DeepAlign: Mitigating Modality Conflict through Modality-Specific Alignment

Shuo Li[1*] Bingchen Miao[2*] Wendong Bu[2*] Juncheng Li[2✉] Hanwang Zhang[1] Fei Wu[2]

[1] Nanyang Technological University, [2] Zhejiang University

SHUO008@e.ntu.edu.sg, hanwangzhang@ntu.edu.sg

{miaobingchen23, wendongbu, junchengli, wufei}@zju.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated promising advancements in augmenting the capabilities of LLMs to comprehend visual input. However, modality misalignment between vision and text remains a key challenge in MLLM, which can be attributed to two aspects: misalignment of modality-specific representations and depletion of modality-specific details. To address the issue of modality misalignment, we propose ***DeepAlign****, a novel multimodal alignment framework to mitigate modality conflict, which employs representation intervention and structure-induced knowledge distillation to prevent the misalignment and depletion of modality-specific information. Extensive experiments demonstrate that DeepAlign significantly mitigates modality conflicts, leading to substantial performance improvements compared to backbone models across multiple vision-language tasks. It also stimulates some emergent abilities in MLLMs, such as multimodal in-context learning on interleaved text-image sequences.*

1. Introduction

Multi-modal Large Language Models (MLLMs) [8, 20, 55] have shown promising advancements in augmenting the capabilities of LLMs to comprehend visual input. Mainstream MLLMs [9, 24, 64] often comprise a pre-trained vision encoder a pre-trained LLM, following a "bridging" paradigm that leverages a lightweight intermediate connection module to align visual features with the textual modality. This architecture enables LLMs to efficiently acquire the ability to tackle a spectrum of multimodal tasks.

However, a natural question arises: *Is this modality alignment approach sufficient for an LLM, originally designed to handle linguistic tasks, to comprehensively grasp the visual information?* Unfortunately, HallusionBench [17] reveals that existing MLLMs exhibit cognitive biases towards the input vision information, leading to misinterpretations that deviate from reality. Additionally, LaVIT [58] argues that the reliance on intermediate connection modules for vision-language alignment is insufficient to leverage the superior reasoning capability of LLMs in the realm of multimodal comprehension. Overall, the visual and textual modalities within current MLLMs are poorly integrated, resulting in modality conflicts [2, 4, 65].

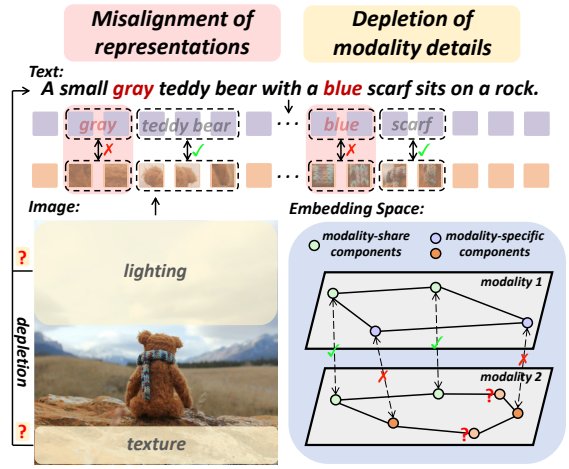

Figure 1. Illustration of two issues that exacerbate modality conflict: 1) Misalignment of modality-specific representations, 2) Depletion of modality-specific details. These two types of conflict can be abstractly represented in the embedding space as shown in the bottom-right corner of the figure.

We posit that the modality conflict primarily stems from a focus on modality-shared information, neglecting the modality-specific knowledge that plays a crucial role in enhancing multimodal comprehension. This issue is further exacerbated by the following two primary concerns:

(1) Misalignment of modality-specific representations: The visual input of the LLM often lacks textual equivalents [39, 50], and current vision-language pretraining emphasizes aligning modality-shared semantics [22, 31, 51] (*e.g.*, generating a short caption for the image), leaving modality-specific information misaligned [49]. For instance, as shown in Fig. 1, though high-level semantics such as "teddy bear" and "scarf" are correctly aligned, the

* Equal Contribution.
✉ Juncheng Li is the Corresponding Author.

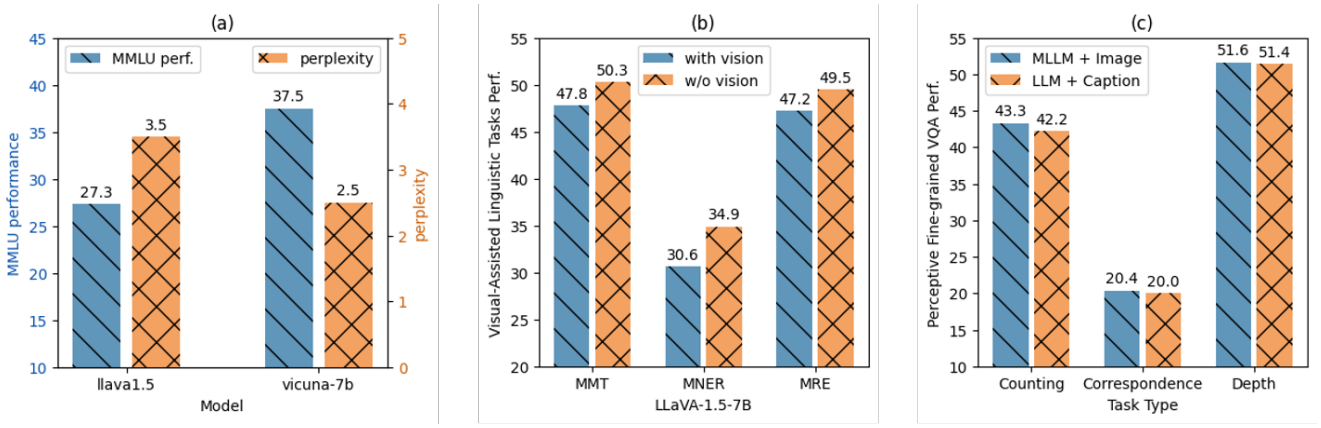

Figure 2. (a) Compare the performance and perplexity of LLaVA-v1.5 and Vicuna on linguistic tasks. (b) The performance of LLaVA-v1.5 on visual-assisted task with/without vision as the input. (c) The performance on fine-grained perceptive VQA tasks with/without replacing the image with its dense caption.

specific attributes like color ("gray" and "blue") are misaligned. This misalignment prevents the visual modality from synergizing with the textual modality and instead impairs the LLM's linguistic reasoning capabilities.

(2) Depletion of modality-specific details: The autoregressive training of multimodal models [12, 52] primarily relies on simple text-side supervision, where losses are computed only for predicting language responses, with visual data reduced to a supplementary role. As a result, the MLLM's comprehension of vision is limited to a modality-shared level, preventing it from capturing visual details that are difficult to articulate in text. As shown in Fig. 1, the complex texture of the rock and the bright lighting of the sky are not described in the accompanying text, representing a loss of visual-specific details. This limitation restricts MLLM's multimodal reasoning capabilities in complex scenarios that require fine-grained perception of images.

We conducted a series of exploratory experiments to substantiate our analysis of the modality conflict. Experimental details are provided in Section 2.

(1) As shown in Fig. 2 (a), the performance of text-only comprehension tasks decreases significantly upon the transition of an LLM (*i.e.*, Vicuna [7]) to an MLLM (*i.e.*, LLaVA-v1.5 [30]). And multimodal pre-training also leads to a notable increase in the perplexity of textual token prediction [6, 41]. These findings suggest that current MLLMs fails to seamlessly integrate the visual modality, adversely impacting the linguistic reasoning capabilities of LLMs.

(2) As shown in Fig. 2 (b), the incorporation of visual modality does not assist the MLLM to more effectively perform visual-assisted linguistic tasks (MNER [47], MMT [54], and MRE [19]). Further evidence is illustrated in Fig. 2 (c), replacing the raw visual image with a comprehensive image caption and forcing the LLM to conduct visual-blind reasoning with the caption, does not lead to a notable drop in perceptive VQA tasks. This implies that the comprehension capability of MLLMs remains at a modality-shared level with the depletion of modality-specific details, which limits its performance in scenarios that require fine-grained perception of vision [34, 46].

To address these challenges, we propose **DeepAlign**, a novel multimodal post-training framework to mitigate modality conflict from dual-levels. To avoid misalignment in modality-specific representations, we adopt representation intervention to bridge the gap caused by modality-specific heterogeneity. By identifying and isolating the modality-specific components within both visual and textual representations, we first quantify the underlying misalignment. And then through internal representation regulation, facilitated by a lightweight adapter, we harmonize the modality-specific components of visual representations with the LLM's textual embedding space. It preliminarily alleviates the misalignment, laying a solid foundation for effective cross-modal integration.

To prevent depletion of modality-specific details, we propose a structure-induced approach, distilling structural knowledge from DINOv2 [10, 38] into MLLM as vision-side supervision.As a vision-only self-supervised model trained without linguistic cues, DINOv2 has enhanced ability to capture visual details with structural awareness.To this end, we enforce semantic interrelations from DINOv2 as an induced structural constraint upon post-MLLM hidden state. We utilize patch-level similarity between DINO embeddings of two intra-image patches to supervise semantic similarity of corresponding post-MLLM visual hidden states.Under this paradigm, MLLM training is no longer limited to text-side supervision. Visual modality also receives its own supervision, enabling the MLLM to apprehend structural semantics within images and perceive more reasoning-aware visual details.

Extensive experiments demonstrate that DeepAlign effectively alleviates the modality conflict and enhances the performance of MLLMs across various vision-language benchmarks. Furthermore, it endows MLLMs with several

emergent capabilities, such as in-context learning and fine-grained demonstrative reasoning.

Overall, our main contributions are three-fold:

- Through quantitative exploratory analysis, we reveal the phenomenon of modality conflict in current MLLMs from two primary aspects.
- We propose DeepAlign, a novel multimodal post-training framework to mitigate modality conflict, which employs representation intervention and structure-induced knowledge distillation to prevent the misalignment and depletion of modality-specific information.
- Experimental results indicate that DeepAlign effectively alleviates modality conflict and enhances the performance of MLLMs in various vision-language tasks.

2. Exploring Modality Conflict

In this section, we describe the details of exploratory experiments. To demonstrate the effects of modality misalignment, we compare the performance of MLLM with its corresponding LLM on linguistic tasks. To illustrate depletion of modality-specific details, we further experiment on visual-assisted linguistic tasks and perceptive VQA tasks.

Linguistic Tasks. We compare the performance of MLLM against its corresponding LLM on the MMLU [18] benchmark for linguistic reasoning. We utilize LLaVA-v1.5-7B [30] as the MLLM, and Vicuna-7B [7] as the LLM. The overall performance is shown in Fig. 2 and LLaVA-v1.5-7B significantly underperforms Vicuna-7B with an average performance decline of 10 points, highlighting the trade-off between multimodal capabilities and pure linguistic reasoning. Furthermore, we randomly sample 1000 text segments from Wikipedia to assess the perplexity of text-token-prediction for both the LLM and MLLM. As illustrated in Fig. 2 (a), the perplexity of LLaVA-v1.5-7B is significantly higher than that of its corresponding LLM. The integration of visual information appears to interfere with the model's text processing capabilities, indicating that after multimodal pretraining, there is a misalignment between the visual and textual modalities. The introduction of vision has disrupted the model's original text comprehension abilities, failing to achieve synergistic gains between different modalities.

Visual-Assisted Linguistic Tasks. Next, we evaluate the performance of MLLM on three visual-assisted linguistic tasks: Multi-modal Named Entity Recognition (MNER) [47], Multimodal Machine Translation (MMT) [54], and Multimodal Relation Extraction (MRE) [19]. We conduct these tasks under two conditions: one using standard multimodal inputs and another using text-only inputs for LLaVa-v1.5. While theoretically these tasks can be completed without the need for images, the incorporation of visual input provides additional contextual insights, which, in theory, can enhance the performance.

However, Fig. 2 (b) shows that MLLM performs worse on all three tasks when provided with multimodal input compared to text-only input. Rather than offering additional insights, the incorporation of visual context appears to have a detrimental effect on MLLM's reasoning capabilities, potentially due to the model's inability to align and integrate multimodal information. This result underscores that MLLM fails to effectively capture the modality-specific details presented in the image, highlighting a limitation in current multimodal learning frameworks.

Perceptive VQA tasks. We further explore how MLLMs integrate visual information in perceptive fine-grained VQA tasks, which are sourced from BLINK [14] benchmark. We focus on three sub-tasks: complex object counting, visual correspondence, and relative depth. These tasks are difficult to solve by solely reducing the evaluation into text-only questions using dense captioning; Instead, they necessitate that the model perceives the content of the raw image to provide accurate answers. In addition to prompting LLaVA-v1.5 [30] to tackle the task in a standard setup, we also replace the input image with textual captions that we obtain by using the BLIP-2 [23] model and instruct a text-only LLM (*i.e.*, Vicuna-7b) [7] to solve the task.

As depicted in Fig. 2 (c), compelling LLM to perform VQA task in a "visual-blind way" does not result in significant performance decrease. This suggests MLLMs' visual comprehension capability remains at a modality-shared level, incapable of perceiving information hard to articulate in text. The lack of visual details, crucial for semantically linking instructions, hinders MLLM's ability to complete fine-grained perceptive multimodal tasks.

3. Method

In this section, we introduce DeepAlign, a novel post-training framework that employs representation intervention and structure-induced knowledge distillation to mitigate the misalignment and depletion of modality-specific information, as illustrated in Fig. 3. For each component, we first outline empirical observations justifying its necessity, then systematically detail the methodology.

3.1. Representation Intervention

Observation. Current vision-language pre-training focuses on learning modality-shared semantics, while modality-specific components remain misaligned. To visualize this misalignment, we select a set of image-text pairs and utilize LLaVA-v1.5 [30] as the backbone to extract their post-MLLM visual and textual hidden states, respectively. We then employ UMAP for dimensionality reduction to visualize the embedding distributions of both textual and visual representations. As shown in Fig. 4 (a), the represen-

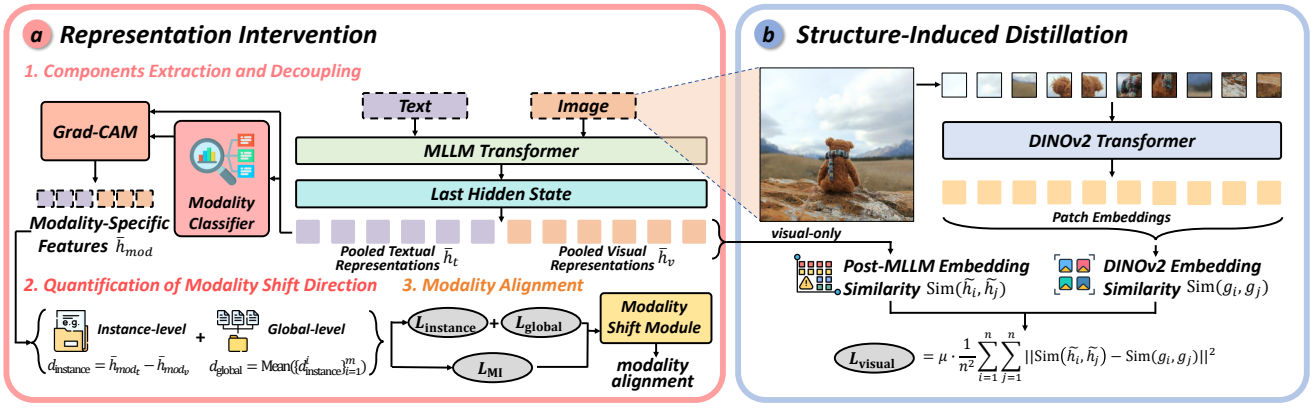

Figure 3. Overview of DeepAlign: 1) **Representation Intervention** that bridges the modality gap by harmonizing visual and textual embedding spaces; and 2) **Structure-Induced Distillation** distills structural knowledge from DINOv2 into the MLLM .

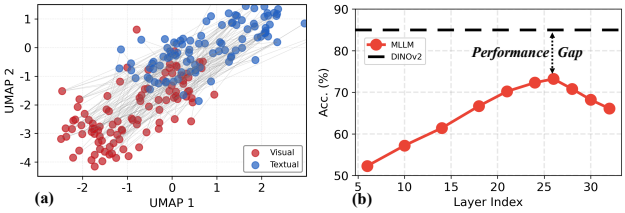

Figure 4. (a) Visualization of the embedding distribution for post-MLLM textual and visual hidden states. (b) Linear probing on ImageNet with MLLM hidden states and DINOv2 embeddings.

tations are distinctly separated, revealing a clear misalignment between the two modalities within the MLLM.

Methodology. To mitigate misalignment between modality-specific representations, we propose a representation intervention framework that bridges the gap by harmonizing visual and textual embedding spaces. Our approach involves three key stages: (1) extraction and decoupling of modality-specific components, (2) quantification of modality shift directions, and (3) alignment via a trainable modality shift module.

First, for a set of image-text pairs, we extract pooled visual and textual representations $(\bar{h}_t, \bar{h}_v) \in R^D$ (D is the embedding dimension) from the internal hidden states of the last token in each MLLM layer. Based on these visual and textual representations, we first train a modality classifier f to distinguish whether a given representation originates from the visual or textual modality. f outputs a predicted score $y = f(\bar{h})$ for each representation $\bar{h}$, indicating the probability that it is derived from the text or the visual modality.

Then motivated by Grad-CAM [42], which is a widely used technique to localize the most important features for classification, we propose to use the gradients of the predicted score $w^{cls} = \nabla_{\bar{h}} y_k = \frac{\partial y_k}{\partial \bar{h}}$ corresponding to the ground-truth modality k as the attention weights to obtain the modality-discriminative (modality-specific) features in channel wise:

$$\bar{h}_{\text{mod}} = sw^{cls} \odot \bar{h}, \tag{1}$$

where $\odot$ represents the Hadamard product, and sw^{cls} is the attention weight vector with s as an adaptive non-negative parameter to modulate the energy $\epsilon(\bar{h}_{mod}) = ||\bar{h}_{mod}||_2^2$ such that $\epsilon(\bar{h}_{mod}) = \epsilon(\bar{h})$:

$$s = \sqrt{\frac{||\bar{h}||_2^2}{||w^{cls} \odot \bar{h}||_2^2}} = \sqrt{\frac{\sum_{d=1}^{D} \bar{h}_d^2}{\sum_{d=1}^{D} (w_d^{cls} \bar{h}_d)^2}} \tag{2}$$

After extracting modality-specific features $\bar{h}_{mod}$, we capture modality shift direction at both instance and global levels. The instance-level modality shift direction is calculated as $d_{\text{instance}} = \bar{h}_{mod_t} - \bar{h}_{mod_v}$. We then aggregate all these instance-level directions to capture the global trend of modality shift across multiple samples, with $d_{\text{global}} = \texttt{Mean}(\{d^i_{\text{instance}}\}^m_{i=1})$.

Combining these two-level shift directions, we can achieve a more comprehensive correction of modality-specific misalignments. Specifically, we further propose a simple but effective modality shift module $\texttt{SHF}(\cdot)$ to align the modality-specific components of visual representations with the LLM's textual embedding space. The modality shift module is implemented by inserting adapter layers into several intermediate transformer layers, which transforms the internal visual representation h_v into a shifted representation $h'_v = \texttt{SHF}(h_v)$ to eliminate the misalignment of modality-specific components. During training, with the previously mentioned text-image pairs, to guide the module in learning the correct modality shift, we construct both global-level and instance-level supervision signals based on the modality shift directions:

$$\begin{aligned} \mathcal{L}_{\text{global}} &= \alpha(1 - \cos_{\text{sim}}(h'_v[-1] - h_v[-1], d_{\text{global}})) \\ \mathcal{L}_{\text{instance}} &= \beta(1 - \cos_{\text{sim}}(h'_v[-1] - h_v[-1], d_{\text{instance}})) \end{aligned} \tag{3}$$

where $h_v[-1]$ is the pooled visual representations from last-visual-token hidden states. Besides, we introduce an additional regularization term to minimize the change in mutual

information between the visual representations $h_v(h'_v)$ and the corresponding text representations h_t before and after modality shifting, which ensures that the shifted representation h'_v aligns more closely with the textual modality while preserving critical information.

$$\mathcal{L}_{\text{MI}} = -\gamma(\text{MI}(h'_v, h_t) - \text{MI}(h_v, h_t)) \quad (4)$$

where α, β and γ are hyperparameters.

3.2. Structure-Induced Distillation

Oberservation. Current multimodal pretraining relies on text-side supervision, lacking supervision over visual semantics. To further investigate how visual features evolve progressively and semantically during MLLM decoding, we leverage linear probing for MLLM internal visual hidden states and DINOv2 representations for ImageNet classification task [11]. As shown in Fig. 4 (b), classification performance achieves a peak when leveraging intermediate layer hidden states for probing, but remains significantly lower than that of DINOv2, indicating a substantial semantic gap. It suggests that MLLMs prioritize modality-shared semantics focuses on text-level, neglecting semantically-rich visual-specific representations needed for image perception. This highlights the necessity to incorporate visual supervision during training to bridge this semantic gap.

Methodology. To percept modality-specific visual details, we propose a structure-induced approach that distills DINOv2's structural knowledge into the MLLM as visual supervision. Trained without linguistic cues, DINOv2 [38] — a vision-only self-supervised model—excels at capturing granular visual details and refined structural awareness. We aim to use the patch-level similarity of DINO embeddings within intra-image patches to guide the semantic similarity of the corresponding post-MLLM visual hidden states, thereby injecting additional fine-grained visual semantics.

Formally, let $\texttt{Enc}$ denote the DINOv2 encoder. Given an image $\mathcal{I}$, we obtain the encoder output $g = \texttt{Enc}(\mathcal{I}) \in R^{n \times D}$ and the post-MLLM hidden states $\widetilde{h} = \texttt{MLLM}(\mathcal{I}) \in R^{n \times D'}$, where n, D, and D' represent the number of patches in the image, the embedding dimension of the encoder, and the embedding dimension of the MLLM, respectively. For both g and $\widetilde{h}$, we can distinctively derive the similarity matrices between any two patches. To impose semantic relationships from DINOv2 as an induced structural constraint on the MLLM hidden states, we apply a Mean Squared Error (MSE) loss to align the post-MLLM similarity matrix with the post-DINOv2 matrix. Thus, we define the visual supervision loss as follows:

$$\mathcal{L}_{\text{visual}} = \mu \cdot \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} ||\texttt{Sim}(\widetilde{h}_i, \widetilde{h}_j) - \texttt{Sim}(g_i, g_j)||^2 \quad (5)$$

where i, j are the patch index, $\texttt{Sim}(\cdot, \cdot)$ is a pre-defined similarity function, and μ is the hyperparameter.

3.3. Final Training Loss

Finally, we combine the three supervision signals (Eq. 3, 4) for representation intervention and the visual supervision loss (Eq. 5) with the conventional text autoregressive loss: $\mathcal{L}_{\text{AR}} = \sum \log P(t_i | t_1, ..., t_{i-1}; \mathcal{I})$ for optimization (t denotes the text token). The modality-shifting module (*i.e.*, the inserted adapter layers) serves as the only trainable parameters. The final training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{AR}} + \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{instance}} + L_{\text{MI}} + \mathcal{L}_{\text{visual}} \quad (6)$$

This loss function balances textual and visual optimization, enabling effective modality alignment.

4. Experiments

Implementation Details. DeepAlign is a model-agnostic method, and we apply it to LLaVA-v1.5-7B [31], Qwen2.5-VL-7B, and InternVL3-8B to explore the broad applicability of our method. The text-image pairs used for post-training are selected from a high-quality subset of CC3M [21] and COCO Caption [5]. All post-training data are derived from the original pre-training and instruction-tuning datasets of the backbone multimodal large language model (MLLM). During post-training, the only trainable parameters are those of the adapter layers (*i.e.*, modality shift module, 200M params), with a peak learning rate of 3×10^{-5}. More details are shown in Appendix A.

Experimental Setup. After successful post-training with DeepAlign, we perform a comprehensive comparison against SOTA baselines on a diverse set of vision-language tasks under a zero-shot setting, as detailed in Section 4.1. In Section 4.2, we re-conduct experiments mentioned in Section 2 to validate that DeepAlign can effectively mitigate modality conflicts. In Section 4.3 we demonstrate that DeepAlign could enable MLLM to exhibit emergent abilities. Finally, we conduct in-depth analysis to thoroughly evaluate the effectiveness of our method in Section 4.4. See more details in Appendix B.

4.1. Zero-shot Vision-Language Comprehension

We first evaluate the effectiveness of DeepAlign on a range of state-of-the-art academic benchmarks, including MLLM Benchmarks (MMBench [33], MMStar [3], MMMU [57], HallusionBench [16], OCRBench [35], MMVet [56]) and VQA Benchmarks (ScienceQA [36], TextVQA [43], RealWorldQA, MTVQA [44]). We comprehensively compare DeepAlign with several baselines also dedicated to achieving better modality alignment through post-training, including RLHF [28], HADPO [62], DataTailor [55], POVID [63], SIMA [48], and VISTA [26].

We report the results of DeepAlign and baseline methods in Table 1. Based on the observation of experimental results, we have summarized the following conclusions:

Table 1. Results of DeepAlign and baselines on MLLM-oriented comprehension, and VQA benchmarks. **Bold** values and underlined values represent the best performance and second best performance across all post-training methods.

Methods	MLLM Benchmarks						VQA Benchmarks			
	MMBench	MMStar	MMMU	HallusionBench	OCRBench	MMVet	ScienceQA	TextVQA	RealWorldQA	MTVQA
GPT-4.5	83.4	69.3	72.1	60.0	845	75.3	/	/	/	/
Gemini-1.5-Pro	82.8	67.1	68.6	55.9	770	74.6	/	/	71.1	/
GPT-4o	84.3	65.1	70.7	56.2	806	74.5	90.7	70.3	76.5	31.2
LLaVA-v1.5-7B	62.1	34.6	33.7	25.2	385	32.2	69.2	49.7	54.8	6.0
+RLHF	62.3	34.8	33.9	25.1	387	32.5	69.5	49.9	55.0	6.1
+HADPO	62.5	36.2	33.6	25.8	390	32.8	70.0	51.3	55.2	6.3
+DataTailor	62.0	34.9	34.0	25.3	386	33.5	69.1	49.8	54.7	7.5
+POVID	63.0	35.5	35.3	26.2	411	33.4	70.5	50.5	55.8	6.8
+SIMA	64.3	34.8	33.5	27.4	402	32.8	71.1	50.2	55.5	6.4
+VISTA	65.4	35.9	37.5	28.9	410	34.1	72.2	52.6	57.4	8.0
+DeepAlign	**68.2**	**40.1**	**39.4**	**31.0**	**427**	**38.5**	**75.3**	**55.7**	**60.2**	**9.3**
Qwen2.5-VL-7B	82.2	64.1	58.0	51.9	888	69.7	89.0	76.7	66.8	29.0
+RLHF	82.3	64.5	58.1	52.0	887	69.8	89.2	76.8	66.9	29.1
+HADPO	82.3	64.2	57.9	51.8	887	69.8	88.8	76.6	66.7	28.9
+DataTailor	82.1	64.6	58.2	51.7	890	69.6	88.9	76.5	66.7	28.8
+POVID	82.6	64.3	59.0	52.5	889	70.2	89.8	77.2	67.5	29.8
+SIMA	82.4	65.2	58.8	52.3	890	70.5	90.0	77.5	67.8	29.5
+VISTA	82.8	65.5	59.2	52.8	889	70.4	90.2	77.8	68.0	30.1
+DeepAlign	**83.2**	**65.7**	**59.5**	**53.6**	**894**	**71.0**	**90.5**	**78.4**	**68.1**	**30.8**
InternVL3-8B	82.1	68.7	62.2	49.0	884	82.8	97.9	82.1	71.4	30.4
+RLHF	82.2	68.8	62.3	48.9	885	82.9	98.0	82.2	71.5	30.5
+HADPO	83.1	68.8	62.1	49.2	885	82.9	97.8	82.0	**72.0**	30.3
+DataTailor	82.0	68.9	62.4	49.1	883	82.7	97.8	82.0	71.3	30.3
+POVID	82.4	68.9	62.5	50.3	886	83.0	98.5	82.3	71.3	30.6
+SIMA	82.2	69.2	63.0	49.5	888	83.2	98.3	82.8	71.8	30.8
+VISTA	83.3	69.4	63.4	50.4	887	83.1	98.7	83.2	71.3	30.6
+DeepAlign	**83.6**	**69.8**	**63.7**	**50.5**	**890**	**83.6**	**99.2**	**83.9**	71.6	**31.4**

1) **DeepAlign is a model-agnostic method that can alleviate the misalignment of modality-specific representations across different backbone MLLMs.** It significantly enhances the performance of original models on various benchmarks. Specifically, when leveraging Qwen2.5-VL-7B as the backbone, it shows notable improvements: +1.0 points on MMBench [33], +1.6 points on MMStar [3], +1.5 points on ScienceQA [36], and +1.7 points on TextVQA [43], with consistent gains across other MLLM and VQA benchmarks.

2) Compared with baseline methods that also aim to achieve better modality alignment through post-training, DeepAlign mostly outperforms them on most benchmarks across the three backbone models (LLaVA-v1.5-7B, Qwen2.5-VL-7B, and InternVL3-8B). This indicates that **DeepAlign more effectively eliminates modality conflicts, thereby achieving superior performance**.

4.2. Mitigation of Modality Conflict

Fine-grained perceptive VQA tasks. We evaluate DeepAlign on several perceptive-related tasks of BLINK [14], where MLLMs must perceive the contents of the image to answer. As shown in Table 2, DeepAlign significantly improves the performance of Qwen2.5-VL-7B on these fine-grained perceptive VQA tasks, with steady gains across all six subtasks. It also helps narrow the performance gap between Qwen2.5-VL-7B and advanced closed-source models, such as Gemini-1.5-Pro and GPT-4o. This demonstrates DeepAlign enhances the model's perception capabilities, enabling it to better capture fine-grained visual semantics—rooted in its structural design that injects additional visual information distilled from DINOv2 into the MLLM, which effectively preserves valuable modality-specific details for perceptive tasks.

Table 2. Evaluations on BLINK [14] benchmark.

Methods	Spatial	Local	Vis.Corr.	Similarity	Counting	Depth
Gemini-1.5-Pro	79.7	63.1	80.2	83.7	57.5	77.4
GPT-4o	82.5	59.8	89.5	80.7	65.8	75.8
Qwen2.5-VL-7B	68.1	54.1	56.4	65.2	50.0	65.0
+POVID	70.5	55.0	58.3	66.8	55.2	66.4
+SIMA	69.2	56.2	60.1	68.5	52.5	67.8
+DeepAlign	**72.3**	**57.5**	**62.5**	**70.2**	**58.3**	**69.2**
InternVL3-8B	70.9	57.4	41.9	76.7	61.7	70.6
+POVID	72.1	58.5	45.5	78.3	65.0	73.5
+SIMA	73.5	59.8	43.2	80.1	63.3	72.1
+DeepAlign	**75.2**	**61.2**	**48.3**	**82.5**	**68.3**	**75.2**

Visual-Assisted Linguistic Tasks. We re-conduct experiments in Secion 2 on the three visual-assisted linguistic tasks (*i.e.*, MNER [47], MMT [54], and MRE [19]). The results in Fig. 5 (a) show that after mitigating modality conflict through DeepAlign, incorporating visual inputs improves the model's performance across all three tasks compared to text-only inputs. This demonstrates that DeepAlign resolves the modality conflict issue, enabling the model to leverage visual information. In contrast to earlier results in Fig. 2 (b) where adding visual inputs degraded perfor-

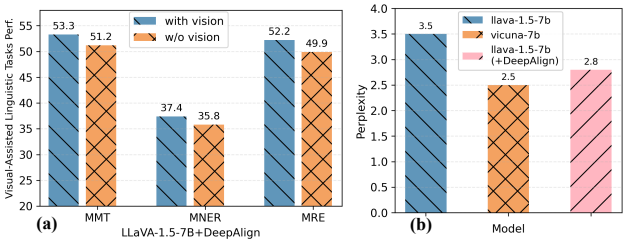

Figure 5. (a) The performance of LLaVA-v1.5-7B+DeepAlign on visual-assisted task with/without vision as the input. (b) Perplexity of text token prediction.

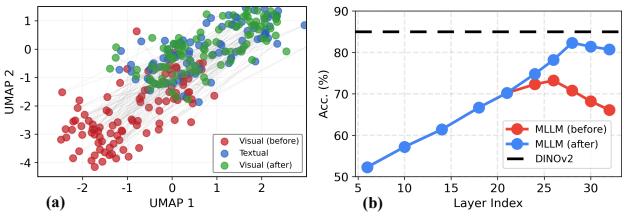

Figure 6. (a) Visualization of the embedding distribution for post-MLLM textual and visual hidden states (before/after applying DeepAlign). (b) Linear probing on ImageNet with MLLM internal hidden states (before/after applying DeepAlign).

mance, DeepAlign guides the MLLM to capture necessary modality-specific details, thereby synergistically improving the model's performance for multimodal understanding.

Perplexity of Text Prediction. We leverage 1000 randomly sampled text segments from Wikipedia to assess the perplexity of text-token prediction for LLaVA-v1.5-7B enhanced by DeepAlign. Fig. 5 (b) shows MLLM perplexity is still higher than its LLM, but notable reduction after post-training MLLM with DeepAlign. This suggests DeepAlign effectively mitigates modality misalignment, compensating for textual capability degradation from visual modality.

4.3. Emergent Abilities

Hallucination Mitigating. DeepAlign enables MLLMs to mitigate hallucinations. As shown in Table 1, after DeepAlign bridges the modality gap, the two backbones exhibit enhanced HallusionBench [27], which is a benchmark designed to evaluate hallucinations. Specifically, it brings $+1.7$ points on Qwen2.5-VL-7B (from 51.9 to 53.6) and $+1.5$ points on InternVL3-8B (from 49.0 to 50.5). Its improvements outperform baselines (POVID, SIMA), confirming DeepAlign effectively reduces hallucinations.

Multimodal In-context Learning. DeepAlign enables more seamless integration between image and text. With a small number of interleaved pairs, we post-train Qwen2.5-VL-7B with DeepAlign—unlocking its multimodal in-context learning (similar gains for LLaVA-v1.5-7B, InternVL3-8B). In Table 3, we evaluate the few-shot in-context learning performance of DeepAlign on OKVQA [37] and VQAv2 [15]. When provided with 4-shot, 8-shot, and 16-shot examples, Qwen2.5-VL-7B enhanced by DeepAlign shows steady performance gains as the number of in-context examples increases.

Table 3. Evaluations on few-shot in-context understanding.

Methods	VQAv2			OKVQA		
	4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
LLaVA-v1.5-7B	60.2	58.8	52.6	52.2	48.4	42.2
+DeepAlign	66.4	67.5	68.0	56.2	57.0	57.5
Qwen2.5-VL-7B	70.3	71.4	72.1	58.7	60.1	60.8
+DeepAlign	72.3	73.1	73.7	61.2	61.9	62.7
InternVL3-8B	72.2	72.9	73.6	60.3	61.4	62.1
+DeepAlign	73.6	74.3	75.1	62.6	63.3	64.2

Table 4. Zero-shot evaluation on DEMON [25] Benchmark.

Methods	MMD	VST	VRI	MMC	KGQA	TRQA	MMR
LLaVA-v1.5-7B	18.8	16.9	20.6	21.2	37.4	31.5	46.2
+DeepAlign	28.8	28.2	27.6	28.8	54.8	48.3	56.1
Qwen2.5-VL-7B	41.2	30.4	30.1	31.6	58.2	50.5	58.8
+DeepAlign	42.8	32.3	31.8	33.2	60.1	52.4	60.6
InternVL3-8B	42.5	32.0	31.6	33.2	59.3	52.1	60.2
+DeepAlign	44.3	33.8	33.2	34.7	61.3	53.8	62.0

Demonstrative Instruction Following on Interleaved Text-Image Sequence. We experiment on the DEMON benchmark [25], which evaluates ability to follow demonstrative instructions on interleaved image-text sequences. As shown in Table 4, DeepAlign significantly enhances Qwen2.5-VL-7B's performance across all DEMON task categories—for example, boosting MMD from 41.2 to 42.8 and KGQA from 58.2 to 60.1. Similar improvements are observed for LLaVA-v1.5-7B and InternVL3-8B, underscoring DeepAlign's ability of long-context understanding.

4.4. In-Depth Analysis

Visualization of Modality Representations. To visualize DeepAlign alleviating modality misalignment, for image-text pairs, we extracted post-MLLM visual hidden states before/after inserting our modality shifting module. Together with post-MLLM text hidden states, we visualize the embedding distribution of these three representations via UMAP. As shown in Fig. 6 (a), distribution gap between text/visual modalities is reduced after DeepAlign, showing our method mitigates misalignment.

Smaller Semantic Gap with DINOv2. With structure-induced distillation, we reuse linear probing on MLLM internal visual hidden states for ImageNet classification (Fig. 6 (b)). We observe two notable changes after DeepAlign: (1) Previously, accuracy dropped rapidly with increasing layer depth post-peak, but with distillation, decline is minimal—MLLM retains visual semantic details in higher layers. (2) Peak performance matches DINOv2, showing our visual supervision infuses necessary visual semantics for fine-grained image classification. This validates DeepAlign enhances MLLM's ability to preserve visual info across layers, bridging multimodal/vision-only gaps.

Effectiveness of Individual Components. To further investigate the effectiveness of individual components, we train the following ablation models: (1) *w/o intervention*:

Table 5. Ablation study of individual components.

Methods	MMBench	MMStar	ScienceQA	TextVQA
LLaVA-v1.5-7B	62.1	34.6	69.2	49.7
+DeepAlign	**68.2**	**40.1**	**75.3**	**55.7**
w/o intervention	63.5	36.2	70.8	51.3
w/o distillation	65.4	37.6	72.5	52.8
w/o global	66.8	38.8	73.9	54.1
w/o mutual	67.5	39.5	74.6	54.9
Qwen2.5-VL-7B	82.2	64.1	89.0	76.7
+DeepAlign	**83.2**	**65.7**	**90.5**	**78.4**
w/o intervention	82.3	64.3	89.3	77.0
w/o distillation	82.5	64.7	89.7	77.4
w/o global	82.8	65.1	90.0	77.9
w/o mutual	83.0	65.5	90.2	78.2
InternVL3-8B	82.1	68.7	97.9	82.1
+DeepAlign	**83.6**	**69.8**	**99.2**	**83.9**
w/o intervention	82.3	68.9	98.1	82.4
w/o distillation	82.7	69.1	98.3	82.6
w/o global	83.0	69.5	98.7	83.2
w/o mutual	83.3	69.6	99.0	83.7

remove the representation intervention framework; (2) *w/o mutual*: remove the loss in Eq. 4 from representation intervention; (3) *w/o global*: retain only instance-level modality shift direction (exclude global shift) in representation intervention; (4) *w/o distillation*: remove structure-induced distillation. We conduct the ablation study across multiple backbone MLLMs, evaluating on MMBench, MMStar, ScienceQA, and TextVQA as shown in Table 5. All ablation models consistently underperform the full DeepAlign across all benchmarks, with more notable performance drops in the *w/o intervention* and *w/o distillation* setups. This highlights the critical role of our proposed components in mitigating misalignment and preserving modality-specific information, which drives better performance on diverse vision-language tasks.

Loss Weight Ablation. We conduct loss weight ablation on Qwen2.5-VL-7B, fixing $\mathcal{L}_{\text{AR}}$ at 1.0 while adjusting four other losses' weights ($\mathcal{L}_{\text{visual}}$, $\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{instance}}$, $\mathcal{L}_{\text{MI}}$), with typical combinations in Table 6. Results show that all 1.0 (our setup) achieve excellent performance: it ranks first on ScienceQA and near the top on MMBench, outperforming combinations with insufficient (e.g., all 0.5) or excessive (e.g., all 1.5) weights. Notably, weight adjustments only cause minor fluctuations, confirming our all 1.0 design is robust, generalizable, and needs no complex tuning.

Modality Collaboration. With DeepAlign's seamless vision-text integration, we observe emerging modality synergy. Specifically, we use fixed pure text instruction-tuning data and gradually increase image-text pair proportion, post-training via DeepAlign and standard fine-tuning. As shown in Fig. 7, with standard fine-tuning, as image-text pair data increases, though performance on vision-language benchmarks (NoCaps [1]) improves, the performance on pure NLP benchmarks (MMLU [18]) significantly drops. In contrast, DeepAlign shows consistent upward performance on both benchmarks with more image-text pairs, indicating preliminary cross-modal enhancement synergy.

Table 6. Ablation study on loss weights (fix $\mathcal{L}_{\text{AR}} = 1.0$).

$\mathcal{L}_{\text{visual}}$	$\mathcal{L}_{\text{global}}$	$\mathcal{L}_{\text{instance}}$	$\mathcal{L}_{\text{MI}}$	MMBench	MMStar	ScienceQA
1.0	1.0	1.0	1.0	83.2	65.7	**90.5**
0.5	0.5	0.5	0.5	80.5	62.8	86.8
1.5	1.5	1.5	1.5	82.8	65.4	89.7
1.5	1.0	1.0	1.0	83.1	65.8	89.5
1.0	0.5	1.0	1.0	82.5	64.4	89.1
1.0	1.0	1.5	1.0	83.0	65.5	90.2
1.0	1.0	1.0	0.5	82.8	65.1	89.6
1.0	1.5	1.5	1.0	**83.3**	65.4	90.0
1.5	1.0	1.0	1.5	83.0	**65.9**	89.8
2.0	1.0	1.0	1.0	82.9	65.5	89.3

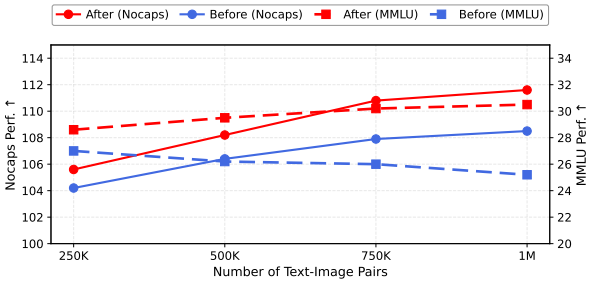

Figure 7. The trend of NoCaps & MMLU performance with the number of text-image training data.

5. Related works

MLLMs have become pivotal for bridging visual-textual modalities and advancing multimodal comprehension [30], yet visual-textual misalignment remains a core challenge. Their generated outputs often fail to fully match input visual data [29, 32], leading to hallucinations that undermine reliability in real-world use. To address this, existing methods adopt preference learning [48], adaptive learning [60], contrastive learning [53, 61], or optimal transport [40] to enhance semantic alignment. However, these approaches overlook modality-specific representations, focusing solely on eliminating inter-modal differences [45, 59], with restrictive objectives that directly compare visual and text features. Unlike them, we decompose representations into shared and specific components: mitigating misalignment in specific representations to resolve conflicts, and introducing visual supervision to preserve modality details [13], enabling harmonious integration in MLLMs.

6. Conclusion

In this paper, we address modality conflict in MLLMs, which stems from focusing on modality-shared information while neglecting modality-specific knowledge. We propose **DeepAlign**, a novel multimodal alignment framework mitigating modality conflict through representation intervention and structure-induced knowledge distillation, preventing misalignment and depletion of modality-specific information. Experimental results demonstrate DeepAlign significantly mitigates modality conflicts, enhancing MLLMs' performance on various vision-language tasks, and empowering them with emergent capabilities like multimodal in-context learning on interleaved image-text contexts.

Acknowledgements. This research is supported by National Research Foundation, Singapore, NRF-NRFI10-2024-0004.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 8

[2] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 5, 6

[4] Xiang Chen, Chenxi Wang, Ningyu Zhang, Yida Xue, YUE SHEN, GU Jinjie, Huajun Chen, et al. Unified hallucination detection for multimodal large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*. 1

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 5

[6] Xiangning Chen, Chen Liang, Da Zhou, Yue Feng, Cho-Jui Wang, and Quoc V Le. Moma: Efficient early-fusion pre-training with mixture of modality experts. *arXiv preprint arXiv:2407.21770*, 2024. 2

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 3

[8] Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Siliang Tang, Hanwang Zhang, and QIANRU SUN. Unified generative and discriminative training for multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:23155–23190, 2025. 1

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[10] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[12] Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024. 2

[13] 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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 8

[14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 3, 6

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017. 7

[16] 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, 2024. 5

[17] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3, 8

[19] Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. Multimodal relation extraction with cross-modal retrieval and synthesis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,

pages 303–311, Toronto, Canada, 2023. Association for Computational Linguistics. 2, 3, 6

[20] Hongzhe Huang, Zhewen Yu, Jiang Liu, Li Cai, Dian Jiao, Wenqiao Zhang, Siliang Tang, Juncheng Li, Hao Jiang, Haoyuan Li, et al. Align ^{2} llava: Cascaded human and large language model preference alignment for multi-modal instruction curation. *arXiv preprint arXiv:2409.18541*, 2024. 1

[21] Sanghyun Jo, Soohyun Ryu, Sungyub Kim, Eunho Yang, and Kyungsu Kim. Ttd: Text-tag self-distillation enhancing image-text alignment in clip to alleviate single tag bias, 2024. 5

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[25] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 7

[26] Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Yuan Li, Zhaopeng Tu, and Xiaolong Li. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*, 2025. 5

[27] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 7

[28] Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism, 2023. 5

[29] Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 8

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3, 8

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 5

[32] Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Andrew Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, Xiaohan Zhang, Lichao Sun, Xiaotao Gu, Hongning Wang, Jing Zhang, Minlie Huang, Yuxiao Dong, and Jie Tang. AlignBench: Benchmarking Chinese alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11621–11640, Bangkok, Thailand, 2024. Association for Computational Linguistics. 8

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5, 6

[34] Yang Liu, Chen Li, and Bolei Zhou. Detr++: Hierarchical decoding for fine-grained visual grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2

[35] 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*, 67(12), 2024. 5

[36] 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 *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[37] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*. 2, 5

[39] Yuqi Pang, Bowen Yang, Haoqin Tu, Yun Cao, and Zeyu Zhang. Language models can see better: Visual contrastive decoding for llm multimodal reasoning. *arXiv preprint arXiv:2502.11751*, 2025. 1

[40] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridging vision and language spaces with assignment prediction, 2024. 8

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[42] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 4

[43] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 5, 6

[44] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024. 5

[45] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 8

[46] Xinyu Wang and Yixuan Jiang. A comprehensive review of multimodal large language models. *arXiv preprint arXiv:2408.01319*, 2024. 2

[47] Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, and Kewei Tu. ITA: Image-text alignments for multi-modal named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3176–3189, Seattle, United States, 2022. Association for Computational Linguistics. 2, 3, 6

[48] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and Cao Xiao. Enhancing visual-language modality alignment in large vision language models via self-improvement, 2024. 5, 8

[49] Zitian Wang, Zehao Huang, Yulu Gao, Naiyan Wang, and Si Liu. Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection. *arXiv preprint arXiv:2408.05945*, 2024. 1

[50] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024. 1

[51] Huiming Xie, Yang Qin, and Shuxue Ding. Hierarchical vision–language pre-training with freezing strategy for multi-level semantic alignment. *Electronics*, 14(4):816, 2025. 1

[52] Jian Yang, Dacheng Yin, Yizhou Zhou, Fengyun Rao, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling. *arXiv preprint arXiv:2410.10798*, 2024. 2

[53] Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, and Wentao Zhang. SEA: supervised embedding alignment for token-level visual-textual integration in mllms. *CoRR*, abs/2408.11813, 2024. 8

[54] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3, 6

[55] Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. Mastering collaborative multi-modal data selection: A focus on informativeness, uniqueness, and representativeness, 2024. 1, 5

[56] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5

[57] 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 *Proceedings of CVPR*, 2024. 5

[58] Shuoxi Zhang, Hanpeng Liu, Stephen Lin, and Kun He. You only need less attention at each stage in vi-

sion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2024. 1

[59] Shichuan Zhang, Sunyi Zheng, Zhongyi Shui, Honglin Li, and Lin Yang. Multi-modal learning with missing modality in predicting axillary lymph node metastasis, 2024. 8

[60] Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. AlignGPT: Multi-modal large language models with adaptive alignment capability, 2024. 8

[61] Yaqi Zhao, Yuanyang Yin, Lin Li, Mingan Lin, Victor Shea-Jay Huang, Siwei Chen, Weipeng Chen, Baoqun Yin, Zenan Zhou, and Wentao Zhang. Beyond sight: Towards cognitive alignment in lvlm via enriched visual knowledge, 2024. 8

[62] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2023. 5

[63] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 5

[64] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 1

[65] Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024. 1