLLava-3B							
TINYLLAVA	1488.30	120.71	185.00	143.33	133.33	180.00	762.37
TINYLLAVA-VDEP (OURS)	**1499.08**	**130.70**	**200.00**	**158.33**	**138.33**	**180.00**	**807.36**
CHANGE	+10.78	+9.99	+15.00	+15.00	+5.00	+0.00	+44.99
LLava-v1.5-7B							
LLAVA	1510.72	135.71	190.00	**158.33**	128.33	175.00	787.37
LLAVA-VDEP (OURS)	**1516.60**	**136.00**	**190.00**	153.30	**135.00**	**180.00**	**794.30**
CHANGE	+5.88	+0.29	+0.00	-5.03	+6.67	+5.00	+6.93

Table 3: Comparison of LLava-VDEP (Ours) and LLava-v1.5 on MME Tasks with Different Model Sizes.

α	RWQA	MME^P	MMB	VQA^{OK}
LLava-VDEP-7B				
w/ 0.1	55.42	1479.00	62.25	57.33
w/ 0.01	56.73	1504.99	62.52	55.70
w/ 0.001	**57.64**	**1515.60**	**62.52**	**57.68**

Table 4: Ablation study on the hyperparameter α

DATA RATIO	RWQA	MME^P	MMB	VQA^{OK}
LLava-VDEP-7B				
w/ 0.5	54.25	1439.16	58.90	55.80
w/ 0.8	57.12	1509.47	59.36	57.26
w/ 1.0	**57.64**	**1515.60**	**62.52**	**57.68**

Table 5: Ablation study on the hyperparameter Data Ratio.

cination in multimodal language models.

5.2 Ablation study

Hyperparameters α. As illustrated in Table 4, with decreasing hyperparameter α, the overall performance of the model exhibits a consistent improvement in performance metrics across multiple benchmarks. This observation suggests that reducing the weight assigned to the image loss notably improves the model's performance. The underlying reason for this phenomenon lies in the disparity between the number of image tokens and text tokens, with the former being significantly larger. This imbalance often leads to a higher proportion of background tokens in image data. When α is relatively large, the model tends to overfit these background tokens, i.e., the model disproportionately focuses on less informative regions of the image, thereby introducing noise that impairs the effectiveness of text alignment during training. By contrast, a smaller α alleviates the constraints of image reconstruction, reducing the influence of background noise and enabling more effective text-image alignment, thereby promoting superior performance in multimodal tasks.

Hyperparameters data ratio. As shown in Table 5, we utilize the VDEP framework to train the model with varying text-to-image data ratios and assess its performance across multiple multimodal benchmarks. By adjusting the ratio of VDEP mode to LLava mode within a batch during pre-training, we control the proportion of image reconstruction data, where a higher ratio indicates a greater amount of image data. The results in Table 5 demonstrate a clear trend: as the proportion of image data increases, the model's overall performance improves consistently across multiple test datasets. This phenomenon is attributed to the greater challenge of simultaneously optimizing regression tasks for both images and text, as it requires balancing competing objectives compared to optimizing only the text regression task. A lack of sufficient image data during pre-training leads to suboptimal learning of all tasks, resulting in weaker alignment between modalities and ultimately degrading the model's overall performance.

6 Conclusion

While previous works have identified the text bias problem in multimodal large language models, we provide a rigorous theoretical analysis revealing that under the assumption of insufficient preservation of visual semantics, the common alignment objective can degenerate into unimodal, text-only training. From an information flow perspective, we further expose the inherent modality imbalance in existing optimization objectives. Therefore, we propose VDEP, a novel framework that explicitly incorporates dynamic visual semantic reconstruction into the training process without modifying model architecture. Experiments on 13 benchmarks show VDEP reduces modality imbalance and boosts visual understanding.

7 Limitation.

Although VDEP exhibits outstanding performance in improving image-text alignment, it relies on the hyperparameter α. While we determine an appropriate range of α for models of varying scales, the optimal value for a given model size remains undetermined. Future work focuses on developing methods to adaptively determine the value of the hyperparameter based on model size and data characteristics. Alternatively, it proposes an effective strategy to eliminate the need for explicit hyperparameter tuning. During pre-training, to improve the effectiveness of image-related tasks while ensuring no degradation in the performance of text-related tasks, we utilize a dataset with double the training samples of the original. As a result, the training time increases by around 3 hours.

Impact Statement

In this paper, we propose a novel paradigm for multimodal alignment, named Vision Dynamic Embedding-Guided Pre-training. Grounded in information theory, this approach incorporates the image reconstruction task as an explicit component of the autoregressive objectives in multimodal large models. This paradigm offers a streamlined and effective framework for aligning MLLMs, emphasizing the critical role and efficacy of image reconstruction in facilitating image-text alignment. The experimental setup and data processing in our study adhere to the principles outlined by the LLava dataset.

References

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. 2024. Unified generative and discriminative training for multi-modal large language models. *arXiv preprint arXiv:2411.00304*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

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

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. 2024. Lmms-eval: Accelerating the development of large multimodal models. * indicates equal contribution.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

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

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Letitia Parcalabescu and Anette Frank. 2024. Do vision & language decoders use images and text equally? how self-consistent are their explanations? *arXiv preprint arXiv:2404.18624*.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, and 1 others. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2(4).

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2024. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2024a. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. 2024b. Reconstructive visual instruction tuning. *arXiv preprint arXiv:2410.09575*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*.

x.ai. 2024. Grok-1.5 vision preview. Accessed: 2025-01-26.

Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, and Wentao Zhang. 2024. Sea: Supervised embedding alignment for token-level visual-textual integration in mllms. *arXiv preprint arXiv:2408.11813*.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, and 1 others. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024b. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*.

Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. 2024. Alignngpt: Multi-modal large language models with adaptive alignment capability. *arXiv preprint arXiv:2405.14129*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

A Implementation Details.

We design a series of experiments to rigorously evaluate the effectiveness of our proposed method across models of varying scales. These experiments involve two models with different parameter sizes: 3B, 7B. For the 3B model, we use the TinyLLava architecture in our experiments. Within this framework, SigLIP is the visual encoder, while Phi-2 is the language model. For the 7B model, we use the pre-trained CLIP ViT-L/14 (336^2) as the visual encoder, combined with the Vicuna v1.5 language model for experiments. pre-training is conducted on the CC-558K dataset with a 1×10^{-3} learning rate. After pre-training, fine-tuning is performed on the mix-665K dataset with a learning rate of 2×10^{-5}. All experiments are conducted on a hardware system with eight NVIDIA A100 GPUs, each with 40GB of memory, to meet the computational requirements. In addition, detailed training steps and specific rules of the implementation plan are fully presented in the appendix. Our training strategy employs a mixed autoregressive pre-training approach with a strict 1:1 ratio of image data to text data. The image data is sourced from the CC-558K dataset, as in pre-training.. During the SFT stage, our experimental settings match the LLava models.

B BenchMarks.

B.1 Visual Question Answering

We conduct experiments on visual question-answering benchmarks, including, OK-VQA, GQA, VQAv2, TextVQA, RealWorldQA, and ScienceQA. OK-VQA includes questions that necessitate external knowledge beyond the multimodal inputs provided. GQA is specifically designed to assess the reasoning capabilities of the model. VQAV2 is one of the most widely used VQA evaluation sets. It covers a wide variety of visual question-answering tasks, and the number of test sets is huge enough to evaluate the visual capabilities of the model very well. TextVQA places a greater emphasis on evaluating the model's ability to comprehend text within natural scenes. RealWorldQA is a benchmark specifically designed to evaluate the spatial understanding capabilities of multimodal AI models in real-world contexts. ScienceQA comprises multimodal multiple-choice questions across a diverse range of science topics. These datasets are strategically selected to evaluate our method's capacity to understand comprehensively and reason across diverse visual contexts and knowledge domains.

OK-VQA: OK-VQA(Outside Knowledge VQA)(Marino et al., 2019) is a visual question answering dataset that requires external knowledge. The answers to the questions cannot be inferred solely from the image but also need to incorporate common sense or world knowledge. This dataset evaluates the model's ability in the intersection of vision and knowledge reasoning.

GQA: GQA(Graph Question Answering)(Hudson and Manning, 2019) generates questions and answers based on image scene graphs, focusing on structured reasoning. It emphasizes logical analysis and challenges the model's depth of understanding of semantics and context.

VQAV2: VQAv2(Goyal et al., 2017) is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer. We used the test split to report Our result.

TextVQA: TextVQA(Singh et al., 2019) focuses on textual information in images, requiring models to recognize and comprehend text within images to answer questions. It drives research on the integration of visual and textual information, expanding the boundaries of visual question answering.

RealWorldQA: RealWorldQA(x.ai, 2024) features images and questions sourced from real-world scenarios, encompassing diverse content from daily life. The dataset imposes higher requirements on the model's generalization ability and adaptability to complex scenes.

ScienceQA: ScienceQA(Lu et al., 2022) is a multimodal question answering dataset combining images and scientific questions, covering multiple scientific topics such as physics and biology. It

bridges AI technology with the field of science education, promoting intelligent question answering applications in educational contexts.

B.2 General Multimodal Benchmarks

We evaluate our proposed method on general multimodal benchmarks, including MME, OCR-Bench, MMBench, SEED-Bench, POPE, AI2D,MMStar, and MMT-Bench. MME measures both perception and cognition abilities on a total of 14 subtasks. MMBench comprehensively evaluates a model's multimodal capabilities in Chinese and English contexts. OCR-Bench contains a collection of 1,000 manually filtered and corrected question-answer pairs, covering five representative text-related tasks. MMStar primarily targets evaluation tasks with a strong reliance on visual information. Candidate samples are initially filtered from existing benchmarks via an automated pipeline, followed by manual verification to ensure that each selected instance exhibits clear visual dependency, minimal data leakage, and requires advanced multimodal reasoning capabilities. SEED-Bench focuses on assessing generative comprehension in multimodal large language models. POPE evaluates the extent of multimodal hallucinations present in a model. AI2D assesses a model's ability to interpret scientific diagram inputs. MM-Vet evaluates the multimodal conversational skills of a model using GPT-4 as a benchmark. MMT-Bench is a comprehensive benchmark developed to evaluate MLLMs across a wide range of multimodal tasks requiring expert knowledge and deliberate visual recognition, localization, reasoning, and planning.

These diverse benchmarks provide a comprehensive framework for evaluating the performance and capabilities of our proposed method in multimodal learning.

MME: MME(Fu et al., 2024), short for Multimodal Evaluation , is a comprehensive multimodal benchmark designed to evaluate the ability of models to understand and process information across multiple modalities, including vision, text, and audio. It provides a standardized framework to measure performance on tasks requiring cross-modal reasoning and understanding, making it an essential tool for assessing the generalization of multimodal large language models (MLLMs).

MMBench MMBench(Multimodal Benchmark)(Liu et al., 2025) is a task-driven benchmark that focuses on systematically evaluating multimodal models across diverse real-world application scenarios, such as visual question answering, image captioning, and video understanding. Its emphasis on practical use cases highlights its importance for assessing the practical utility of MLLMs.

SEED: SEED(Spatial and Entity-aware Evaluation Dataset)(Li et al., 2023a) is a benchmark specifically designed to evaluate the spatial and entity reasoning capabilities of multimodal models. By incorporating complex spatial relationships and entity-based queries, SEED tests a model's ability to perform fine-grained reasoning, which is critical for tasks such as scene understanding and object-oriented question answering.

POPE: POPE(Perceptual and Object-aware Performance Evaluation)(Li et al., 2023b) focuses on evaluating the perceptual understanding and object-centric reasoning of multimodal models. It emphasizes tasks like object detection, recognition, and spatial awareness, making it a key benchmark for assessing models' performance in visually grounded tasks.

AI2D: AI2D(Allen Institute for AI Diagram Dataset)(Kembhavi et al., 2016) is a dataset centered on diagram understanding, designed to evaluate models' abilities to process non-photographic visual content. It focuses on reasoning over diagrams and charts, making it vital for tasks requiring scientific and technical visual comprehension.

OCRB: OCRB (Optical Character Recognition Benchmark)(Mishra et al., 2019) is a specialized benchmark for assessing a model's ability to recognize and interpret text in images. It focuses on OCR-related tasks, such as text detection, transcription, and contextual understanding, which are crucial for applications like document analysis and scene-text understanding.

MMStar: MMStar(Chen et al., 2024), an elite vision-indispensable multi-modal benchmark comprising 1,500 samples meticulously selected by humans. MMStar benchmarks 6 core capabilities and 18 detailed axes, aiming to evaluate LVLMs' multi-modal capacities with carefully balanced and purified samples.

MMMU: MMMU(Multimodal Multitasking Understanding)(Yue et al., 2024) evaluates the multitasking capabilities of multimodal models by testing their performance on multiple simultaneous tasks across different modalities. This benchmark is essential for assessing the adaptability and efficiency of models in dynamic, multitask scenarios.

MMTB: MMTB(Multimodal Task Bench-

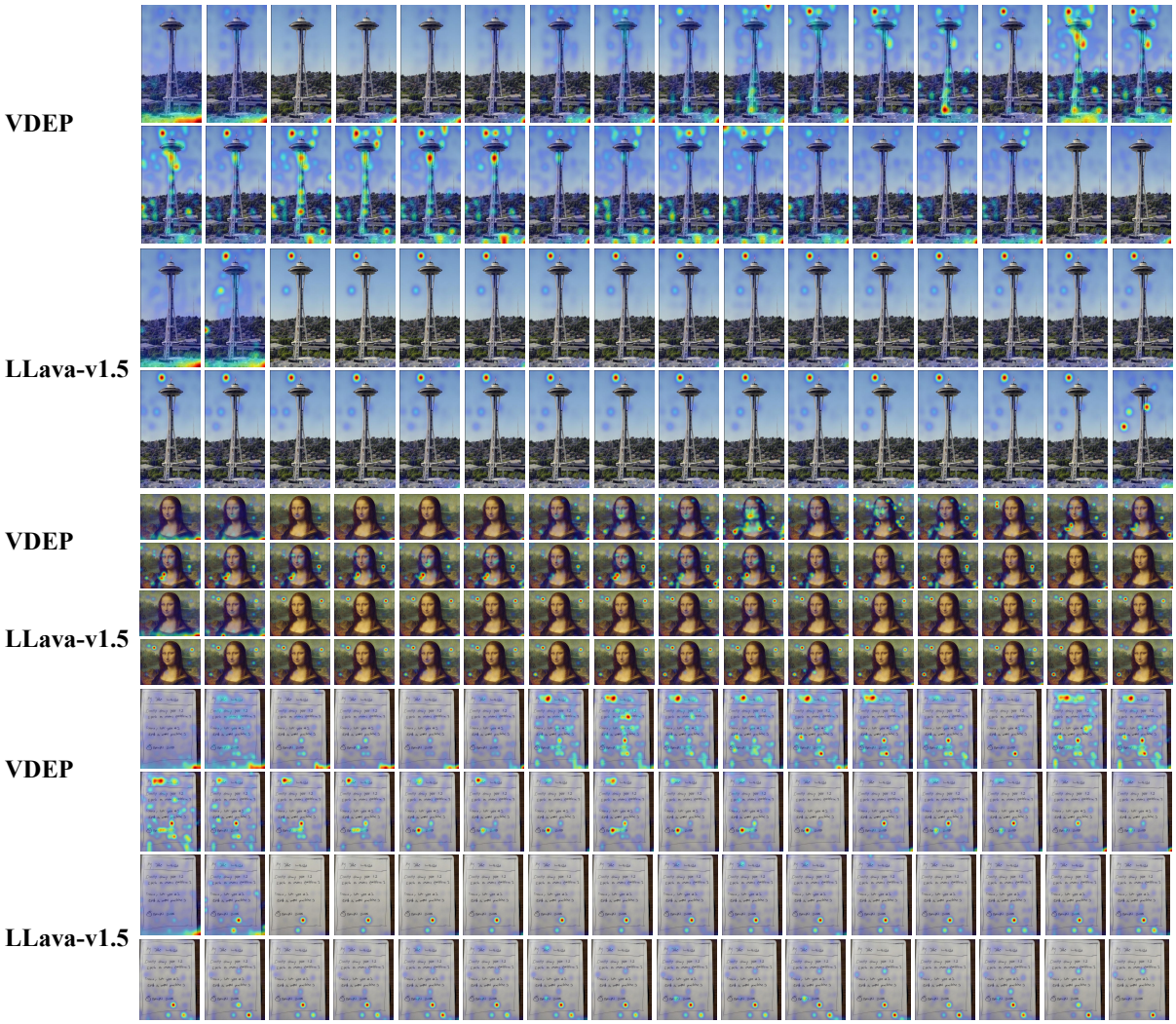

Figure 4: Layer-wise attention visualization of visual-to-instruction information flow. Displayed from top to bottom are the attention heatmaps from LLava-v1.5-7B and LLava-v1.5-7B-VDEP, respectively. The example is derived from LLava-Bench (Liu et al., 2024b) and the query is "Describe this photo in detail".

mark)(Ying et al., 2024) is a broad benchmark designed to evaluate the performance of multimodal models on a wide range of tasks, including vision-and-language navigation, multimodal reasoning, and image captioning. Its diversity makes it a strong indicator of a model's overall multimodal proficiency.

OCRB: OCRB (Optical Character Recognition Benchmark)(Mishra et al., 2019) is a specialized benchmark for assessing a model's ability to recognize and interpret text in images. It focuses on OCR-related tasks, such as text detection, transcription, and contextual understanding, which are crucial for applications like document analysis and scene-text understanding.

C Case Study

As shown in 4, we can see from the visualization of cases in different scenarios that the VDEP method significantly enhances the MLLM's perception of fine-grained content in vision, and the visual semantics are retained in depth at different layers of the LLM. This is also consistent with our theoretical verification.

D Proof: Using $\mathcal{L}_2$ Distance to Estimate Mutual Information Between Embeddings

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables representing embedding vectors in a continuous space. We aim to show that minimizing the squared $\mathcal{L}_2$ distance between $\mathbf{X}$ and $\mathbf{Y}$ can serve as a proxy to maximizing their mutual information $\mathcal{I}(\mathbf{X}; \mathbf{Y})$.

D.1 Mutual Information Definition

Recall that the mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathcal{I}(\mathbf{X};\mathbf{Y}) = \mathcal{H}(\mathbf{X}) - \mathcal{H}(\mathbf{X}|\mathbf{Y}),$$

where $\mathcal{H}(\mathbf{X})$ is the differential entropy of $\mathbf{X}$, and $\mathcal{H}(\mathbf{X}|\mathbf{Y})$ is the conditional differential entropy of $\mathbf{X}$ given $\mathbf{Y}$.

D.2 Conditional Entropy and Reconstruction Error

Intuitively, if $\mathbf{Y}$ can reconstruct $\mathbf{X}$ with high fidelity, then the uncertainty of $\mathbf{X}$ given $\mathbf{Y}$ is low, i.e., the conditional entropy $\mathcal{H}(\mathbf{X}|\mathbf{Y})$ is small.

Suppose we measure the reconstruction error between $\mathbf{X}$ and $\mathbf{Y}$ by the expected squared Euclidean ($\mathcal{L}_2$) distance:

$$\mathcal{L} = \mathbb{E}\left[\|\mathbf{X} - \mathbf{Y}\|_2^2\right].$$

When $\mathcal{L}$ approaches zero, it means $\mathbf{Y}$ nearly perfectly reconstructs $\mathbf{X}$, and thus the uncertainty of $\mathbf{X}$ given $\mathbf{Y}$ becomes very small.

D.3 Linking $\mathcal{L}_2$ Distance to Conditional Entropy

To make this intuition more precise, assume the reconstruction error $\mathbf{X} - \mathbf{Y}$ follows a Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}$, i.e.,

$$p(\mathbf{X}|\mathbf{Y}) = \mathcal{N}(\mathbf{Y}, \mathbf{\Sigma}).$$

Under this assumption, the conditional differential entropy of $\mathbf{X}$ given $\mathbf{Y}$ is

$$\mathcal{H}(\mathbf{X}|\mathbf{Y}) = \frac{1}{2}\log\left((2\pi e)^d \det \mathbf{\Sigma}\right),$$

where d is the dimensionality of $\mathbf{X}$.

Since the expected squared error $\mathcal{L}$ equals the trace of the covariance matrix,

$$\mathcal{L} = \mathrm{Tr}(\mathbf{\Sigma}),$$

minimizing $\mathcal{L}$ corresponds to reducing the overall variance of the reconstruction error.

As $\mathcal{L} \to 0$, the covariance matrix $\mathbf{\Sigma}$ approaches the zero matrix, and thus $\det \mathbf{\Sigma} \to 0$. Because the logarithm of the determinant tends to negative infinity, the conditional differential entropy satisfies

$$\mathcal{H}(\mathbf{X}|\mathbf{Y}) \to -\infty.$$

This reflects a fundamental property of differential entropy: when a continuous distribution becomes degenerate (variance tends to zero), its differential entropy tends to negative infinity. In other words, the uncertainty of $\mathbf{X}$ given $\mathbf{Y}$ vanishes in the limit of perfect reconstruction.

D.4 Interpretation and Practical Implications

Although the conditional differential entropy tends to negative infinity mathematically, this corresponds to the intuitive notion that the uncertainty of $\mathbf{X}$ given $\mathbf{Y}$ becomes negligible. In practice, this means that minimizing the reconstruction error $\mathcal{L}$ effectively reduces the conditional entropy $\mathcal{H}(\mathbf{X}|\mathbf{Y})$.

Since mutual information can be expressed as

$$\mathcal{I}(\mathbf{X};\mathbf{Y}) = \mathcal{H}(\mathbf{X}) - \mathcal{H}(\mathbf{X}|\mathbf{Y}),$$

reducing $\mathcal{H}(\mathbf{X}|\mathbf{Y})$ by minimizing $\mathcal{L}$ increases the mutual information $\mathcal{I}(\mathbf{X};\mathbf{Y})$.

D.5 Conclusion

Therefore, minimizing the squared $\mathcal{L}_2$ reconstruction loss between embeddings $\mathbf{X}$ and $\mathbf{Y}$ serves as a practical surrogate for maximizing their mutual information. Formally,

$$\lim_{\mathcal{L}\to 0} \mathcal{I}(\mathbf{X};\mathbf{Y}) = \mathcal{H}(\mathbf{X}),$$

which corresponds to the maximal mutual information achievable when $\mathbf{X}$ is fully determined by $\mathbf{Y}$.

This justifies the use of the squared $\mathcal{L}_2$ distance as a feasible and effective loss function to estimate and maximize mutual information between embedding vectors.

E Detailed experiments.

We present comprehensive ablation results derived from LLava-v1.5 to substantiate the experimental conclusions in the main text. Additionally, we performed ablation studies on the image loss function to demonstrate the simplicity and effectiveness of the L2 loss.

Data Ratio	AI2D	MM-Vet	MMMU	MMT	GQA	VizWizQA	VQA^{T}	SQA^{I}
LLava-v1.5-VDEP-7B								
w/ 0.5	54.02	29.00	31.20	46.30	61.65	49.82	46.33	68.62
w/ 0.8	55.18	28.20	**31.30**	46.72	61.65	45.40	46.27	**69.16**
w/ 1.0	**56.57**	**30.60**	30.80	**48.00**	**62.50**	**50.37**	**46.76**	68.36

Table 6: Ablation study on the hyperparameter Data Ratio, which represents the proportion of different VDEP and LLava patterns in the pre-training stage.

α	AI2D	MM-Vet	MMMU	MMT	GQA	VizWizQA	VQA^{T}	SQA^{I}
LLava-v1.5-VDEP-7B								
w/ 0.1	55.57	30.50	30.60	47.64	61.45	46.76	46.52	67.72
w/ 0.01	**56.64**	**32.20**	**31.30**	**48.48**	**62.63**	**52.72**	46.94	67.77
w/ 0.001	56.57	30.60	30.80	48.00	62.50	50.37	**46.76**	**68.36**

Table 7: Ablation study on the hyperparameter α, which represents the variation of the image loss weight.

Loss	Perception	Commonsense QA	Coarse-grained Perception Tasks				Total
		(Reasoning)	Existence	Count	Position	Color	Scores
LLava-v1.5-VDEP-7B							
1/L2	**1518.34**	133.57	190.00	**163.33**	135.00	**180.00**	**801.90**
Sigmoid(L2)	1478.45	133.57	190.00	145.00	**138.33**	175.00	781.90
L2	1515.60	**136.00**	190.00	153.30	135.00	**180.00**	794.30

Table 8: Ablation study on the hyperparameter Loss Function on MME.

Loss	VQA^{OK}	GQA	VizWiz	VQA^{T}	RWQA	SQA^{I}
LLava-v1.5-VDEP-7B						
1/L2	56.11	62.47	**51.37**	46.56	54.38	**69.01**
Sigmoid(L2)	57.37	**62.95**	49.87	46.67	**57.90**	68.32
L2	**57.68**	62.50	50.37	**46.76**	57.64	68.36

Table 9: Ablation study on the hyperparameter Loss Function on VQA.

Loss	MMBench		AI2D	MM-Vet	MMMU	MMTB	OCRB	POPE
	EN	CN						
LLava-v1.5-VDEP-7B								
1/L2	65.97	**58.52**	**57.09**	31.10	**31.20**	47.93	320	85.62
Sigmoid(L2)	66.20	58.24	56.47	**31.70**	31.00	**48.32**	**334**	**85.98**
L2	**66.81**	58.23	56.57	30.60	30.80	48.00	326	85.95

Table 10: Ablation study on the hyperparameter Loss Function on benchmarks for insruction-following LMMs.