SCALABLE INFORMATION FLOW DISTILLATION FOR LARGE MULTIMODAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The application of Multimodal Large Language Models (MLLMs) in resource-constrained scenarios is significantly limited due to their enormous parameter size and computational complexity. Knowledge distillation aims to transfer knowledge from large models to smaller ones, striking a balance between performance and capacity. Recent works have gradually shifted their focus from response-based distillation to intermediate-layer-based distillation, with the core objective being to explore the intrinsic working mechanisms of MLLMs. In this paper, we propose a new distillation paradigm based on the information flow of MLLMs. Specifically, through systematic analysis, we reveal the information flow pathways between text, images and the last token (used for generation) as they pass through the intermediate layers. Consequently, we identify a three-stage workflow for MLLMs: understanding the text firstly, then processing the multimodal data and finally focusing on generating a response. Moreover, this information flow exhibits a proportional scaling property with respect to the number of layers across models of different sizes. Based on these findings, we propose a simple distillation scheme that compresses the information flow of large models into small models. Extensive experiments verify the effectiveness of the proposed method.

1 INTRODUCTION

Multimodal large language models (MLLMs) Wang et al. (2024); Liu et al. (2024b) have demonstrated outstanding performance in various multimodal tasks, such as visual question answering (VQA), thanks to the scaling law. However, their performance requires not only vast amounts of training data but also the support of a large number of parameters and significant computational resources. In resource-constrained edge scenarios, such as those with limited storage space and weak computing cores, their practical value is significantly diminished. Consequently, research into lightweight MLLMs has attracted widespread attention. Among these approaches, knowledge distillation, as one of the mainstream techniques, has demonstrated promising results. Its objective is to transfer knowledge from large-scale MLLMs to small-scale MLLMs Shu et al. (2025); Cai et al. (2025), enabling them to combine high performance with compactness.

A straightforward distillation approach involves adding a loss that aligns the ouput probability distribution on top of the autoregressive loss of MLLMs Hinton et al. (2015). While the end-response-based distillation paradigm has yielded promising results, the knowledge signals are concentrated in the tail, resulting in overly lax supervision of the intermediate layers. Recent studies have shifted the focus from end-point distillation to mid-layer distillation, achieving knowledge transfer precisely by interpretable analysis for working mechanism of MLLMs. For example, CompoDistill Kim et al. (2026) finds that visual signals in the mid-layers are crucial for compositional reasoning tasks, and Align-KD Feng et al. (2025) reveals the importance of features from the first layer.

In this paper, we provide a novel information flow-based distillation strategy. We draw inspiration from recent work on the interpretability of MLLMs Neo et al. (2025); Zhang et al. (2025b). The pattern of interaction between multimodal information across the various layers of MLLMs, termed *information flow*. In the early LLaVA Liu et al. (2023) models (based on LLama-2 Llama (2023)), the information flow exhibits the following pattern: at shallow layers, global image information flows towards the text; at intermediate layers, local image information flows towards the text; and at final layers, text information flows towards the last token (for generation). However, when using

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the more advanced large language model, LLama-3 Grattafiori et al. (2024), as the foundation, the phenomenon of global image information flowing towards the text at shallow layers is significantly weakened. We observe the same change in the Qwen series Bai et al. (2023) and find that the hierarchical processing of global and local image information is not pronounced, tending instead to be unified at intermediate layers. We hypothesise that, at shallow layers, information may flow within the same modality. We therefore conduct analytical experiments on intra-modal flow and discover that, in the shallow layers of MLLMs, information flows primarily within the text. In the intermediate layers, information flow occurs not only within the text and images but also across modalities. In the final layers, information flows from the text to the last token. Consequently, we identify a three-stage workflow for MLLMs: they first undertake a preliminary understanding of the text, then perform multimodal processing, and finally focus on generating the response.

Furthermore, we investigate how the information flow changes across MLLMs of different sizes. We find that the three-stage information flow remains stable across models of all sizes, and, surprisingly, the length of each stage is proportional to the total number of layers of the model. This suggests that within a given family of MLLMs, when the number of layers changes, each stage scales proportionally, rather than sacrificing or favoring any particular stage. In addition, we explore the impact of information flow on the performance of MLLMs. We design a simple causal intervention experiment to observe performance changes by introducing perturbations to the model's information flow. We use Gaussian noise and the information flow from larger-scale MLLMs as perturbation factors, respectively. After applying Gaussian noise, model performance decreases, whereas applying the information flow from larger-scale MLLMs results in a slight improvement in model performance. This suggests that the information flow is a key factor to model performance and that the flow pattern of larger-scale MLLMs serves as a valuable teacher signal worth learning from.

Based on the above discussion, we propose an intuitive distillation scheme, named **InfDistill**, to transfer information flow from large-scale MLLMs to small-scale MLLMs. Our method supervises the information flow at each layer of the student model by adjusting the attention weight matrices. Specifically, following the three stages of information flow, we align attention in key regions for layers at specific stages. In the shallow layers, only the attention within the text is used for alignment; in the intermediate layers, all attention is aligned; and in the final layers, only the attention of the answer tokens toward the text is aligned. Since the number of layers in the teacher and student models typically differs, we design a parameter-free teacher matching strategy. Considering that the information flow is scaled proportionally across layers in different models, we directly scale the teacher layer numbers uniformly to the same scale as the student layers. The new teacher layer indices are uniformly distributed along the student layer index axis. We search for teachers within a fixed radius centered on the student index. The attention from multiple teacher layers is averaged to establish a one-to-one correspondence with the student layers. Our method effectively improves the performance of student models and achieves optimal or competitive results on multiple benchmarks.

Our contributions are summarized as follows: (1) We reveal the evolution of information flow in the latest MLLMs and identify their three-stage workflow. (2) We find that information flow consistently exists across MLLMs of different sizes within the same family, and that the duration of each stage is proportional to the number of model layers. (3) We demonstrate that the information flow of the model is a key factor influencing performance and that the information flow in large-scale MLLMs provides valuable guidance for smaller-scale ones. (4) Based on these findings, we propose a simple distillation strategy which transfers the information flow to the student model as knowledge. Correspondingly, we propose a parameter-free teacher matching strategy. Extensive experiments validate the effectiveness of the proposed method.

2 An Exploration for Scalable Information Flow

2.1 Preliminary

Multimodal large language models typically accept both images and text as input, with the two directly concatenated and the image tokens prefixed. Let $\mathbb{I} = [\mathbb{V}, \mathbb{T}]$ denote the input, where $\mathbb{V} = [v_1, ..., v_{n_v}]$ represents the image tokens and $\mathbb{T} = [t_1, ..., t_{n_t}]$ represents the text tokens. The last text token is used to predict the next token, denoted as $\hat{t}$. A large language foundation model typically consists of an embedding network f_e, a classification network f_c, and multiple transformer blocks

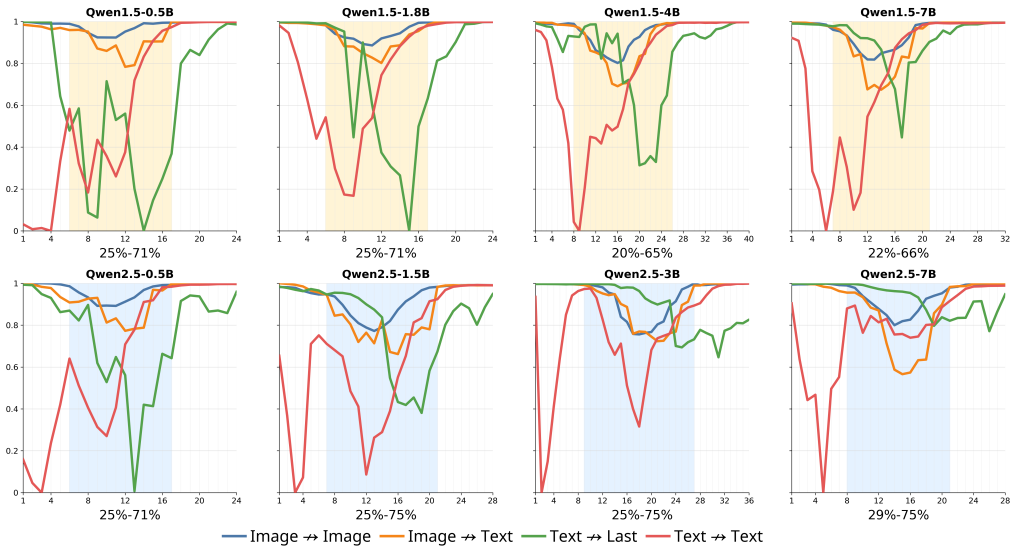


Figure 1: The normalized prediction probability changes on eight mainstream Qwen series MLLMs. The horizontal axis represents the model layer number, and the vertical axis represents the normalized prediction probability. $\nrightarrow$ means blocking attention. A significant decrease in prediction probability generally indicates a greater impact. The orange and blue backgrounds highlight the stages where all types of blockages have a significant impact.

$f_b = \{f_{mha}, f_{ffn}\} \times m$ where f_{mha}, f_{ffn} denote multi-head attention and feed-forward network, respectively. Thus, the prediction probability of the next token can be expressed as:

$$p(t_{n+1}|t_1, ..., t_n) = softmax(f_c(\Pi_{l=1}^{m} f_b^l(f_e(\mathbb{I})))). \tag{1}$$

It can be observed that information is passed between tokens via the multi-head attention module, which works by calculating the similarity between queries and keys to assign attention weights. Assuming the queries and keys in layer l are denoted as Q^l and K^l, respectively, the attention distribution can be expressed as:

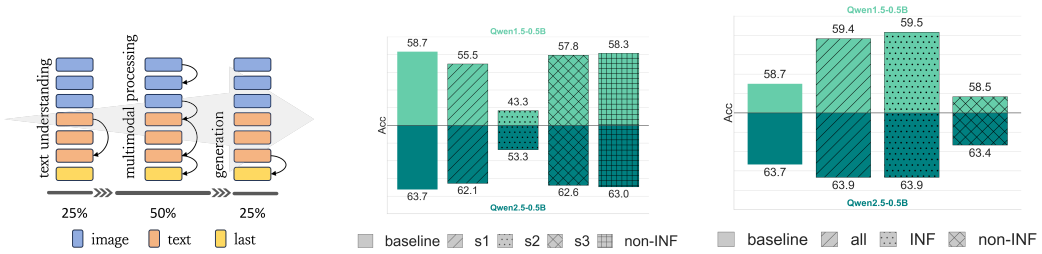
$$A^l = softmax(\frac{Q^l(K^l)^T}{\sqrt{d_k}} + M), \tag{2}$$

where d_k denotes the hidden dimension and M denotes causal mask. To investigate the impact of information flow on the model, we can disable specific attention regions to observe changes in output probabilities. For example, by blocking the $\mathbb{V} \nrightarrow \mathbb{T}$ attention weights in layer l and observing the downward trend in ground-truth output probabilities, one can determine whether the image to text information flow in that layer is critical.

2.2 Scalable Information Flow

In early LLama-2-based MLLMs, the $Image \nrightarrow Text$ information flow occurs preferentially in the shallow layers. However, this phenomenon is significantly weakened in more advanced LLama-3-based MLLMs Zhang et al. (2025b). We hypothesize that more advanced models tend to pass image information to text in the intermediate layers, while intra-modal information flow may occur in the shallow layers. To this end, we conduct systematic analytical experiments on the Qwen series models. As shown in Figure 1, we set up four types of blocking, $[Image \nrightarrow Image, Image \nrightarrow Text, Text \nrightarrow Last, Text \nrightarrow Text]$ (the $Image \nrightarrow Last$ type is removed because it has virtually no effect on the results), where $Last$ means the last token for generation.

We perform experiments on eight Qwen models. As shown in Figure 1, in the shallow layers of the model, only the $Text \nrightarrow Text$ type of blocking significantly affects the output probabilities. Moving into the intermediate layers, all types of blocking lead to a decline in output probabilities. In the final layers, only the $Text \nrightarrow Last$ type causes a noticeable decrease in output probabilities.



(a) (b) (c)

Figure 2: (a) The three-stage workflow of MLLMs. (b) Gaussian noise perturbations. Evaluation is performed on MME using Qwen1.5-0.5B (top) and Qwen2.5-0.5B (bottom). *s1, s2, s3* denote noise is applied in stage 1, 2, 3, respectively. *non-INF* denotes noise is applied to parts beyond the information flow. (c) Attention perturbations. We mix the attention weights of Qwen1.5-0.5B with Qwen1.5-4B (top), and the attention weights of Qwen2.5-0.5B with Qwen2.5-3B. *all* means all attention weights are applied. *INF, non-INF* refer to only the information flow being considered and only the parts beyond the information flow being considered, respectively.

Therefore, we conclude that information flows primarily within the text in the shallow layers of the model; subsequently, intense information exchange occurs in the intermediate layers, including both intra-modal and cross-modal flows between image and text; finally, in the final layers, information flows mainly from the text to the last token which used for generation.

In addition, we highlight the intermediate layers where information flows most intensely. We observe that, although the total number of layers varies across models, this phase is generally handled by the middle 50% of the model's layers. Furthermore, its starting and ending points are roughly at the 25% and 75% marks of the model's layers. This means that as the number of layers of the model changes, the information flow scales proportionally. The model does not sacrifice or favor any particular phase.

Based on the above analysis, we identify the three-stage workflow of MLLMs, as shown in Figure 2a. The first stage, accounting for approximately 25% of the model's layers, is responsible for understanding the text; the second stage, accounting for approximately 50% of the model's layers, is responsible for multimodal processing; and the third stage, accounting for the remaining 25%, focuses on generating responses.

2.3 The Effect of Information Flow on Performance

A natural question arises: How does the information flow affect the model's performance? To answer this question, we conduct a simple causal intervention experiment. Using Qwen1.5-0.5B and Qwen2.5-0.5B Bai et al. (2023) as baselines, we implement interventions to the attention weights related to information flow and observe the performance changes on MME Fu et al. (2026). Stage 1 involves the intra-text attention matrix, Stage 2 involves the full attention matrix, and Stage 3 involves the attention matrices for the text and the last token.

First, we apply mild Gaussian noise to the target (to prevent the model from collapsing). As shown in Figure 2b, we apply Gaussian noise to the attention weights of stages 1, 2, and 3, as well as to those not involved in the information flow. It can be observed that, compared to the baseline, perturbing the three stages of the information flow leads to a performance decline. In contrast, perturbing the non-information flow results in a significantly smaller decline. This suggests that the information flow has a greater impact on performance than the non-information flow.

Next, we mix the attention weights of large-scale MLLMs with small-scale MLLMs. As shown in Figure 2c, we directly inject the attention weights from Qwen1.5-4B into Qwen2.5-0.5B, while also injecting the attention weights from Qwen2.5-3B into Qwen2.5-0.5B. The mix scope is set to the entire attention matrix, the information flow, and the non-information flow, respectively. As can be observed in the figure, model performance improved simply through attention injection. Furthermore, the improvement achieved through information flow injection is comparable to or even greater than that achieved by injecting the entire attention matrix. We believe this is because

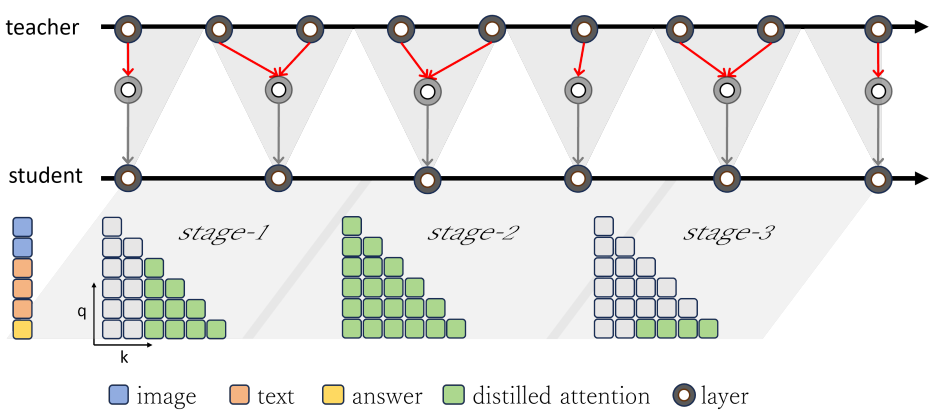


Figure 3: An overview of our method. First, we select teacher layers for each layer of the student model with an adaptive teacher matching strategy. Multiple teachers are averaged to ensure one-to-one distillation. Then, drawing on the three stages of the information flow, we apply different distillation strategies.

the entire attention matrix contains excessive noise. In contrast, non-information flow injection causes a performance decline. This suggests the potential of information flow injection to enhance the performance of student models.

3 METHODOLOGY: INFDISTILL

Based on the above discussion, we propose an intuitive distillation framework designed to transfer the information flow from large-scale MLLMs to small-scale MLLMs. More specifically, we encourage the student model to learn the attention distributions present in the information flow of the large model. Our method is called InfDistill and an overview is presented as Figure 3.

Adaptive Teacher Matching Strategy. Typically, the number of layers of the teacher model does not match the number of layers of the student model. The first problem we need to solve is matching teacher layers to each student layer. Previous studies typically employ sliding windows Kim et al. (2026); Lee et al. (2025) to cluster teacher layers, but this approach does not work for our method. The reason is that previous studies only needed to match teachers to a small number of layers of student model, whereas we need to match teachers to every single student layer. This means we cannot find a suitable window size and step size (e.g., Qwen1.5-4B with 40 layers vs. Qwen1.5-0.5B with 24 layers). Furthermore, designing the window and step size is a complex engineering task. To address this, we propose a simple parameter-free matching strategy. Since the information flow scales with the number of layers, we directly scale the teacher layers to the same scale as the student layers. Let the teacher layer indices be $a = [1, 2, ..., n]$, the student layer indices be $b = [1, 2, ..., m]$. Then the scaled teacher indices are denote as:

$$\hat{a}_i = 1 + (a_i - 1)\frac{m-1}{n-1}, \tag{3}$$

where a_i denotes the i-th number of teacher layer. For the j-th student layer, its teacher layers are:

$$a_t^j = \{a_i \in a | b_j - 0.5 \leq \hat{a}_i < b_j + 0.5\}. \tag{4}$$

Three-Stage Distillation Strategy. Based on the information flow mechanism, we propose a three-stage distillation strategy. In the first stage, the model focuses on understanding the text, therefore, the attention matrix within the text is used for distillation. We denote the student and teacher attentions as $A_s^l[\mathbb{T} : \mathbb{T}]$ and $A_t^l[\mathbb{T} : \mathbb{T}]$, respectively. For multiple teacher layers, we use average pooling to ensure one-to-one distillation, denoted as $\hat{A}_t^l[\mathbb{T} : \mathbb{T}] = \frac{1}{|a_t^j|}\sum A_t^l[\mathbb{T} : \mathbb{T}], l \in a_t^j$. In the second stage, all attention must be taken into account, therefore, full attention distillation is performed in this stage. The student and averaged teacher attentions are denoted as A_s^l and $\hat{A}_t^l$. In the third stage,

the model focuses on generating answers. Consequently, we only distill the attention matrix between the text and the answer. The student and averaged teacher attentions are denoted as $A_s^l[\mathbb{A} : \mathbb{T}]$ and $\hat{A}_t^l[\mathbb{A} : \mathbb{T}]$, where $\mathbb{A}$ denotes the answer set. The distillation loss is represented as:

$$\mathcal{L}_{inf} = 1 - \frac{1}{|s|}[\sum^{l \in s_1} sim(A_s^l[\mathbb{T} : \mathbb{T}], \hat{A}_t^l[\mathbb{T} : \mathbb{T}]) + \sum^{l \in s_2} sim(A_s^l, \hat{A}_t^l) + \sum^{l \in s_3} sim(A_s^l[\mathbb{A} : \mathbb{T}], \hat{A}_t^l[\mathbb{A} : \mathbb{T}])], \tag{5}$$

where s_1, s_2, s_3 denote layer sets for stage 1, 2, 3 and $s = s_1 + s_2 + s_3$ and $sim()$ denotes similarity function. In this paper, we employ the cosine similarity.

Training Scheme. We adopt a three-step training scheme and integrate two standard training losses. The first training loss is the autoregressive loss of the large language model, denoted as $\mathcal{L}_{base}$. Additionally, following previous studies Shu et al. (2025); Cai et al. (2025), we introduce a standard distillation loss, which uses the Kullback-Leibler (KL) divergence to align the probability distributions of the outputs, denoted as:

$$\mathcal{L}_{kd} = \mathbb{E}_{(x,y)} \left[\sum_{t=1}^{|y|} \mathrm{KL}(p_T(\cdot \mid x, y_{<t}) \,\|\, p_S(\cdot \mid x, y_{<t})) \right], \tag{6}$$

where x, y denote the input and output, respectively.

Pre-Training (PT). In the first step, we train the projector, which is responsible for projecting visual features into the language space. Both the visual encoder and the Large Language Model (LLM) are frozen. The training loss is $\mathcal{L}_{base}$.

Supervised Fine-Tuning (SFT). In the second step, we fine-tune the projector and the LLM to enable the LLM to answer visual questions. The training loss is $\mathcal{L}_{base}$.

Distilled Fine-Tuning (DFT). In the third step, we perform distillation-based fine-tuning on the projector and the LLM. The goal is to use the information flow from large-scale MLLMs to guide the generative patterns of the student model. The training loss is $\mathcal{L}_{base} + \mathcal{L}_{kd} + \mathcal{L}_{inf}$.

Discussion: It is worth noting that our goal is not applying the 25%-50%-25% three-stage information flow to all MLLMs, as that would be challenging. Our goal is to provide a feasible strategy for distillation which involves detecting and transferring information flow. We are the first to validate the scalability of information flow across MLLMs of different sizes and evaluated their impact on model performance. Subsequent distillation experiments have also confirmed that information flow transfer is an effective method.

4 Experiments

4.1 Setup

Implementation Details and Evaluation Benchmarks. To ensure a fair comparison, we follow the experimental setup of LLaVA-KD. Specifically, we use the pre-trained SigLIPB/14@384px Zhai et al. (2023) as the visual encoder and use base models from the Qwen1.5 and Qwen2.5 series as language models. The projector is a two-layer MLP with a GELU activation layer. We leverage LLaVA1.5-558k Liu et al. (2024b) to train the projector during the PT phase, and LLaVA-mix-665k to train the projector and LLM during the SFT and DFT phases. The learning rate is set to 1e-3 during the PT phase and to 2e-5 during the SFT and DFT phases. The image resolution is 384×384. We use the AdamW optimizer, with a cosine decay learning rate scheduler. The weight decay is 0.0 and the warm-up ratio is 0.03. The global batch size is 256 during the PT phase and 128 during the SFT and DFT phases. We conduct the experiments on 8 NVIDIA H200 GPUs. To comprehensively evaluate the effectiveness of our method, we select eight multimodal datasets for evaluation, including GQA Hudson & Manning (2019), ScienceQA (SciQA) Lu et al. (2022), TextVQA Singh et al. (2019), MME Fu et al. (2026), MMBench (MMB) Liu et al. (2024c), MMBench-CN (MMB^{CN}) Liu et al. (2024c), POPE Li et al. (2023), MMMU Yue et al. (2024).

Comparison Methods. Native models: Imp Shao et al. (2025), Bunny He et al. (2024), Mini-Gemini Li et al. (2025), MoE-LLaVA Lin et al. (2026), SPHINX Liu et al. (2024a),

Table 1: Benchmarked results with SoTA MLLMs.#Samples denotes training data samples. The gray background indicates distillation methods. The best and second-based results are marked in bold and underlined, respectively.

	Method	LLM	#Samples	GQA	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	AVG
~7B	LLaVA-1.5	Vicuna-7B	1.2 M	62.0	66.8	58.2	75.5	64.3	58.3	85.9	34.4	63.2
	InstructBLIP	Vicuna-7B	130 M	49.2	60.5	50.1	–	36.0	–	–	–	–
	Qwen-VL	Qwen-7B	1500 M	59.3	67.1	63.8	–	38.2	7.4	–	–	–
	Qwen-VL-Chat	Qwen-7B	1500 M	57.5	68.2	61.5	74.4	60.6	56.7	–	35.9	–
	mPLUG-Owl2	LLaMA2-7B	400 M	56.1	68.7	54.3	72.5	66.5	–	85.8	32.7	–
~4B	TinyLLaVA	Qwen1.5-4B	1.2 M	63.4	72.9	59.0	69.3	67.9	67.1	85.2	38.9	65.5
	TinyLLaVA	Qwen2.5-3B	1.2 M	63.2	76.0	61.5	73.9	71.8	69.5	86.4	40.3	67.8
	TinyLLaVA	Phi2-2.7B	1.2 M	62.0	69.1	59.1	73.2	66.9	–	86.4	38.4	–
	Bunny	Phi2-2.7B	2.6 M	62.5	70.9	56.7	74.4	68.6	37.2	–	38.2	–
	Imp-3B	Phi2-2.7B	1.5 M	63.5	72.8	59.8	–	72.9	46.7	–	–	–
	MobileVLM	MLLaMA-2.7B	1.2 M	59.0	61.0	47.5	64.4	59.6	–	84.9	–	–
	MoE-LLaVA	Phi2-2.7B	2.2 M	62.6	70.3	57.0	–	68.0	–	85.7	–	–
	MiniCPM-V	MiniCPM-2.4B	570 M	51.5	74.4	56.6	68.9	64.0	62.7	79.5	–	–
	LLaVADI	MLLaMA-2.7B	1.2 M	61.4	64.1	50.7	68.8	62.5	–	86.7	–	–
~2B	Imp-2B	Qwen1.5-1.8B	1.5 M	61.9	66.1	54.5	65.2	63.8	61.3	86.7	–	–
	Bunny-2B	Qwen1.5-1.8B	2.6 M	59.6	64.6	53.2	65.0	59.1	58.5	85.8	–	–
	Mini-Gemini-2B	Gemma-2B	2.7 M	60.7	63.1	56.2	67.0	59.8	51.3	85.6	31.7	59.4
	MoE-LLaVA-2B	Qwen1.5-1.8B	2.2 M	61.5	63.1	48.0	64.6	59.7	57.3	87.0	–	–
	TinyLLaVA	Qwen2.5-1.5B	1.2 M	62.0	72.0	57.4	**72.5**	68.6	63.0	85.5	**37.0**	64.8
	TinyLLaVA	Qwen1.5-1.8B	1.2 M	55.5	65.3	47.7	61.2	57.1	55.5	83.4	34.1	57.5
	LLaVA-MOD	Qwen1.5-1.8B	5 M	58.7	68.0	58.5	66.7	66.3	61.9	87.0	–	–
	LLaVA-KD	Qwen1.5-1.8B	1.2 M	62.3	64.7	53.4	69.1	64.0	63.7	86.3	33.6	62.1
	LLaVA-KD	Qwen2.5-1.5B	1.2 M	**62.5**	71.6	59.7	70.0	**71.0**	66.6	86.7	35.8	65.5
	InfDistill	Qwen1.5-1.8B	1.2 M	60.6	65.6	54.3	69.7	65.8	62.6	**88.0**	34.7	62.7
	InfDistill	Qwen2.5-1.5B	1.2 M	61.7	**72.4**	**59.7**	71.2	70.0	**67.5**	87.4	35.7	**65.7**
~1B	SPHINX-Tiny	TinyLLaMA-1.1B	15 M	58.0	21.5	**57.8**	63.1	52.3	56.6	82.2	–	–
	TinyLLaVA	Qwen1.5-0.5B	1.2 M	57.4	60.9	47.4	59.8	55.0	52.4	83.7	31.6	56.0
	TinyLLaVA	Qwen2.5-0.5B	1.2 M	58.3	59.1	49.2	61.5	58.9	54.2	86.1	**33.6**	57.6
	LLaVADI	MLLaMA-1.4B	1.2 M	55.4	56.0	45.3	58.9	55.0	–	84.7	–	–
	LLaVA-MOD	Qwen1.5-0.5B	5 M	56.2	**62.8**	53.9	**65.3**	58.8	50.4	–	–	–
	LLaVA-KD	Qwen1.5-0.5B	1.2 M	59.6	60.6	49.9	64.5	60.1	55.5	85.9	30.2	58.3
	LLaVA-KD	Qwen2.5-0.5B	1.2 M	**59.8**	60.6	52.0	64.7	**61.3**	**57.0**	86.4	28.3	58.8
	InfDistill	Qwen1.5-0.5B	1.2 M	58.4	59.2	49.7	64.5	59.0	54.6	**87.1**	32.1	58.1
	InfDistill	Qwen2.5-0.5B	1.2 M	59.6	61.5	51.2	64.9	60.6	56.2	86.9	32.1	**59.1**

TinyLLaVA Zhou et al. (2024), InstructBLIP Dai et al. (2023), mPLUG-Owl2 Ye et al. (2024), LLaVA1.5 Liu et al. (2024b), QwenVL Bai et al. (2023), MobileVLM Chu et al. (2023), MiniCPM-V Yao et al. (2024). Distilled models: LLaVAMoD Shu et al. (2025), LLaVA-KD Cai et al. (2025), LLaVADI Xu et al. (2024).

4.2 Comparison with SoTAs.

Table 1 reports the overall comparison between our method and sota MLLMs. For the 1B models, InfDistill with Qwen2.5-0.5B achieves 59.1 average score, the highest among all 1B models. This surpasses the TinyLLaVA baselines using Qwen1.5-0.5B and Qwen2.5-0.5B by 3.1 and 1.5 points, respectively. It also exceeds the distillation method LLaVA-KD (Qwen2.5-0.5B) by 0.3 on average socre. For the 2B models, InfDistill with Qwen2.5-1.5B reaches 65.7 average score, setting a new state of the art among 2B MLLMs. It outperforms TinyLLaVA (Qwen2.5-1.5B) by 0.9 points and LLaVA-KD (Qwen2.5-1.5B) by 0.2 points, and exceeds Mini-Gemini-2B by a substantial margin of 6.3 Avg. On individual benchmarks, it ranks first or tied-first in SciQA, TextVQA, and MMB^{CN}, while placing second in MME and MMB. With Qwen1.5-1.8B, InfDistill achieves 62.7 average score, beating TinyLLaVA (Qwen1.5-1.8B) and LLaVA-KD (Qwen1.5-1.8B) by 5.2 and 0.6 points, respectively. It is worth noting that LLaVA-KD requires two distillation stages (DFT-SFT-DFT), whereas our method requires only one distillation stage (PT-SFT-DFT). Our method offers higher efficiency and better performance.

4.3 Ablation Studies

Ablation experiments on stages and full attention. We conduct ablation experiments on staged and full attention distillation. As shown in Table 2, removing any single stage results in a significant drop in performance, confirming that these three stages are mutually reinforcing. The largest drop occurs when Stage-2 is removed, as it involves the most intense information flow among the three stages. Furthermore, knowledge distillation on the full attention matrix yields an average score of 65.2, which is 0.5 points lower than our scheme. This suggests that indiscriminately learning

Table 2: Ablation experiments on stages and full attention. w/o indicates removing one stage during distillation. *Full attention* indicates that distillation is performed on the entire attention matrix.

Method	GQA	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	Avg
PT-SFT	60.1	70.0	56.9	69.1	68.7	63.1	85.8	38.8	64.1
w/o Stage-1	61.3	71.9	59.5	70.2	70.7	67.4	86.0	35.1	65.3
w/o Stage-2	61.3	71.0	59.1	69.4	69.4	66.2	86.1	34.6	64.6
w/o Stage-3	61.4	71.8	59.2	72.0	71.3	66.1	86.5	35.1	65.4
Full attention	61.7	71.6	59.1	70.6	70.4	66.8	86.3	35.0	65.2
All stages	61.7	72.4	59.7	71.2	70.0	67.5	87.4	35.7	65.7

Table 3: Ablation studies on information flow loss $\mathcal{L}_{inf}$.

Method	$\mathcal{L}_{kd}$	$\mathcal{L}_{inf}$	GQA	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	Avg
PT-SFT	×	×	55.6	58.6	45.9	58.7	54.7	49.4	84.9	30.6	54.8
DPT-DFT	✓	×	56.4	59.2	48.2	58.1	55.0	47.4	86.7	32.7	55.5
DPT-DFT	×	✓	56.2	58.2	49.4	63.1	55.8	52.5	87.3	30.8	56.7
PT-SFT-DFT	✓	×	58.6	58.8	49.1	62.9	59.3	55.0	85.9	30.3	57.5
PT-SFT-DFT	×	✓	58.5	58.0	49.3	63.6	59.3	54.6	86.9	31.9	57.8
PT-SFT-DFT	✓	✓	58.4	59.2	49.7	64.5	59.0	54.6	87.1	32.1	58.1

the entire attention weights introduces redundant or noisy supervision, whereas selective attention distillation is more effective. Additionally, our approach is more memory- and time-efficient than full attention distillation.

Ablation studies on information flow loss $\mathcal{L}_{inf}$. Table 3 presents the ablation experiments for the proposed information flow loss $\mathcal{L}_{inf}$. It can be observed that, compared to the PT-SFT baseline, introducing any of the distillation losses yields a consistent improvement, with the best results achieved when using both $\mathcal{L}_{kd}$ and $\mathcal{L}_{inf}$ simultaneously. Notably, $\mathcal{L}_{inf}$ alone already demonstrates significant benefits. Specifically, under the DPT-DFT setting, using $\mathcal{L}_{inf}$ alone increases the Avg score from 55.5 to 56.7 compared to using $\mathcal{L}_{kd}$ alone, and improves the MME and MMB^{CN} scores by 5.0 and 5.1 points, respectively. Under the PT-SFT-DFT setup, the performance of using $\mathcal{L}_{inf}$ alone is also slightly higher than that of using $\mathcal{L}_{kd}$ alone. This demonstrates that information flow distillation effectively improves the performance of student models and can be effectively integrated with the end-response-based distillation paradigm.

Table 4: Detailed ablation experiments.

(a) Ablation studies on similarity function.

Method	Avg
KL Divergence	55.3
MSE	57.5
Cosine Similarity	58.1

(b) Ablation studies on matching strategies.

Method	Avg
radius=1.0	57.4
radius=0.5	58.1

(c) Experiments on teacher models with different sizes.

LLM of the Teacher	LLM of the Student	Avg
Qwen2.5-7B	Qwen2.5-0.5B	58.4
Qwen2.5-4B	Qwen2.5-0.5B	59.1
Qwen2.5-1.5B	Qwen2.5-0.5B	58.8
Qwen2.5-7B	Qwen2.5-1.5B	65.6
Qwen2.5-4B	Qwen2.5-1.5B	65.7

Detailed ablation studies. We conduct several ablation experiments on specific details. (1) We perform ablation experiments on the choice of similarity functions, and the results are shown in Table 4a. We compare the performance of KL divergence, MSE, and cosine similarity. We can see that cosine similarity achieved the best results, indicating that it is better suited for measuring the alignment between information flow distributions, whereas KL divergence is prone to optimization instability due to differences in distribution scale. (2) For the adaptive teacher matching strategy, we conduct ablation experiments on the matching radius in Equation 4. Specifically, we increase the radius from 0.5 to 1.0. This means that a single teacher layer may correspond to multiple student layers. Table 4b shows a slight decline in performance. This indicates that stricter local matching provides more precise supervision signals and avoids the noise introduced by overly broad matching. (3) We conduct ablation experiments on the teacher model. As shown in Table 4c, for Qwen2.5-0.5B, the 4B teacher achieves the best performance, outperforming both the 7B and 1.5B teachers. For

Table 5: Ablation studies of different training recipes.

Method	GQA	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	Avg
PT-SFT	55.6	58.6	45.9	58.7	54.7	49.4	84.9	30.6	54.8
DPT-DFT	57.3	58.0	49.0	63.8	57.1	53.7	86.3	31.3	57.1
DPT-SFT-DFT	58.5	59.0	49.1	61.9	57.8	53.5	86.8	30.6	57.2
DPT-DFT-SFT	56.6	58.1	48.4	61.6	57.1	54.2	86.0	31.8	56.7
PT-SFT-DFT	58.4	59.2	49.7	64.5	59.0	54.6	87.1	32.1	58.1

Qwen2.5-1.5B, the 4B teacher also performs slightly better than the 7B teacher. This indicates that a larger teacher is not necessarily better. A moderate-sized 4B teacher strikes a better balance between knowledge capacity and distillation difficulty, resulting in a more effective match with the student model in terms of information flow distribution.

Ablation studies of different training recipes.. Table 5 compares the effects of different training recipes. Using the standard PT-SFT as the baseline, all recipes show improvements after distillation. It is worth noting that compared to DPT-SFT-DFT, placing DFT before SFT (DPT-DFT-SFT) results in an Avg of only 56.7, indicating that the subsequent SFT may mask some of the distillation signals, thereby weakening the final performance. In contrast, the PT-SFT-DFT recipe achieves the best results. This demonstrates that establishing a solid multimodal foundation through PT-SFT first, followed by information flow distillation in the final stage, can further refine cross-modal alignment without compromising the knowledge gained from pretraining and supervised fine-tuning.

5 Related Works

Interpretability for MLLMs. Recent years have seen a growing body of research on the interpretability of MLLMs, which facilitate the application of MLLMs. LVLM-Interpret Ben Melech Stan et al. (2024) developes a toolkit for causal interpretation to investigate the decision-making processes of MLLMs. Furthermore, researchers Neo et al. (2025); Zhang et al. (2025a) have explored how LLMs process multimodal data, revealing the mechanism that early-stage image processing and late-stage prediction. With regard to model safety, researches on the logit distribution of output tokens Zhao et al. (2024) and safe activation layers Xu et al. (2025) have yielded promising results. More specifically, some studies have focused on unraveling the mechanisms of multimodal information storage Basu et al. (2024), interaction Neo et al. (2025), and flow Zhang et al. (2025b) within each layer. We have further refined the layer-based interpretability framework of MLLMs by investigating the intra-modal flow mechanisms and scalability of information flow.

Knowledge Distillation for MLLMs. Aligning the probability distributions of outputs is a widely used distillation method Hinton et al. (2015). LLaVADI Xu et al. (2024) explores the key factors influencing distillation. LLaVA-MOD Shu et al. (2025) introduces MoE distillation to train experts. LLaVA Cai et al. (2025) proposes distillation of image outputs and the relationships between them. VLsI Lee et al. (2025) introduces the intermediate verbalizers to perform distillation on selected layers. Align-KD Feng et al. (2025) finds that cross-modal attention matching in the first layer is crucial. CompoDistill Kim et al. (2026) finds that image attention is key to influencing compositional reasoning tasks. Unlike previous work, we propose a new strategy that realizes a layer-to-layer distillation paradigm by revealing information flow.

6 Conclusion

In this work, we are committed to validating the feasibility of a novel distillation strategy. Inspired by information flow and refined by us across MLLMs of various scales, we identify a three-stage working mechanism for MLLMs and demonstrate the scalability of information flow. We then uncover the potential of information flow to enhance model performance through causal interventions and implement it in the proposed distillation method. Extensive experiments validate the effectiveness of the proposed method. However, the generalizability of information flow to MLLMs beyond the Qwen series has not been fully explored. Additionally, further research is needed for cases involving more complex modalities, such as video and audio.

AI USE STATEMENT

In this work, we used generative AI tools for generating portions of code. We have not used generative AI tools for generating research ideas, designing or executing experiments, analyzing data, organizing or framing the paper, or drafting the manuscript, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for translating and polishing some sentences in the manuscript to improve readability and professionalism. We have reviewed all AI-assisted work. Specifically, the AI-generated code was checked and verified for correctness by the authors, and all AI-assisted translations and edits were reviewed and revised by the authors to ensure that they accurately convey our intended meaning. The core research process including the conception of ideas, execution of experiments, organization and framing of the paper, and most of the writing was carried out by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:7400–7426, 2024.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8182–8187, June 2024.

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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 239–249. IEEE, 2025.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4178–4188. IEEE, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. *arXiv preprint arXiv:1503.02531*, 2015.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Jiwan Kim, Kibum Kim, Sangwoo Seo, and Chanyoung Park. Compodistill: Attention distillation for compositional reasoning in multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 71814–71844, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29545–29557. IEEE, 2025.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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*, pp. 292–305, 2023.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024b.

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*, pp. 216–233. Springer, 2024c.

Hugo Touvron Llama. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2407.21783*, 2023.

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

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*, 27:2961–2974, 2025.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Weilong Dai, Zhelun Yu, Wanggui He, et al. Llava-mod: Making llava tiny via moe-knowledge distillation. In *International Conference on Learning Representations*, volume 2025, pp. 9386–9404, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. Cross-modal safety mechanism transfer in large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 58315–58336, 2025.

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.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. Mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13040–13051. IEEE, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Information flow in LVLMs across reasoning tasks. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2289–2299, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.115. URL https://aclanthology.org/2025.naacl-long.115/.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791. IEEE, 2025b.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pp. 127–142. Springer, 2024.

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

A Appendix