COMPODISTILL: ATTENTION DISTILLATION FOR COMPOSITIONAL REASONING IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, efficient Multimodal Large Language Models (MLLMs) have gained significant attention as a solution to their high computational complexity, making them more practical for real-world applications. In this regard, the knowledge distillation (KD) approach has emerged as a promising alternative, which transfers the rich visual and linguistic knowledge from a larger model (teacher) to a smaller model (student). However, we observe that existing KD methods struggle to effectively distill the teacher MLLM's rich *visual perception* abilities to the student, a challenge that has been largely overlooked in previous studies. Through a systematic analysis, we identify visual attention misalignment between student and teacher as the main cause of this issue. Based on this insight, we propose CompoDistill, a novel KD framework that explicitly aligns the student's visual attention with that of the teacher to enhance the student's visual perception abilities. Our extensive experiments show that CompoDistill significantly improves performance on compositional reasoning tasks that require visual perception abilities while maintaining strong performance on visual question answering tasks, as done in existing studies. Furthermore, CompoDistill demonstrates effectiveness with a more advanced backbone, highlighting its generalizability.

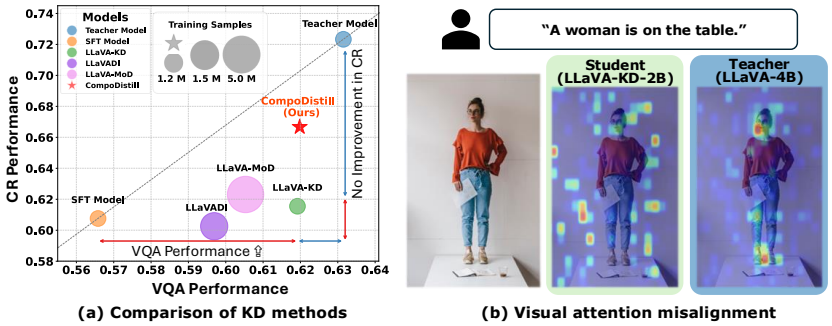

Figure 1: (a) Comparison of the teacher (LLaVA-4B), SFT (LLaVA-2B), and various 2B KD methods distilled from the same teacher. (b) *Visual attention misalignment* between the student (LLaVA-KD-2B) and teacher (LLaVA-4B), where the student attends to irrelevant image regions for the text query *A woman is on the table*, in contrast to the teacher. For more examples including our proposed method, CompoDistill, please refer to Appendix A.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Liu et al., 2023; Lin et al., 2024b) have demonstrated superior performance on various vision-language tasks over vision-language models (Radford et al., 2021; Zhai et al., 2023), marking a significant step forward in multimodal learning. Yet, their advancement has largely been driven by scaling laws, which has resulted in ever-larger and more computationally demanding architectures (OpenAI, 2023; Qwen et al., 2025). The substantial computational and memory costs associated with such massive models pose considerable challenges for practical deployment. This has, in turn, intensified interest in developing **efficient MLLMs** that can maintain strong multimodal capabilities while mitigating resource requirements.

To this end, **knowledge distillation (KD)** (Cai et al., 2024; Shu et al., 2024; Xu et al., 2024) has emerged as a promising approach, transferring rich knowledge from a larger *teacher* model to a smaller *student* model. These KD-based methods have demonstrated effectiveness in visual question

answering (VQA) tasks (e.g., GQA (Hudson & Manning, 2019)), which require *visual recognition* ability, outperforming the standard supervised fine-tuning (SFT model[1] in Figure 1(a)).

Despite the progress of KD methods in visual recognition ability, a crucial question remains unexplored: *Is visual perception ability equally well distilled?* *Visual recognition* refers to tasks like object classification, where the goal is to identify objects in images based on their features. On the other hand, *Visual perception* involves more complex abilities, such as understanding relationships among objects and accurately capturing their attributes—capabilities essential for real-world multimodal applications[2]. To answer the question, in Figure 1(a), we compare the state-of-the-art KD methods (Shu et al., 2024; Cai et al., 2024; Xu et al., 2024) on VQA and compositional reasoning (CR) datasets, where VQA serves as a benchmark for evaluating visual recognition ability and CR is designed to assess visual perception ability[3]. We observe that, although existing KD methods show significant performance improvements on the VQA, their performance on the CR is on par with the SFT model, which is unexpected. Based on this observation, *we argue that existing KD methods struggle to effectively distill visual perception ability from the teacher model.*

Beyond a simple performance comparison, we investigate the failure of the existing KD methods in distilling visual perception ability and analyze the attention maps of both the student (i.e., LLaVA-KD-2B (Cai et al., 2024)) and the teacher (i.e., LLaVA-4B (Liu et al., 2023)) models. Specifically, by visualizing their focus areas for a given text, we aim to identify differences in attention, which we believe are closely related to perception ability. As shown in Figure 1(b), we observe that the student model struggles to focus on the relevant regions, while the teacher model captures them effectively, revealing a clear mismatch in the attention distributions between the teacher and the student models. The discrepancy in attention distributions, which we term ***visual attention misalignment***, reveals that although KD methods aim to transfer knowledge from the teacher to the student through knowledge distillation, the student fails to inherit the teacher's powerful visual attention mechanism, thereby limiting the effective distillation of visual perception ability.

In this regard, we hypothesize that the limited performance improvement of KD methods in CR tasks arises from the visual attention misalignment between the teacher and student models. To empirically validate our hypothesis, we conduct controlled experiments examining the relationship between attention behavior and visual task performance in Section 3. Our results demonstrate that this misalignment is directly responsible for the unexpectedly low performance of the existing KD methods in CR tasks, and this study is the first to establish that addressing visual attention misalignment is the key to enhancing the student model's visual perception ability.

Motivated by these findings, we propose CompoDistill, a novel KD framework aimed at effectively distilling the teacher's rich visual perception abilities to the student model by addressing visual attention misalignment. We introduce the **Visual ATtention alignment (VAT)** module to explicit align the visual attention of the student model with that of the teacher model, using a simple yet effective group matching strategy, in which each student layer is matched with a group of teacher layers in a one-to-many manner, to handle the difference in the number of LLM layers between the teacher and student. However, the teacher's visual attention is optimized for its own vision-language space, making a simple transfer via the VAT module ineffective within the student's distinct and incompatible space. This mismatch creates a conflict between the student's feature space and the imposed attention mechanism, ultimately restricting the model's inherent perceptual abilities. To this end, we propose the **Teacher Adapter Fetch (TAF)** module to bridge this feature space gap and enable synergy with the VAT module. Building on this, we introduce a meticulously designed three-stage training strategy that leverages both modules to comprehensively distill visual perception.

Through extensive experiments on multiple CR datasets, we demonstrate that CompoDistill significantly outperforms existing KD methods, demonstrating the effectiveness of CompoDistill in enhancing the student model's visual perception ability. Meanwhile, CompoDistill maintains competitive performance on VQA datasets, preserving its visual recognition abilities. Furthermore, we demonstrate the effectiveness of CompoDistill with a more advanced backbone, highlighting its generalizability.

[1] Sharing the same architecture and size as the student models in KD, this model is trained only with visual instruction tuning Liu et al. (2023) via supervised fine-tuning. We henceforth denote it as the SFT model.

[2] Regarding a detailed comparison of the differences between the two abilities, please refer to Appendix B.

[3] VQA datasets include GQA (Hudson & Manning, 2019), TextVQA (Singh et al., 2019), and MME (Fu et al., 2024), and CR datasets include SugarCrepe (Hsieh et al., 2023), SADE (Ma et al., 2023), BiVLC (Miranda et al., 2024), and Winoground (Thrush et al., 2022). Results in Figure 1(a) correspond to the average performance across the datasets.

We summarize our contributions as follows: (1) We identify, for the first time, that existing KD methods in MLLMs fail to distill visual perception ability from the teacher and provide a detailed attention-based analysis, offering insights on how to enhance this ability. (2) We propose CompoDistill, a novel KD framework that transfers both visual recognition and perception abilities through the Visual ATtention alignment and the Teacher Adapter Fetch Module. (3) We achieve significant improvements in CR tasks by effectively distilling the visual perception ability, while maintaining competitive performance in VQA tasks by preserving the visual recognition ability.

2 PRELIMINARY

Multimodal Large Language Models (MLLMs). Following the LLaVA (Liu et al., 2023) design, MLLMs consist of three main components: a pre-trained LLM $LM_\theta(\cdot)$, a vision encoder $V_\phi(\cdot)$, and an adapter $P_\psi(\cdot)$. Given a text query Q and an image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the image height and width, the vision encoder extracts patch-level features $\mathbf{z}_p = V_\phi(I) \in \mathbb{R}^{N_v\times d_p}$, with N_v being the number of visual patches and d_p the hidden dimension of V_ϕ. The adapter projects them into the language space as $\mathbf{x}_v = P_\psi(\mathbf{z}_p) \in \mathbb{R}^{N_v\times d}$, where d is the hidden dimension of LM_θ. Hereafter, we refer to $\mathbf{x}_v$ as the *visual tokens*. Meanwhile, the query Q is tokenized into embeddings $\mathbf{x}_t \in \mathbb{R}^{N_t\times d}$, where N_t is the number of text tokens. The combined sequence $[\mathbf{x}_v, \mathbf{x}_t] \in \mathbb{R}^{(N_v+N_t)\times d}$ is processed by the LM_θ to compute the probability of the target answer as:

$$p(\mathbf{y}_{1:K}) = \prod_{i=1}^{K} p(\mathbf{y}_i \mid \mathbf{x}_v, \mathbf{x}_t, \mathbf{y}_{<i}), \tag{1}$$

where $\mathbf{y}_{<i}$ is the sequence of the answer tokens up to the i-th token and K is the answer length. During training, the cross-entropy loss derived from Equation 1 is minimized, denoted as $\mathcal{L}_{LM}$.

3 WHY IS THE VISUAL PERCEPTION ABILITY NOT DISTILLED PROPERLY?

In Section 1, we noted that existing KD methods, despite improvements in terms of VQA, show unexpected results on compositional reasoning (CR) tasks, which we attribute to ***visual attention misalignment***. This section provides an in-depth analysis to elucidate the underlying causes of this issue. Our analysis first identifies the key factor for distilling visual ability through an attention-based analysis (Section 3.1), then demonstrates this factor's direct impact on performance (Section 3.2). Finally, we validate that alleviating the visual attention misalignment associated with this key factor facilitates a more effective transfer of visual perception (Section 3.3).

3.1 IDENTIFYING THE KEY FACTOR: TEACHER-STUDENT ATTENTION SIMILARITY OVER VISUAL TOKENS IN THE VISUAL UNDERSTANDING LAYER

In this section, we aim to identify a critical factor that can serve as a criterion for successfully distilling visual abilities (i.e., visual recognition and perception). To achieve this, we examine the layer-wise functions of the LLM within MLLMs, building on insights from prior works (Chen et al., 2025a; Yoon et al., 2025). This is followed by an attention-based analysis to pinpoint the key factor.

Layer-wise Functionality in MLLMs. Following previous studies (Chen et al., 2025a; Yoon et al., 2025), we adopt a layer-wise view of MLLM processing: *Early layers* align heterogeneous modalities with the LLM's text space, *Intermediate layers* integrate these signals for fine-grained semantic understanding, and *Later layers* generate the final response. Our focus is on the intermediate layers[4], which we term the **visual understanding layers**, since these layers play a critical role in forming foundational visual reasoning and comprehension (Chen et al., 2025b), which are closely tied to the visual abilities (recognition and perception) central to our study.

Attention Analysis for VQA and CR Tasks. Building on our focus on the visual understanding layers, we design an experiment to investigate if teacher-student attention alignment can explain the difference in the degree of performance improvement between VQA and CR tasks, as observed in Figure 1(a). To this end, we compare the attention distribution of the teacher model (i.e., LLaVA-4B (Liu et al., 2023) with that of two distinct models: a student model distilled directly from the teacher (i.e., LLaVA-KD-2B (Cai et al., 2024)) and an SFT model (i.e., LLaVA-2B) that is architecturally identical to the student model but trained without distillation.

As shown in Figure 2(a), we measure the attention similarity to quantify the alignment between a given model (student or SFT) and the teacher. This is computed as the cosine similarity of

[4]Following (Neo et al., 2025), intermediate layers refer to the 30–70% range of the total LLM layers.

their attention distributions, where the answer token $\mathbf{y}_i$ is the *query (Q)* and the visual tokens ($\mathbf{x_v}$) or text tokens ($\mathbf{x_t}$) are the *key (K)*. Our primary analysis centers on the attention distributions over visual tokens during answer generation, as these are critical for the visual abilities being studied. For a more comprehensive comparison, we also investigate attention distribution over text tokens. We analyze the attention similarity of the student and SFT models to the teacher. Specifically, we investigate how these similarities in attention behavior are related with each model's performance in VQA and CR task, which brings us closer to identifying the key factor. To guide this investigation, we formulate two key research questions:

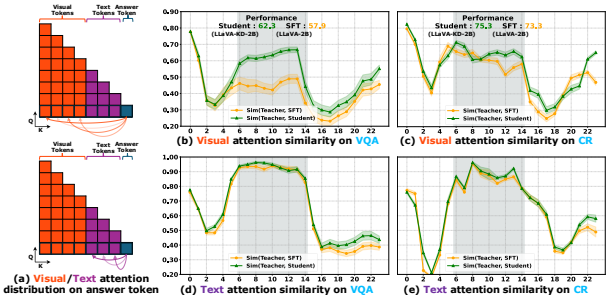

Figure 2: (a) Attention of the answer token over visual tokens and text tokens during its generation. The transparency indicates the degree of attention the answer token gives to these tokens. (b-e) Layer-wise Teacher-Student and Teacher-SFT attention similarities over visual tokens ((b), (c)) and text tokens ((d), (e)). GQA is used for VQA ((b), (d)), and SugarCrepe is used for CR ((c), (e)). Results on other datasets are shown in Appendix C.

Q1) Where do the performance gains of KD on VQA (visual recognition) tasks originate? For VQA, where the student model demonstrates clear performance improvements, Figure 2(b) shows that, within the *visual understanding layers*, the teacher-student attention similarity over visual tokens is significantly higher than that between the teacher-SFT. However, no such improvement is observed for text tokens (Figure 2(d)). This finding brings us to identify the clue of the key factor: *if the teacher-student visual attention similarity in the visual understanding layers is high, distillation is effectively achieved, resulting in performance improvement of the student* .

Q2) Why does KD fail to deliver comparable improvements on CR (visual perception) tasks? We extend our analysis to CR tasks, where the student model performs on par with the SFT model. As shown in Figure 2(c) and (e), we observe that the teacher-student attention similarity is comparable to the teacher-SFT attention similarity in the *visual understanding layers*, even over the visual tokens, which contrasts with our observation on VQA. That is, in the *visual understanding layers*, the student model shows no better alignment with the teacher model in terms of the attention over visual tokens than the SFT model, despite such alignment being crucial for success in the downstream performance as shown in the case of VQA. Drawing on the findings from ***Q1*** and ***Q2***, we argue that the inability of existing KD methods to effectively distill visual perception ability stems from the moderate level of teacher–student attention similarity in the *visual understanding layers*.

3.2 Relationship between Teacher-Student Attention Similarity and Downstream Performance

While the teacher-student attention similarity over visual tokens appears to be a key factor in determining whether visual abilities have been distilled, its direct relationship with downstream performance remains unclear. In other words, it is uncertain whether the higher teacher–student attention similarity (green line), relative to the teacher–SFT attention similarity (yellow line) in the visual understanding layers (Figure 2(b)), explains the superior performance of the student model on the VQA dataset. To this end, we quantify the relationship between the teacher-student *visual attention similarity* and the *answer token probability* given in Equation 1 on the VQA dataset (Figure 3), where visual recognition ability has been successfully distilled from the teacher model. More precisely, we randomly sampled 5,000 instances from the GQA test set, grouped them according to the student–teacher similarity over visual tokens during the answer generation step, and then measured the average probability assigned to the ground-truth answers. The results reveal a clear positive trend: higher attention similarity is consistently associated with higher answer probabilities, providing direct evidence that aligning the student's attention with the teacher's attention is a key factor driving better performance on VQA.

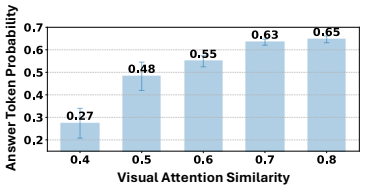

Figure 3: Relationship between attention similarity and performance on the GQA (VQA) dataset.

3.3 A Simple Solution: Replacing Student's Visual Attention with Teacher's

Is Teacher Attention Really Beneficial? A natural question arises: *Does increasing the teacher-student attention similarity over visual tokens —a key factor for VQA performance—also enhance performance on CR tasks?* To test this, we perform a direct intervention during inference on CR. Before generating the answer, we substitute the student's original attention over visual tokens with the average of the teacher's and student's attention, thereby increasing the student's attention similarity to the teacher[5].(Figure 4).This yields modest but consistent performance gains in the SugarCrepe dataset (Swap, Replace and Add are three types of CR sub-tasks in the dataset). Although small, these improvements confirm that the student benefits from incorporating the teacher's attention, reinforcing our hypothesis that attention alignment plays a crucial role in effective knowledge transfer.

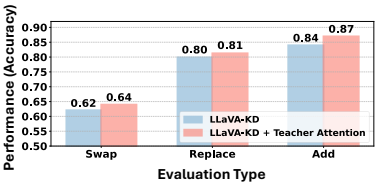

Figure 4: Change in performance the student attention is substituted with the teacher attention on the SugarCrepe (CR) dataset.

In summary, our analysis identifies *visual attention misalignment* as the key barrier to effectively distilling the teacher's perception abilities to the student, highlighting the crucial need to accurately transfer the teacher's visual attention.

4 METHODOLOGY: COMPODISTILL

Motivated by the findings in Section 3, our framework, called CompoDistill, is designed to effectively transfer the teacher's visual perception abilities by addressing the visual attention misalignment. We first introduce our two core components: the **Visual ATtention alignment (VAT)** module (Section 4.1), which aligns the student's visual attention mechanism with that of the teacher, and the **Teacher Adapter Fetch (TAF)** module (Section 4.2), which ensures the student processes visual space consistently with the teacher. These modules are then integrated into a meticulously designed three-stage training strategy (Section 4.3). The overall framework is shown in Figure 5.

4.1 VISUAL ATTENTION ALIGNMENT (VAT) MODULE

Attention Distillation Loss. To transfer the teacher's attention mechanism to the student, we utilize the attention matrix of the Transformer (Vaswani et al., 2017) within the LLM layer, which represents the importance of each token relative to other tokens. The attention matrix for the input tokens (i.e., $\mathbf{x}_v$ and $\mathbf{x}_t$) is computed as: $A = \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\right) \in \mathbb{R}^{(N_v+N_t)\times(N_v+N_t)}$, where $\mathbf{Q} \in \mathbb{R}^{(N_v+N_t)\times d}$ and $\mathbf{K} \in \mathbb{R}^{(N_v+N_t)\times d}$ are the query and key matrices derived from the input tokens, respectively. From the overall attention matrix A, as discussed in Section 3, our goal is to distill the *attention over visual tokens* from the teacher to the student in the visual understanding layers, leading us to extract a sub-matrix $\tilde{A} \in A$ that focuses on visual attention. Specifically, for each visual understanding layer l, this sub-matrix includes only the columns (keys) of A_l corresponding to the visual tokens, resulting in $\tilde{A}_l = A_l[:, : N_v] \in \mathbb{R}^{(N_v+N_t)\times N_v}$. Based on the visual attention-related sub-matrices of the teacher ($\tilde{A}^t_l$) and student ($\tilde{A}^s_l$), we compute the cosine distance between them for the attention distillation loss as $1 - \text{sim}\left(\tilde{A}^t_{l_t}, \tilde{A}^s_{l_s}\right)$, where l_t and l_s denote the index of the visual understanding layers of the teacher and student, respectively.

Group Layer Matching. However, since the teacher has more layers than the student, directly matching a student layer index l_s with a teacher layer index l_t is challenging. A naive solution is to map each l_s to one l_t by uniformly sampling teacher layers according to the ratio of their depths. For example, if the student has 5 layers and the teacher has 10 layers, then teacher layers $\{1, 3, 5, 7, 9\}$ are selected and aligned with the 5 student layers. However, this approach is suboptimal, as it overlooks the richer perception abilities distributed across the teacher's layers and does not ensure accurate alignment[6]. Instead, we propose a simple yet effective **Group Layer Matching** strategy, where each l_s is aligned with a group of $\{l_t\}$ in a one-to-many manner, enabling the student to capture broader teacher knowledge while roughly preserving the layer order.

[5]We initially attempted to completely replace the student's attention with that of the teacher. However, this rather led to a slight performance degradation compared with vanilla LLaVA-KD, which we attribute to a mismatch between the student's feature space and the teacher's attention.

[6]In Table 3, we show that this approach is less effective than the proposed Group Layer Matching strategy.

Formally, let the sequence of student's visual understanding layer indices be $L_S = \{l_s^1, \ldots, l_s^j, \ldots, l_s^k\}$ and that of the teacher be $L_T = \{l_t^1, \ldots, l_t^j, \ldots, l_t^m\}$, where j is the j-th layer in the visual understanding layers, and k and m denote the total number of visual understanding layers for the student and teacher, respectively, with $k < m$. For each student layer l_s^j, we define a corresponding group of teacher layers, G_j, consisting of n[7] consecutive teacher layers, formed using a sliding window approach. For example, the group of teachers G_1 assigned for the first student layer l_s^1 could be $\{l_t^1, l_t^2, \ldots, l_t^n\}$, while the group G_2 for the second student layer l_s^2 would be $\{i_t^2, i_t^3, \ldots, i_t^{n+1}\}$, and so on. So, the total number of groups is the same as the number of student's target layers, k.

To distill the visual attention of a teacher group G_j into student layer l_s^j, we formulate the objective by minimizing the cosine distance between the attention matrix of student layer, $\tilde{A}_{l_s^j}^s$, and the averaged attention matrices of its corresponding teacher group, $\bar{A}_j^t$, as follows:

$$\mathcal{L}_{ADL} = 1 - \frac{1}{k}\sum_{j=1}^{k} \text{sim}\left(\bar{A}_j^t, \tilde{A}_{l_s^j}^s\right), \bar{A}_j^t = \frac{1}{n}\sum_{l \in G_j} \tilde{A}_l^t. \quad (2)$$

Furthermore, beyond the naive one-to-one matching approach, our proposed group matching strategy is superior to the more advanced adaptive matching method (Lee et al., 2025b) in terms of effectiveness, as demonstrated in Section 5.3.

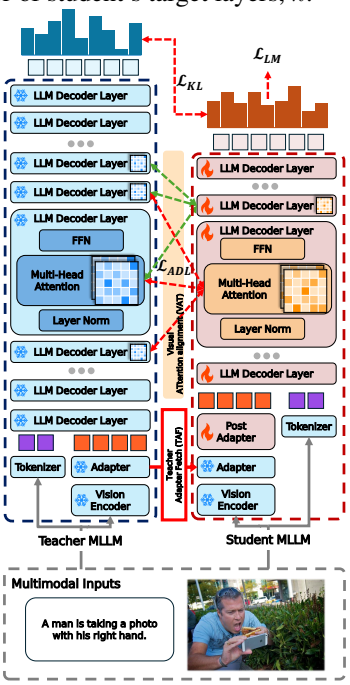

Figure 5: Overview of CompoDistill. It consists of the VAT module and the TAF module.

4.2 Teacher Adapter Fetch (TAF) Module

The adapter, responsible for projecting the visual space into the language space of the LLM, is crucial for generating the visual tokens with which the attention mechanism processes. In this regard, given that the teacher's attention mechanism transferred via VAT is tightly coupled with the output of its own adapter, a vision-language space mismatch occurs when this attention is imposed on a student with a different, incompatible space, hindering effective knowledge transfer. To address this, we introduce the **Teacher Adapter Fetch (TAF)** module, which directly leverages the teacher's frozen, pretrained adapter ($P_{\psi_t}^t$) and adds a lightweight trainable MLP ($P_{\psi_s}^s$) only for dimensional alignment. This ensures the student processes the visual input through the same lens as the teacher, making the attention transfer effective. Formally, the student's visual token is expressed as follows:

$$\mathbf{x}_v = P_{\psi_s}^s\left(P_{\psi_t}^t(\mathbf{z}_p)\right) \in \mathbb{R}^{N_v \times d^s}, \quad (3)$$

where d_s is the hidden dimension of the student LLM. This approach allows for the student to bridge the gap between the teacher-imposed attention mechanism and its own visual feature space.

Discussion. It is important to note that attention distillation itself is not a novel concept, as similar approaches have been explored in various domains (Wang et al., 2020; Sajedi et al., 2023; Zhou et al., 2025), including diffusion models, dataset distillation, and language models. However, its application to MLLMs is particularly challenging due to mismatched visual feature spaces between teacher and student. To this end, we introduce the TAF module as a simple yet effective solution that makes attention distillation practical in this setting. Lastly, our work is not intended to propose a fundamentally new distillation method; rather, its primary contribution lies in identifying a critical factor that can serve as a criterion for successfully distilling visual abilities in MLLMs, i.e., visual attention misalignment, as explained in Section 3.

4.3 Three-Stage Distillation Framework

Our knowledge distillation framework consists of three stages, built upon two foundational training objectives. The first objective is language modeling autoregressive loss $\mathcal{L}_{LM}$, as defined in Section 2. Moreover, following previous studies (Shu et al., 2024; Cai et al., 2024), we employ

[7] To ensure the use of all teacher layers L_T, we define n in closed form as $m - k + 1$.

a Kullback-Leibler (KL) divergence loss ($\mathcal{L}_{KL}$) to align the student's predictive distribution (p_s) with the teacher's (p_t), encouraging the student to mimic the teacher's final output at the logit level:

$$\mathcal{L}_{KL} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N} p_t(\mathbf{y}_n|\mathbf{x}_v, \mathbf{x}_t, \mathbf{y}_{<k}) \log \frac{p_t(\mathbf{y}_n|\mathbf{x}_v, \mathbf{x}_t, \mathbf{y}_{<k})}{p_s(\mathbf{y}_n|\mathbf{x}_v, \mathbf{x}_t, \mathbf{y}_{<k})}, \tag{4}$$

where $\mathbf{y}_n$ is the n-th vocabulary token and N is the total vocabulary size.

Stage 1: Distilled Pre-Training (DPT). The first stage aims to align the visual feature space with the language space. To this end, instead of initializing an adapter from scratch, we construct it using our **Teacher Adapter Fetch** module, which reuses the teacher's pretrained adapter, which is kept frozen during training. While the vision encoder and LLM remain frozen, the student adapter $P^s_{\psi_s}$ is optimized with the language modeling loss ($\mathcal{L}_{LM}$) and the KL-divergence loss ($\mathcal{L}_{KL}$).

Stage 2: Distilled Fine-Tuning (DFT). In the second stage, we aim to enhance the student's visual perception by aligning its visual attention mechanism over visual tokens with that of the teacher via the **Visual ATtention Alignment** module. During this stage, both the student LLM and the student adapter $P^s_{\psi_s}$ are fine-tuned. The overall objective is defined as $\mathcal{L}_{LM} + \mathcal{L}_{KL} + \mathcal{L}_{ADL}$.

Stage 3: Supervised Fine-Tuning (SFT). Finally, motivated by prior work (Lee et al., 2025b), we fine-tune only the student using the autoregressive language modeling loss ($\mathcal{L}_{LM}$), with both the student LLM and the student adapter $P^s_{\psi_s}$ trained in the same manner as in Stage 2. This final stage consolidates the rich knowledge transferred from the teacher during Stages 1 and 2 into the student's own parameters and further strengths its instruction-following capability.

5 Experiments

5.1 Experimental Settings.

Evaluation Benchmarks. For evaluation, we use two categories of vision-language tasks. ***General VQA***: We evaluate visual recognition abilities using well-established benchmarks including VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), and VizWiz (Bigham et al., 2010) for question answering, TextVQA (Singh et al., 2019) for scene text comprehension, and MME (Fu et al., 2024) for comprehensive multimodal evaluation. ***Compositional Reasoning***: To evaluate visual perception abilities, we use several challenging benchmarks: SugarCrepe (Hsieh et al., 2023), SADE (Ma et al., 2023), BiVLC (Miranda et al., 2024), and Winoground (Thrush et al., 2022). We use accuracy as the evaluation metric. Refer to Appendix D for more details regarding the baselines.

Implementation Details. Both the student and teacher employ SigLIP (Zhai et al., 2023) vision encoder and the Qwen 1.5 (Yang et al., 2024) LLM series, with the student having 1.8B parameters and the teacher having 4B parameters. Regarding the details of the training datasets and hyperparameter settings, please refer to Appendix E.

5.2 Quantitative Results on VQA and CR tasks

In Table 1, we compare CompoDistill with a diverse range of MLLMs of different sizes on VQA and CR tasks, which evaluate visual recognition and perception abilities, respectively. We have the following key observations: **1)** Existing KD methods[8] outperform the SFT model (LLaVA-2B) on VQA (visual recognition) tasks, but not on CR (visual perception) tasks, where their performance remains comparable to that of standard 2B models. This indicates that KD methods struggle to distill visual perceptual abilities. **2)** CompoDistill overcomes this limitation by significantly improving CR performance to a level competitive with 4B models—an achievement not attained by previous KD methods—while simultaneously maintaining strong, state-of-the-art performance on VQA. This demonstrates the ability of CompoDistill to enhance visual perception while maintaining visual recognition. **3)** CompoDistill achieves these results with high data efficiency, relying on just 1.2M training samples. This is a sharp contrast to other models requiring much larger datasets (e.g., LLaVA-MoD with 5M, MiniCPM-V with 570M), proving the effectiveness of CompoDistill toward a truly efficient MLLM.

5.3 Ablation Studies

Effect of Core Components. In Table 2, we conduct ablation studies to understand the effect of each component (i.e., VAT and TAF modules). Note that the variant without either component (row

[8]To ensure fair comparisons, all KD models were distilled from the same teacher model (i.e., LLaVA-4B) and share the same LLM backbone (i.e., Qwen 1.5).

Table 1: Comparison CompoDistill with KD-based methods and other MLLMs on VQA and CR. # Samples : Training data samples. †:Reproduced results using the official source code. The best and second-based models are marked in bold and underlined, respectively for models sizes under 2B.

Size	Method	LLM	# Samples	Visual Question Answering						Compositional Reasoning				
				VQAv2	VizWiz	GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
~7B	LLaVA-7B†	Qwen1.5-7B	1.2 M	79.8	53.1	62.3	60.1	68.6	64.8	87.3	78.5	65.7	58.4	72.5
	LLaVA-7B†	Qwen2.5-7B	1.2 M	81.4	50.6	63.8	64.6	77.5	67.6	93.1	82.9	68.3	68.2	78.1
	CogVLM-7B	Vicuna-7B	1500 M	82.3	-	64.9	78.2	71.8	-	81.8	63.2	64.5	57.2	66.7
	Qwen2.5-VL-7B	Qwen2-7B	1500 M	82.8	61.7	63.9	73.9	83.9	73.2	88.5	77.4	68.4	75.3	77.4
	Deepseek-VL-7B	DLLM-7B	2000 M	-	49.9	61.3	64.7	73.4	-	86.2	78.8	66.2	61.7	73.2
~4B	LLaVA-4B† (*Teacher*)	Qwen1.5-4B	1.2 M	79.1	48.0	62.1	56.7	67.2	62.6	83.0	75.5	64.8	57.8	70.3
	Qwen2.5-VL-3B	Qwen2-3B	1.2 M	80.4	54.1	60.9	68.5	72.8	67.3	57.1	65.9	67.0	63.5	63.4
	Imp-3B	Phi2-2.7B	1.5 M	81.2	54.1	63.5	59.8	72.3	66.2	78.1	61.1	59.5	38.3	59.3
	Bunny-3B	Phi2-2.7B	2.6 M	79.8	43.8	62.5	56.7	74.4	63.4	75.8	67.2	59.7	53.1	64.0
	MoE-LLaVA-3B	Phi2-2.7B	2.6 M	79.9	43.7	62.6	57.0	71.6	63.0	80.5	70.4	64.4	54.1	67.3
	MobileVLM-3B	MobileLLaMA-2.7B	1.3 M	-	-	61.1	57.5	72.0	-	67.9	72.1	57.2	43.9	60.3
	MiniCPM-V	MiniCPM-2.4B	570 M	-	60.2	52.1	73.2	70.5	-	90.0	70.6	65.3	55.3	70.3
~2B	MiniGemini-2B	Gemma-2B	2.7 M	-	41.5	60.7	56.2	67.0	-	67.0	62.0	58.0	50.5	59.4
	Deepseek-VL-1.3B	DLLM-1.3B	2000 M	-	36.8	59.3	**58.4**	65.3	-	36.6	38.6	39.9	60.8	44.0
	MobileVLM-1.7B	MobileLLaMA-1.4B	1.2 M	-	-	59.3	52.1	65.1	-	69.8	64.0	53.8	43.4	57.7
	Imp-2B	Qwen1.5-1.8B	1.5 M	**79.2**	39.2	61.9	54.5	65.2	60.0	71.0	59.6	58.5	51.1	60.1
	Bunny-2B	Qwen1.5-1.8B	2.6 M	76.6	34.2	59.6	53.2	65.0	57.7	68.6	66.5	57.8	48.5	60.3
	MoE-LLaVA-2B	Qwen1.5-1.8B	2.2 M	76.2	32.6	61.5	48.0	64.6	56.6	80.8	62.5	62.0	50.6	63.9
	LLaVA-2B (*SFT* model)	Qwen1.5-1.8B	1.2 M	73.6	31.2	57.9	50.6	61.2	54.9	73.3	63.9	58.5	47.2	60.7
	LLaVA-KD-2B	Qwen1.5-1.8B	1.2 M	78.8	**44.8**	**62.3**	53.4	68.9	61.6	75.3	63.6	57.9	49.3	61.5
	LLaVADI-2B†	Qwen1.5-1.8B	1.2 M	75.8	35.2	58.7	49.4	63.8	56.6	76.9	61.5	54.1	48.4	60.2
	LLaVA-MoD-2B†	Qwen1.5-1.8B	5.0 M	76.3	43.0	58.8	52.7	63.4	58.9	76.9	63.8	60.3	49.4	62.6
	CompoDistill-2B	Qwen1.5-1.8B	1.2 M	78.8	42.5	62.2	56.4	**70.1**	**61.9**	**82.9**	**69.4**	**63.3**	**51.2**	**66.7**

Table 2: Ablation for VAT and TAF modules.

Row	VAT module	TAF module	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
(a)	✗	✗	57.8	50.1	62.5	56.8	76.3	64.9	61.7	48.7	62.9
(b)	✓	✗	54.3	54.6	64.8	57.9	81.0	66.4	62.1	50.5	65.0
(c)	✗	✓	60.8	56.5	66.5	61.3	78.0	67.0	62.4	48.1	63.8
(d)	✓	✓	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7

(a)) follows the proposed three-stage framework (Section 4.3), using only the language modeling loss (Section 2) and logit-based KD loss (Equation 4). ***Effect of VAT module***: We observe that the VAT module significantly improves the performance especially on CR tasks (row (a) vs. (b) and row (c) vs. (d)), demonstrating the effectiveness of enhancing visual perception abilities. This confirms that explicit visual attention alignment is crucial for distilling visual perception ability, as discussed in Section 3. ***Effect of TAF module***: We observe that equipping the TAF module improves the performance on VQA and CR tasks (row (a) vs. (c) and row (b) vs. (d)), indicating that bridging the student's feature space with the imposed attention mechanism is crucial for effective knowledge transfer. Finally, the fully-fledged model (row (d)) achieves the best performance on both VQA and CR tasks, demonstrating the benefit of the VAT and TAF modules.

Table 3: Detailed ablation for VAT module. The blue line performs the same as CompoDistill.

(a) **Attention loss type**

Attention Loss	VQA Avg	CR Avg
✗	61.3	63.8
MSE	60.3	65.2
KL. Div.	60.7	65.5
Cos. Sim.	62.9	66.7

(b) **Target layers**

Target Layers	VQA Avg	CR Avg
Early (∼ 30%)	61.2	63.7
Later (70% ∼)	61.7	64.6
All	62.4	66.6
Intermediate	62.9	66.7

(c) **Layer matching strategy**

Match strategy	VQA Avg	CR Avg
Simple	61.5	65.6
Adaptive	62.0	65.7
Group	62.9	66.7

Fine-grained Analysis on VAT module. We perform fine-grained ablation studies to study the impact of specific design choices—namely, the attention loss type, the target matching layer, and the layer matching strategy—within the VAT module. **1)** We first analyze the impact of the ***attention loss function*** (Table 3a). While any form of attention loss improves CR over the baseline, using Cosine Similarity (Cos. Sim.) significantly outperforms both MSE and KL Divergence. This result suggests that it is more crucial for the student to learn the relative importance among visual patches rather than forcing an exact match of their absolute attention scores. **2)** Next, we investigate ***which layers to target for distillation*** (Table 3b). Performing distillation on the intermediate layers (30-70%) yields the highest performance. This finding confirms our analysis that *visual understanding layers* are crucial for visual-semantic integration and highlights the effectiveness of distilling specifically from these layers, a strategy consistent with prior research (Kaduri et al., 2024). **3)** Lastly, we evaluate different ***layer matching strategies*** (Table 3c). Our proposed Group matching achieves the best performance compared to both Simple matching (which uniformly samples teacher layers) and Adaptive matching (which finds optimal pairs based on layer distance). This suggests that grouping

layers provides a more stable and effective signal for transferring the teacher's complex attention behaviors, especially when student and teacher architectures differ in depth.[9]

Further Benefit on Other Complex Task. Beyond compositional reasoning, we explore a further benefit of enhancing visual perception for complex tasks. Specifically, we expect that CompoDistill can help mitigate the relational hallucinations, thanks to its ability to accurately understand object relationships, as discussed in Section 1. To test this, we evaluate CompoDistill on the R-Bench (Wu et al., 2024) and Reefknot (Zheng et al., 2025) benchmarks, both of which focus on evaluating relational hallucinations. As shown in Table 4, CompoDistill significantly outperforms other KD methods and achieves performance nearly on par with the teacher. The results highlight that enhancing the visual perception ability can be beneficial to not only compositional reasoning tasks but also other complex tasks that require precise understanding of relationships among objects.

Table 4: Comparison for relational hallucination using F1 score.

Model	R-Bench ↑	Reefknot ↑
Teacher (LLaVA-4B)	79.1	67.9
Student (LLaVA-2B)	74.3	61.3
LLaVA-KD-2B	76.5	60.3
LLaVA-MoD-2B	76.2	63.4
CompoDistill-2B	78.6	66.7

6 Scalability Experiments

Richer Data Improves Distillation Performance. We analyze the effect of data scale in Table 5. Performance improves significantly when moving from SFT to distillation, and increases further when the training dataset size is doubled[10]. This highlights that both the quality and quantity of training data are crucial for effective knowledge transfer, with CR showing particular sensitivity to data scaling.

Table 5: Data Scaling Experiment.

Student	# Sample	VQA Avg	CR Avg
Qwen1.5-1.8B (SFT)	1.2 M	56.6	60.7
Qwen1.5-1.8B	1.2 M	62.9	66.7
Qwen1.5-1.8B	2.4 M	63.3	69.9

Larger Teachers Produce Stronger Students. Next, we examine the influence of teacher model size in Table 6. Our experiments show that students consistently benefit from larger teachers, regardless of the student's own size. For instance, a 1.8B student distilled from a 7B teacher outperforms one distilled from a 4B teacher. This demonstrates that a higher-capacity teacher is crucial for transferring stronger reasoning abilities, providing a more effective source of knowledge for the student.

Table 6: Experiments with different size of teacher/student.

Student	Teacher	VQA Avg	CR Avg
Qwen1.5-1.8B (SFT)		56.6	60.7
Qwen1.5-1.8B	Qwen1.5-4B	62.9	66.7
Qwen1.5-1.8B	Qwen2.5-7B	63.4	67.8
Qwen1.5-0.5B (SFT)		52.0	48.4
Qwen1.5-0.5B	Qwen1.5-1.8B	54.7	51.1
Qwen1.5-0.5B	Qwen1.5-4B	56.6	54.5

Our Distillation Method Generalizes Across Backbones. Finally, we test our framework's generalizability by replacing the Qwen backbone with the MobileLLaMA family (MLLaMA) in Table 7. Our distillation method remains effective even with this entirely different architecture. This result confirms the robustness and generalizability of our approach, demonstrating that its principles are not tied to a specific model family but are broadly applicable.

Table 7: Experiments with a different LLM backbone.

Student	Teacher	VQA Avg	CR Avg
MLLaMA-1.7B (SFT)		49.7	43.7
MLLaMa-1.7B	MLLaMa-3B	53.1	48.9

A comprehensive discussion of related works is provided in Appendix I.

7 Conclusion

In this work, we aim to enhance the visual perception abilities of Knowledge Distillation(KD)-based Multimodal Large Language Models (MLLMs), which has been largely overlooked by the previous KD studies. To this end, we conduct a systematic analysis and identify visual attention misalignment as a key factor hindering the effective distillation of visual perception from teacher to student. Building on this analysis, we propose CompoDistill, a novel KD framework that incorporates a Visual ATtention alignment (VAT) module to explicitly address this misalignment. Furthermore, we introduce the Teacher Adapter Fetch (TAF) module to ensure that teacher-imposed attention mechanism is compatible with the student's feature space, making synergy with VAT module. Through extensive experiments on VQA and CR benchmarks, we demonstrate that CompoDistill significantly enhances visual perception abilities while preserving strong visual recognition abilities, as achieved in existing KD works. Regarding the limitation and future work, please refer to Appendix J.

We believe that this work provides a novel perspective on the student's visual attention misalignment and makes a contribution to the pursuit of efficient MLLMs, especially in KD-based research, by establishing the first dedicated direction toward enhancing visual perception abilities.

[9] For a detailed explanation of the layer matching strategies, see Appendix F.

[10] For details on the training data, see Appendix G.

Ethics Statement

Our research contributes to the development of more efficient and accessible Multimodal Large Language Models (MLLMs). By enabling the creation of smaller, yet highly capable student models, our work helps lower the significant computational barriers to deploying advanced AI, promoting its broader adoption. In compliance with the ICLR Code of Ethics, and to the best of our knowledge, we have not faced any ethical issues during this research. Additionally, all datasets and baselines used in our experiments are freely accessible to the public.

Reproducibility Statement

To ensure reproducibility of experiment results throughout the paper, we describe the details of experimental setting and training details in 5.1 and Appendix E, respectively. Furthermore, we provide a source code in https://anonymous.4open.science/r/CompoDistill-D07F.

References

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

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342, 2010.

Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1511–1520, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.130/.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm, 2024. URL https://arxiv.org/abs/2312.06742.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*, 2025a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025b. URL https://arxiv.org/abs/2503.01773.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. Mobilevlm v2: Faster and stronger baseline for vision language model, 2024. URL https://arxiv.org/abs/2402.03766.

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension, 2017. URL https://arxiv.org/abs/1710.10723.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4178–4188, 2025.

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. URL https://arxiv.org/abs/2306.13394.

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*, pp. 6904–6913, 2017.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Drew A Hudson and Christopher D Manning. 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*, pp. 6700–6709, 2019.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens, 2025. URL https://arxiv.org/abs/2411.16724.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020. URL https://arxiv.org/abs/1909.10351.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models, 2024. URL https://arxiv.org/abs/2411.17491.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering, 2018. URL https://arxiv.org/abs/1801.08163.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models, 2024. URL https://arxiv.org/abs/2404.07204.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022. URL https://arxiv.org/abs/2111.15664.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

Byung-Kwan Lee, Ryo Hachiuma, Yong Man Ro, Yu-Chiang Frank Wang, and Yueh-Hua Wu. Genrecal: Generation after recalibration from large to small vision-language models. *arXiv preprint arXiv:2506.15681*, 2025a.

Byung-Kwan Lee, Ryo Hachiuma, Yu-Chiang Frank Wang, Yong Man Ro, and Yueh-Hua Wu. Vlsi: Verbalized layers-to-interactions from large to small vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29545–29557, 2025b.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024a.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26689–26699, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL https://arxiv.org/abs/2403.05525.

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, 2022. URL https://arxiv.org/abs/2209.09513.

Teli Ma, Rong Li, and Junwei Liang. An examination of the compositionality of large generative vision-language models. *arXiv preprint arXiv:2308.10509*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL https://arxiv.org/abs/1906.00067.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Bivlc: Extending vision-language compositionality evaluation with text-to-image retrieval. *Advances in Neural Information Processing Systems*, 37:101880–101904, 2024.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025. URL https://arxiv.org/abs/2410.07149.

OpenAI. Gpt-4v(ision) technical work and authors, 2023. URL https://openai.com/contributions/gpt-4v/. Accessed: 2025-09-18.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models, 2025. URL https://arxiv.org/abs/2503.17349.

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

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

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17097–17107, 2023.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. URL https://arxiv.org/abs/2206.01718.

Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, Zhenzhong Kuang, and Jiajun Ding. Imp: Highly capable large multimodal models for mobile devices. *IEEE Transactions on Multimedia*, 2025.

Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Lei Zhang, Guanghao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024.

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

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression, 2019. URL https://arxiv.org/abs/1908.09355.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024. URL https://arxiv.org/abs/2311.03079.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models, 2024. URL https://arxiv.org/abs/2406.16449.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. Llavadi: What matters for multimodal large language models distillation. *arXiv preprint arXiv:2407.19409*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, Sunghwan Hong, and Seungryong Kim. Visual representation alignment for multimodal large language models, 2025. URL https://arxiv.org/abs/2509.07979.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. URL https://arxiv.org/abs/1608.00272.

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

Tianyang Zhao, Kunwar Yashraj Singh, Srikar Appalaraju, Peng Tang, Vijay Mahadevan, R Manmatha, and Ying Nian Wu. No head left behind–multi-head alignment distillation for transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7514–7524, 2024.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models, 2025. URL https://arxiv.org/abs/2408.09429.

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18270–18280, 2025.

Supplementary Material

- CompoDistill: Attention Distillation for Compositional Reasoning in Multimodal LLMs -

A Additional Visualizations on Attention Misalignment

In this section, we provide additional qualitative examples of the **attention misalignment** between the student and the teacher model, a concept introduced in Section 1. To facilitate comprehension, we have highlighted crucial textual components in red, specifically phrases that require relational understanding or accurate attribute recognition. For a comprehensive comparison, we present results for an additional student model (LLaVADI-2B) and our method (CompoDistill), alongside the baseline LLaVA-KD-2B and LLaVA-4B (Teacher). Notably, the visualizations show that CompoDistill produces an attention map remarkably similar to the teacher's, correctly focusing on the key visual areas relevant to the text. These comparisons are illustrated in Figure 6 and Figure 7.

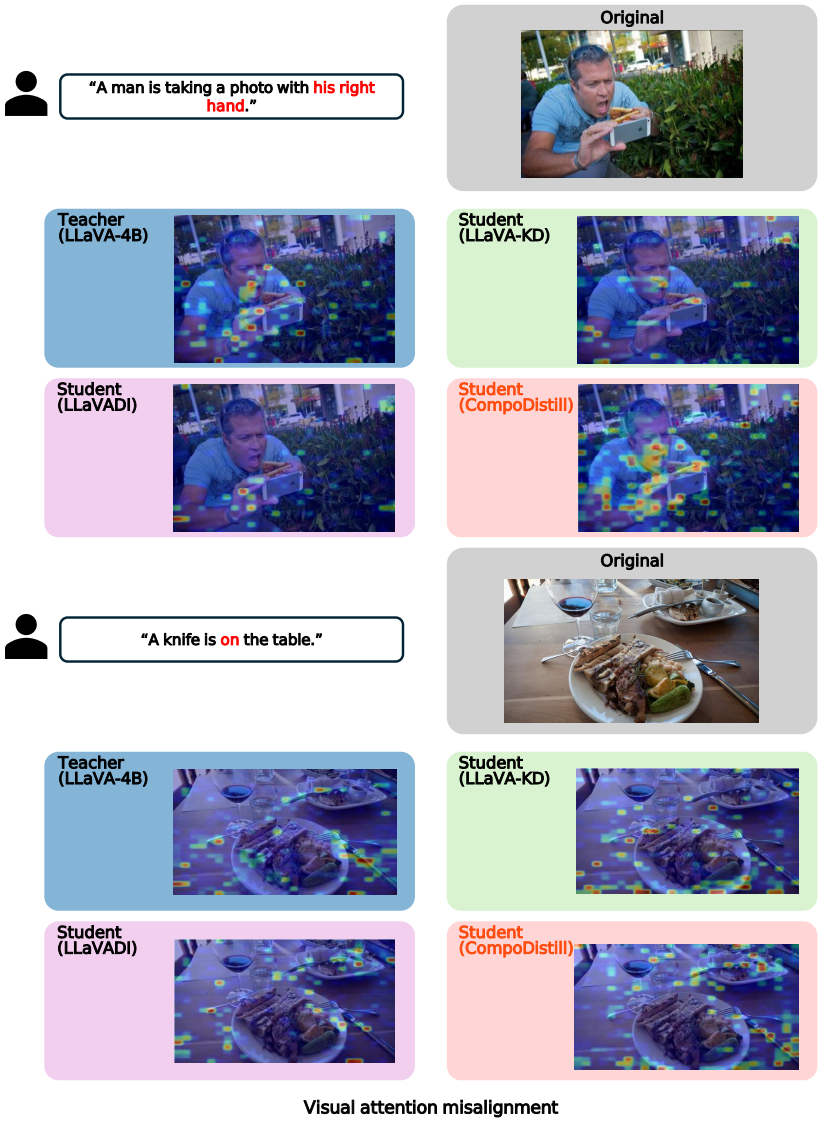

Figure 6: Examples of attention misalignment.

B Differences between Visual Recognition and Perception Abilities

This section provides a detailed explanation of the distinction between visual recognition and visual perception abilities, which are evaluated using Visual Question Answering (VQA) and Composi-

Figure 7: Examples of attention misalignment.

tional Reasoning (CR) datasets, respectively. Also in Figure 8, we illustrated an example of VQA and CR question.

B.1 Visual Recognition

Visual recognition is the foundational ability of a model to **identify and categorize objects, scenes, and their basic attributes within an image.** It fundamentally answers the question, such as "What is in this image?" This process relies on matching learned visual patterns—such as textures, shapes, and colors—to specific labels or concepts. For example, when a model identifies a four-legged furry animal as a "dog," it is performing visual recognition. This ability is analogous to building a vocabulary of the visual world.

Most standard VQA datasets (e.g., VQAv2 (Goyal et al., 2017), Vizwiz (Bigham et al., 2010), and GQA (Hudson & Manning, 2019)) are primarily designed to evaluate this recognition capability. The questions in these datasets are typically direct and fact-based, probing for the presence, count, or simple properties of objects. They can often be answered correctly if the model successfully

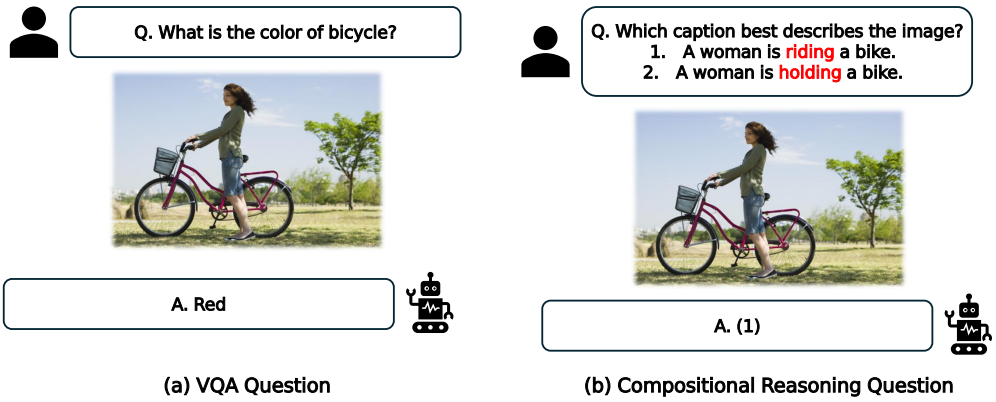

Figure 8: An example of a standard VQA question, which requires simple object identification, versus a Compositional Reasoning (CR) question, which requires accurately distinguishing between two detailed and confusable multiple choices.

recognizes the key objects and their most salient features, without needing to understand complex inter-object relationships.

B.2 Visual Perception

Visual perception is a more advanced cognitive ability that goes beyond simple identification. It involves **interpreting and understanding the relationships between objects, their precise attributes, their spatial arrangement, and the overall context of the scene**. If recognition is about "what," perception is about "how" and "why"—how objects are arranged, how they interact, and why the scene is composed in a particular way. This requires the model to not just list the contents of an image, but to build a coherent, structured understanding of it.

Compositional Reasoning (CR) datasets are specifically designed to evaluate this perceptual ability. The questions are structured to be challenging for models that rely solely on simple keyword matching or recognition. To answer correctly, a model must accurately bind specific attributes to their corresponding objects and correctly interpret the spatial or semantic relationships between them. These questions often test a model's robustness where the correct objects and attributes are present but not in the configuration described by the question.

C Additional Experiments on Attention Similarity

To further validate our findings in Section 3.1, we extend our attention similarity analysis to two additional benchmark datasets: VQAv2 (Goyal et al., 2017) for VQA and Winoground (Thrush et al., 2022) for CR, replicating the experimental setup used for GQA (Hudson & Manning, 2019) and SugarCrepe (Hsieh et al., 2023).

The results, visualized in Figure 9, demonstrate a consistent trend with the conclusions drawn in our main analysis. The analysis reaffirms the critical role of attention over visual tokens in performance improvement. Specifically, on the Winoground (Thrush et al., 2022) (CR) task, where the student model's performance gain over the SFT baseline was marginal, we again observe no significant increase in teacher-student attention similarity. This result corroborates our main argument that higher attention similarity over visual tokens in the visual understanding layers is a key factor for the effective distillation of visual perception.

D Baseline Methods

We include several baselines from two main groups. The first group consists of Knowledge Distillation-based methods, including LLaVADI-2B (Xu et al., 2024), LLaVA-KD-2B (Cai et al., 2024), and LLaVA-MoD-2B (Shu et al., 2024). **To ensure fair comparisons, all KD models were dis-**

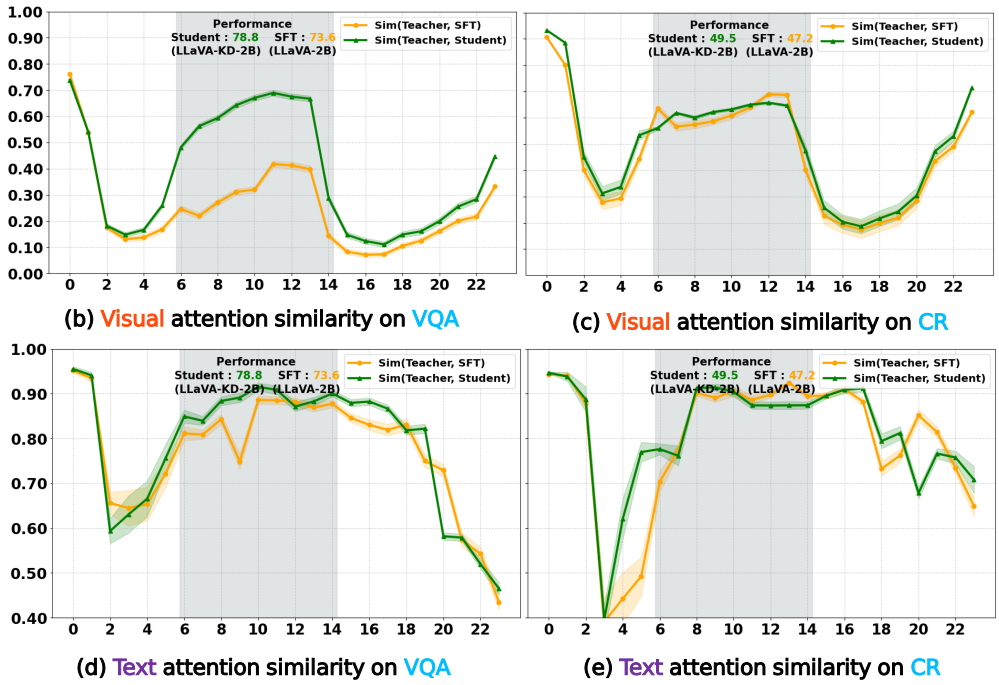

Figure 9: Layerwise attention similarity of visual tokens and text tokens between student/SFT models and the teacher model.

tilled from the same teacher model (i.e., LLaVA-4B) and share the same LLM backbone (i.e., Qwen 1.5). The second group comprises a broad range of General MLLMs with parameter sizes in the range of 1.3B–3B, comparable to that of the compared KD methods, for direct performance comparison. This includes models such as Imp-2B (Shao et al., 2025), Bunny-2B (He et al., 2024), MoE-LLaVA-2B (Lin et al., 2024a), LLaVA-2B (Liu et al., 2024), and Deepseek-VL-1.3B (Lu et al., 2024), MobileVLM-1.7B (Chu et al., 2024), as well as larger models (4B-7B) like MiniCPM-V-2.4B (Hu et al., 2024), CogVLM-7B (Wang et al., 2024) and Qwen2.5-VL-7B (Bai et al., 2025) to provide state-of-the-art context.

E Additional Implementation Details

E.1 Training Strategy and Hyper-parameters

Our model is initialized with a SigLIP-B/14@384 vision encoder and a Qwen1.5-1.8B LLM. For adapter, we use 3-layer MLP. Across all training stages, we use a consistent setup: Each stage is trained for one epoch using AdamW optimizer (Loshchilov & Hutter, 2017). The AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a cosine decay learning rate scheduler, a weight decay of 0.0, and a warm-up ratio of 0.03. We process images at a resolution of 384x384, with input sequence lengths of 729 for the vision encoder and 2048 for the LLM. All training is performed with Float16 numerical precision and Zero2 model parallelism with 8 NVIDIA L40S 48GB GPUs. For the DPT stage, we use the LLaVA-1.5-558K (Liu et al., 2024) dataset with a batch size of 256 and a learning rate of 1e-3. For the DFT and SFT stages, we use the LLaVA-mix-665K (Liu et al., 2024) dataset with a batch size of 128 and a learning rate of 1e-4.

Our training is divided into three distinct stages, each lasting for one epoch:

Distilled Pre-Training (DPT): In this initial stage, our primary goal is to align the visual representations with the language embedding space. To achieve this, we freeze both the vision encoder and the LLM, and exclusively optimize the parameters of the adapter. This stage uses a global batch size of 256.

Distilled Fine-Tuning (DFT): The scope of training is expanded in the second stage to distill knowledge more deeply into the language model. The vision encoder remains frozen, but we now co-optimize both the LLM and the adapter. For this stage, the global batch size is reduced to 128.

Supervised Fine-Tuning (SFT): In the final stage, the model, initialized from the DFT checkpoint, is further refined. The training configuration mirrors the DFT stage: the vision encoder is frozen, while the LLM and adapter are co-optimized. This stage also uses a global batch size of 128 to complete the training process.

The detailed training hyper-parameters are show in Table 8.

Table 8: Training hyper-parameters of each stage.

Configuration	**Distilled Pre-Training**	**Distilled Fine-Tuning**	**Supervised Fine-Tuning**
LLM	✗	✓	✓
Vision Encoder	✗	✗	✗
Adapter	✓	✓	✓
LLM init.	Qwen1.5-1.8B	Qwen1.5-1.8B	From DFT
Vision Encoder init.		SigLIP-B/14@384	
Image resolution		384 x 384	
Vision Encoder Sequence length		729	
LLM sequence length		2048	
Optimizer		AdamW	
Optimizer hyper-parameter		$\beta_1 = 0.9, \beta_2 = 0.98$	
Learning rate		2e-4	
Learning rate scheduler		Cosine decay	
Weight decay		0.0	
Training epoch		1.0	
Warm-up ratio		0.03	
Global batch size	256	128	128
Numerical precision		Float16	
Model parallelism		Zero2	

F Explanation of the Baselines for Layer Matching Strategy

In this section, we provide a detailed explanation of the different layer matching strategies evaluated in our main ablation study (Table 3c). Layer matching is a critical component of attention distillation, as it defines the correspondence between teacher and student layers for knowledge transfer. While our proposed method, **Group matching**, proved to be the most effective, we also explored two alternative baseline strategies to thoroughly investigate the design space. We will describe each of these in turn:

First, **Simple matching**, a straightforward approach that uniformly samples teacher layers to match the number of student layers. Second, **Adaptive matching**. Expanding on (Lee et al., 2025b), this method first computes a matrix of Kullback-Leibler (KL) distances between every student target layer and every teacher target layer. Each student layer is then greedily paired with the teacher layer that has the minimum KL distance, under a consecutive constraint. This constraint ensures that a given student layer can only select a teacher layer that comes after the one selected by the previous student layer, thereby maintaining the sequential integrity of the model's architecture.

G Explanation for the Extended Training Data

This section provides detailed information on the training data used for the data scaling component of our main scalability experiments. Our baseline training utilizes a base dataset of approximately 1.2M samples, comprising the **LLaVA-1.5-558K** (Liu et al., 2024) for Distilled Pre-Training (DPT) and **LLaVA-mix-665K** (Liu et al., 2024) for the Distilled Fine-Tuning (DFT) and Supervised Fine-Tuning (SFT) stages.

To scale the data in Sec 6, we incorporate an additional 1.2M samples, originally curated for the Dense-to-Sparse Distillation (D2S) in LLaVA-MoD (Shu et al., 2024). As detailed in Table 9, this extended dataset is a diverse mixture covering a wide range of tasks, including General QA, Grounding, Science, Chart & Document understanding, OCR, and Knowledge-based reasoning.

Table 9: Training dataset of scaling data experiment.

Stage	Task	Dataset
Distilled Pre-Training (DPT)	Captioning	LLaVA-1.5-558K (Liu et al., 2024)
Distilled Fine-Tuning (DFT)	Conversation	LLaVA-mix-665K (Liu et al., 2024)
Supervised Fine-Tuning (SFT)	Conversation	LLaVA-mix-665K (Liu et al., 2024)
Data Scaling Experiment	Captioning	ShareGPT4V-100K, TextCaps
	Conversation	LLaVA-mix-665K (Liu et al., 2024)
	General QA	GQA (Hudson & Manning, 2019), VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019)
	Grounding	VG (Krishna et al., 2016), RefCoCo (Yu et al., 2016)
	Science	AI2D (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022)
	Chart & Doc	DVQA (Kafle et al., 2018), ChartQA (Masry et al., 2022), DocQA (Clark & Gardner, 2017)
	OCR	OCRVQA (Mishra et al., 2019), SynthDoG-EN (Kim et al., 2022)
	Knowledge	A-OKVQA (Schwenk et al., 2022), GeoQA+ (Cao & Xiao, 2022)

H Additional Experiments Results

H.1 Detailed Experiment Results on VAT module

H.1.1 Detailed Experiment Results on Attention Loss Type

Table 10: Ablation for various attention loss type.

Attention Loss Type	Attention Matching	Attention Layers	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Ablation on attention loss type											
w/o Attention Loss			60.8	56.5	66.5	61.3	78.0	67.0	62.4	48.1	63.8
MSE.	Group	intermediate	59.9	55.3	65.7	60.3	79.9	68.3	62.8	49.7	65.2
KL. Div.	Group	intermediate	60.7	55.6	65.9	60.7	80.1	67.9	63.5	50.5	65.5
Cos. Sim.	Group	intermediate	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7

We provide a detailed analysis of the impact of different attention loss functions, as summarized in Table 10. Our goal was to determine the most effective way to quantify the difference between student and teacher attention maps for distillation. The results show that while all tested loss functions improve performance on Compositional Reasoning (CR) tasks over the baseline, their effects on Visual Question Answering (VQA) are mixed. Notably, both Mean Squared Error (MSE) and KL Divergence (KL. Div.) slightly degrade VQA performance.

Cosine Similarity (Cos. Sim.) proved to be the most effective loss function, improving performance across both the VQA and CR domains. We believe this is because it is more crucial for the student model to learn the relative importance of different visual patches, rather than replicating the exact absolute values (MSE) or the probability distribution (KL. Div.) of the teacher's attention scores.

H.1.2 Detailed Experiment Results on Attention Target Layers

Table 11: Ablation for attention target layers.

Attention Loss Type	Attention Matching	Attention Layers	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Ablation on target layers for attention distillation											
Cos. Sim.	Group	early (∼ 30%)	61.3	55.5	66.8	61.2	80.9	67.4	59.8	46.7	63.7
Cos. Sim.	Group	later (70% ∼)	61.2	55.7	68.4	61.7	81.8	66.8	59.7	50.0	64.6
Cos. Sim.	Group	all	62.1	55.8	69.1	62.4	83.1	68.6	63.7	50.4	66.6
Cos. Sim.	Group	intermediate	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7

We investigate which layers are the most effective targets for attention distillation, with detailed results in Table 11. We compared four strategies: distilling from early layers (∼30%), later layers (70%∼), all layers, and our primary approach of targeting the intermediate layers (**visual understanding layers**).

The results clearly indicate that targeting the visual understanding layers yields the best overall performance on both VQA and CR tasks. While distilling from all layers provides strong and sometimes comparable results, this approach is less computationally efficient. The more focused intermediate

strategy ultimately achieves a better trade-off, slightly outperforming the all-layer approach in overall scores without the associated computational overhead. Conversely, targeting only the early or later layers leads to a noticeable drop in performance.

This confirms our analysis that the intermediate layers, which function as the core **visual understanding layers**, are the most critical for visual-semantic integration. Distilling specifically from this block provides the most potent and effective signal for knowledge transfer, a finding consistent with prior research (Kaduri et al., 2024; Neo et al., 2025).

H.1.3 Detailed Experiment Results on Layers Matching Strategy

Table 12: Ablation for attention matching strategy.

Attention Loss Type	Attention Matching	Attention Layers	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Ablation on layers matching for attention distillation											
Cos. Sim.	Simple	intermediate	60.2	56.3	68.0	61.5	81.6	67.1	63.2	50.7	65.6
Cos. Sim.	Adaptive	intermediate	61.4	55.9	68.6	62.0	82.1	68.3	62.8	49.5	65.7
Cos. Sim.	Group	intermediate	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7

This part details our ablation on different strategies for matching student and teacher layers, with full results in Table 12 and detailed explanation about each strategy is in Section F. We compared our proposed **Group matching** against two strong baselines: **Simple matching** (uniform sampling of teacher layers) and **Adaptive matching** (pairing based on feature distance). The data shows that while both Simple and Adaptive strategies improve performance over a no-distillation baseline, our Group matching consistently achieves the best results across all benchmarks. Notably, our method demonstrates a clear performance advantage over the next best method, Adaptive matching. Furthermore, our approach is more computationally efficient, as it avoid the complex calculations required to dynamically match layers inherent to adaptive strategies. This suggests that grouping layers provides a more stable and robust signal for the student. By averaging the behavior of a block of teacher layers, our method likely smooths out layer-specific idiosyncrasies and transfers a more generalized attention strategy.

H.2 Detailed Experiment Results on Relational Hallucination

Table 13: Relational Hallucination Experiment.

Model	R-Bench			Reefknot		
	Precision	Recall	F1 Score	Perception	Cognitive	Total
Teacher (LLaVA-4B)	66.4	97.8	79.1	45.3	78.0	67.9
Studnet (LLaVA-4B)	60.5	96.3	74.3	40.1	70.8	61.3
LLaVA-KD-2B	63.5	96.2	76.5	41.8	68.8	60.3
LLaVA-MoD-2B	62.5	97.6	76.2	42.0	73.1	63.4
CompoDistill-2B	65.7	97.6	78.6	43.2	77.3	67.9

As mentioned in the main text, we conducted experiments on relational hallucination benchmarks to demonstrate a further benefit of our proposed method. Table 13 presents the detailed quantitative results of this evaluation on the R-Bench (Wu et al., 2024) and Reefknot (Zheng et al., 2025) benchmarks.

The results substantiate our claim, showing that CompoDistill significantly outperforms other knowledge distillation methods on the R-Bench benchmark and nearly closes the performance gap to the teacher model. More strikingly, on the Reefknot benchmark, our method achieves performance on par with the teacher model, showcasing its ability to handle complex relational challenges. These findings provide strong evidence that enhancing visual perception via CompoDistill effectively mitigates relational hallucinations and boosts performance on complex visual reasoning tasks.

H.3 (Additional) Detailed Experiments Results on Training Strategy.

In this section, we provide additional experimental results to analyze the effectiveness of our proposed three-stage training strategy (DPT-DFT-SFT). To demonstrate its efficacy, we investigate the specific impact of each individual stage, with detailed results presented in Table 14.

First, the Distilled Pre-Training (DPT) stage shows a significant impact, particularly on Visual Question Answering (VQA) tasks. As seen by comparing configurations with and without DPT (e.g., (1) vs. (3)), replacing standard pre-training with our distilled approach consistently yields substantial improvements in VQA scores. This highlights DPT's role in building a strong visual knowledge from the teacher.

Next, the Distilled Feature-Tuning (DFT) stage is crucial for enhancing Compositional Reasoning (CR) capabilities. The inclusion of DFT leads to the most significant gains in CR performance across all benchmarks. We attribute this improvement to the effective transfer of the teacher's fine-grained attention patterns through our attention distillation process, which is vital for understanding complex object relationships.

Finally, the Supervised Fine-Tuning (SFT) stage is indispensable. The results from the configuration without SFT (4) show a catastrophic failure on VQA tasks, as the model completely fails to follow task instructions. This demonstrates that SFT is an absolutely critical final step for aligning the model's distilled knowledge with the specific formats and demands of downstream tasks.

Table 14: Ablation for training recipe and teacher adapter fetch module. †:Fail to follow instructions (i.e., answer only single word).

	Training Recipe	Teacher Adapter	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Ablation on different training recipe											
(1)	PT + SFT	✓	57.9	50.6	61.2	56.6	73.3	63.9	58.5	47.2	60.7
(2)	PT + DFT	✓	59.5	54.6	64.8	59.6	79.1	67.3	61.8	50.3	64.6
(3)	DPT + SFT	✓	60.4	56.3	64.4	60.4	72.6	61.6	58.3	49.9	60.6
(4)	DPT + DFT	✓	1.8†	28.0†	34.9†	21.6†	70.2	60.3	57.3	50.8	59.7
(5)	DPT + SFT + DFT	✓	60.6	56.1	65.1	60.6	81.1	67.9	63.2	48.9	65.3
(*)	DPT + DFT + SFT	✓	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7

H.4 Detailed Experiments Results on Scalability

H.4.1 Detailed Experiments Results on Data Scaling

This section provides a detailed analysis of our experiments on scaling data quality and data quantity, with full results presented in Table 15. For a clear point of comparison, we also include the performance of a baseline model trained only with Supervised Fine-Tuning (SFT) on the same initial data volume.

Our findings clearly demonstrate the two-fold benefit of our approach. First, by using our distillation method, the model's performance significantly improves over the SFT baseline, even with the same amount of data. This confirms that distillation provides a higher-quality training signal. Second, when we use more of this high-quality data (doubling the training samples), the model's performance is further enhanced. These results validate that our distillation approach offers a substantial performance gain, which is then amplified by increasing the quantity of the training data. A detailed explanation of the training dataset configuration is provided in Appendix G.

Table 15: Data scaling experiment.

Student LLM	Teacher LLM	# Training Samples	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Data scaling experiment											
Qwen1.5-1.8B (SFT)		1.2 M	57.9	50.6	61.2	56.6	73.3	63.9	58.5	47.2	60.7
Qwen1.5-1.8B	Qwen1.5-4B	1.2 M	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7
Qwen1.5-1.8B	Qwen1.5-4B	2.4 M	62.7	56.8	70.6	63.3	89.9	67.8	68.9	53.3	69.9

H.4.2 Detailed Experiments Results on Student and Teacher LLM Size

This section provides a more detailed analysis of the relationship between teacher and student model sizes in our distillation framework, with full results presented in Table 16.

We first examine a 1.8B parameter student model. As the table shows, distilling from a larger 7B teacher yields a stronger student model compared to distilling from a 4B teacher, with improved performance on both VQA and CR tasks. It is worth noting that we selected a Qwen2.5-7B model as the larger teacher. This decision was made because preliminary evaluations showed that the performance of the Qwen1.5-7B model was not significantly higher than that of the Qwen1.5-4B version;

using the more capable Qwen2.5 architecture ensured a more significant and meaningful gap in teacher capacity for this experiment.

This trend is further validated with a smaller, 0.5B parameter student. The results show a clear progression: the 0.5B student first shows a dramatic improvement over its SFT-only baseline when distilled from a 1.8B teacher. Performance increases again, and quite substantially, when the same 0.5B student is distilled from an even larger 4B teacher. This demonstrates that even smaller student models can effectively absorb the enhanced capabilities of higher-capacity teachers.

In summary, these experiments consistently show that a larger, more capable teacher model is a critical factor in producing a stronger student, regardless of the student's own size. A higher-capacity teacher provides a richer and more effective source of knowledge, successfully transferring more powerful reasoning abilities through our distillation process.

Table 16: Student and teacher LLM Size experiment.

Student LLM	Teacher LLM	# Training Samples	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Large teacher experiment											
Qwen1.5-1.8B (SFT)		1.2 M	57.9	50.6	61.2	56.6	73.3	63.9	58.5	47.2	60.7
Qwen1.5-1.8B	Qwen1.5-4B	1.2 M	62.2	56.4	70.1	62.9	82.9	69.4	63.3	51.1	66.7
Qwen1.5-1.8B	Qwen2.5-7B	1.2 M	62.9	57.1	69.8	63.4	84.1	70.7	63.8	52.6	67.8
Qwen1.5-0.5B (SFT)		1.2 M	54.9	45.5	55.5	52.0	52.9	51.4	39.5	49.8	48.4
Qwen1.5-0.5B	Qwen1.5-1.8B	1.2 M	57.3	48.9	58.1	54.7	57.3	52.8	44.0	50.4	51.1
Qwen1.5-0.5B	Qwen1.5-4B	1.2 M	60.1	50.3	59.5	56.6	59.5	59.0	49.1	50.5	54.5

H.4.3 Detailed Experiments Results on different LLM Backbones

To verify the generalizability of our distillation framework, we conducted an experiment using an entirely different architectural backbone, the MobileLLaMA (Chu et al., 2024) family. This section provides a detailed analysis of these results, which are presented in the Table 17.

Specifically, we applied our distillation method to a 1.4B parameter MobileLLaMA as the student model, using a 2.7B parameter MobileLLaMA model as the teacher. We then compared its performance to a baseline 1.4B student trained with only Supervised Fine-Tuning (SFT). The results clearly show that our distillation method remains highly effective on this new architecture. The distilled student model significantly outperforms the SFT-only baseline across both Visual Question Answering (VQA) and Compositional Reasoning (CR) benchmarks.

This successful application demonstrates the robustness and architectural independence of our approach. It confirms that the core principles of our distillation method are not specifically tailored to the Qwen (Qwen et al., 2025) model family but can be broadly applied to improve the performance of different MLLM backbones, validating the generalizability of our findings.

Table 17: Different LLM backbones experiment.

Student LLM	Teacher LLM	# Training Samples	Visual Question Answering				Compositional Reasoning				
			GQA	TextVQA	MME	Avg	Sugarcrepe	SADE	BiVLC	Winoground	Avg
Different LLM backbone experiment											
MobileLLaMA-1.4B (SFT)		1.2M	53.3	43.8	52.2	49.7	50.1	50.0	37.5	37.5	43.7
MobileLLaMA-1.4B	MobileLLaMA-2.7B	1.2M	57.4	45.2	56.8	53.1	55.6	54.2	39.8	46.5	48.9

I Related Works

Multimodal Large Language Models. Visual instruction tuning (Liu et al., 2023) has propelled MLLMs to strong performance on diverse benchmarks (Chen et al., 2024; Yang et al., 2024; OpenAI, 2023), yet persistent weaknesses remain in fine-grained visual tasks (Tong et al., 2024; Qi et al., 2025). Early efforts to address these limitations focused on scaling up vision encoders (Lu et al., 2024; Kar et al., 2024) or designing more expressive projectors (Cha et al., 2024; Liu et al., 2024). More recently, the focus has shifted to the internal information flow, with research identifying critical failures such as misaligned attention (Jiang et al., 2025; Neo et al., 2025; Darcet et al., 2024) and the dilution of visual features in intermediate layers (Kaduri et al., 2024; Yoon et al., 2025; Chen et al., 2025b). In contrast to these approaches, our analysis focuses on the visual attention dynamics between student and teacher models trained via direct knowledge distillation.

Knowledge Distillation. Knowledge distillation (KD) compresses a large teacher model into a smaller, efficient student (Hinton et al., 2015), a technique widely applied to LLMs via logit matching (Sun et al., 2019; Jiao et al., 2020). In the multimodal domain, methods have adapted this for visual grounding (Cai et al., 2024; Feng et al., 2025) or used modular strategies to overcome architectural limits (Shu et al., 2024). The scope has since expanded to distillation across model families (Lee et al., 2025b;a) and a deeper focus on internal mechanics. For instance, recent alignment-oriented methods like VIRAL (Yoon et al., 2025) regularize intermediate representations to preserve visual semantics, moving beyond simple logit-level supervision. While these studies primarily propose new distillation techniques, our work takes a different approach. We instead provide a detailed analysis to identify the specific bottlenecks that hinder the effectiveness of knowledge distillation in MLLMs.

Compositional Reasoning Benchmarks. Compositional reasoning, the ability to understand the interplay of objects and their relations, remains a significant hurdle for vision–language models. Initial benchmarks like Winoground (Thrush et al., 2022) first exposed these weaknesses in early models. Building on this, a new generation of more robust benchmarks has emerged to address evaluation biases. These include SugarCrepe (Hsieh et al., 2023), which uses LLM-generated hard negatives; SADE (Ma et al., 2023), which specifically diagnoses and mitigates the language bias found in generative models through a debiased test set that neutralizes syntactic shortcuts; and BiVLC (Miranda et al., 2024), which focuses on bidirectional retrieval with synthetic negatives.

J Limitation and Future work

A potential limitation of CompoDistill lies in its inability to capture the distinct information carried by each head within multi-head attention (Zhao et al., 2024; Kang et al., 2025), as it distills only the average of the teacher's visual attention across all heads. This simplification may lead to information loss by overlooking the diverse roles played by individual heads. Moreover, CompoDistill assumes that the teacher and student MLLMs belong to the same LLM series, as consistency in vocabulary is required for logit-based distillation (Equation 4 in the main paper). This constraint poses a challenge when attempting to distill visual knowledge from teachers belonging to different model families.

As future work, we aim to incorporate the head-specific characteristics of the teacher's visual attention into the distillation process, enabling the student to capture more fine-grained and nuanced visual cues.