From Attenuation to Attention: Variational Information Flow Manipulation for Fine-Grained Visual Perception

Jilong Zhu[1,2,3], **Yang Feng**[1,2,3]*
[1] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] State Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences (ICT/CAS)
[3] University of Chinese Academy of Sciences, Beijing, China
zhujilong22s@ict.ac.cn, fengyang@ict.ac.cn

Abstract

While Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities in general visual understanding, they frequently falter in fine-grained perception tasks that require identifying tiny objects or discerning subtle visual relationships. We attribute this limitation to Visual Attenuation: a phenomenon where sparse fine-grained visual signals are prematurely suppressed or diluted by dominant textual tokens during network propagation, resulting in a "loss of focus" during the deep-level decision-making process. Existing input-centric solutions fail to fundamentally reverse this intrinsic mechanism of information loss. To address this challenge, we propose the Variational Information Flow (VIF) framework. Adopting a probabilistic perspective, VIF leverages a Conditional Variational Autoencoder (CVAE) to model the visual saliency relevant to the question-answer pair as a latent distribution. As a plug-and-play module, VIF can be integrated into existing architectures. Extensive evaluations across diverse benchmarks—covering General VQA, fine-grained perception, and visual grounding—demonstrate that VIF yields competitive improvements over previous methods, validating its effectiveness in enhancing the fine-grained perception of MLLMs. Codes are available at https://github.com/ictnlp/VIF.

1 Introduction

In recent years, Multi-modal Large Language Models (MLLMs), represented by the LLaVA (Liu et al., 2023b; Huang et al., 2025), InternVL (Chen et al., 2024; Wang et al., 2025a), and QwenVL (Bai et al., 2023; Wang et al., 2024; Bai et al., 2025b,a) series, have demonstrated remarkable proficiency in general visual understanding and reasoning. However, as the focus shifts from macroscopic scene description to **Fine-Grained Visual Perception**, a critical bottleneck becomes apparent. Current MLLMs often underperform when tasked with discerning minute objects or interpreting subtle visual relationships, limiting their broader applicability.

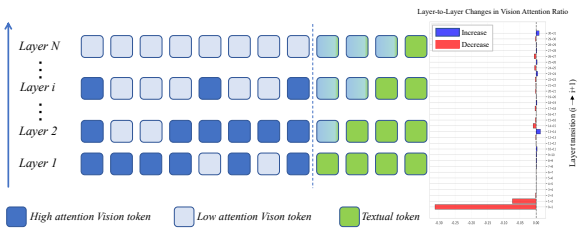

Figure 1: **Illustration of Visual Attenuation. (Left)** Schematic view showing visual tokens (blue) fading as they propagate through deep layers, while textual tokens (green) dominate. **(Right)** Quantitative layer-to-layer changes in vision attention ratio, averaged over 500 randomly sampled instances. The sharp drop in early layers indicates a premature loss of visual details.

This performance degradation stems from the information flow mechanism of MLLMs according to our analysis experiments (details will be shown in Section 2). In the current information flow mechanism of MLLMs, visual and textual inputs are fed into the model independently. After undergoing initial processing in the shallow layers, semantic representations for both modalities are learned at the middle layers. However, as information ascends to the deep layers, textual representations continuously absorb visual information, leading to a gradual attenuation of visual signals. This structural bias ultimately causes the model to prioritize textual representations over visual ones during the downstream textual response generation. Notably, textual representations typically integrate only those visual cues that are relevant to textual instructions, whereas the visual information critical for response generation is often not explicitly reflected in such instructions. Further experimental investigations reveal that, during the textual response generation phase, visual representations not only receive a marginal proportion of

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18586–18597

attention weights but also tend to exhibit a uniform distribution, ultimately resulting in the defocusing of visual information.

Building on the aforementioned findings, enabling fine-grained image understanding requires assigning greater attention to response-relevant visual regions during the generation process. Furthermore, attention over visual representations should be selectively prioritized rather than uniformly distributed. However, these objectives are challenged given the inherent architectural constraints of existing MLLMs. First, MLLMs lack access to response information during inference, making it difficult to identify visual information relevant to the target response. Second, the current formulation of attention distribution lacks explicit constraints, which constitutes a critical bottleneck in guiding the model to focus on specialized visual cues.

Existing research(e.g., Dense Connector (Yao et al., 2024), MMFuser (Cao et al., 2024)) on fine-grained visual perception primarily focuses on enhancing the fidelity of visual inputs. While these methods effectively augment the amount of visual information at the input stage, they still suffer from visual attenuation during cross-modal transmission where visual representations receive insufficient attention, and textual representations fail to capture the fine-grained visual details essential for generating accurate responses.

On these grounds, we introduce a novel **Variational Information Flow (VIF)** framework, designed to regulate the flow of visual information for seamless adaptation to both fine-grained and macroscopic visual tasks. To address the challenge of identifying response-correlated visual information that cannot be directly inferred from textual instructions, we employ variational inference to guide the model toward essential visual cues. In the training phase, we supply the ground-truth response, which facilitates the derivation of two heterogeneous attention distributions over visual representations: a posterior attention distribution, which takes both textual instructions and responses as query, and a prior attention distribution conditioned solely on textual instructions. By aligning these two distributions, VIF equips the MLLM with the capability to prioritize response-relevant visual attention during inference, even when only textual instructions are provided as queries.

To further enhance the model's focus on critical visual cues, we formalize the attention distribution using a Gaussian Mixture Model (GMM), where each Gaussian component in this model serves as a spotlight, re-activating the key visual information that might otherwise be neglected in the response generation process. Extensive experiments demonstrate that VIF, as a plug-and-play module, simultaneously supports fine-grained perception and broader macroscopic tasks, consistently outperforming existing input feature augmentation methods across both general and fine-grained benchmarks, validating the efficacy of our probabilistic information reconstruction paradigm.

Our main contributions are as follows:

i. We identify Visual Attenuation in the deep layers of MLLMs, revealing both a quantitative decline in attention weights and a structural collapse in spatial focus, highlighting the limitations of methods that only enhance input features.

ii. We propose the VIF framework, which employs CVAE-based posterior learning to dynamically reconstruct visual attention and capture task-specific visual saliency through robust probabilistic modeling.

2 Preliminary Experiments

To further investigate the Visual Attenuation, we conducted an in-depth statistical analysis of the internal attention dynamics of MLLMs based on 500 randomly sampled instances.

2.1 Vision Attention Ratio Analysis

As shown in the right panel of Figure 1 and the red curve in Figure 2, we observe a pervasive phenomenon of Visual Attenuation: although visual tokens initially command high attention weights in shallow layers, their influence suffers a precipitous decline in deep layers. Critical visual cues are overwhelmed by dominant textual tokens and prematurely discarded, hindering the model's ability to discern subtle details.

2.2 Vision Attention Distribution Analysis

Further spatial analysis reveals that, beyond the mean attention ratio decline, the internal allocation structure of visual attention also collapses. As illustrated by the violin plots in Figure 2, the long-tailed distribution shrinks in deep layers, with attention weights converging to homogenized low values that approximate a uniform spatial distribution. This indicates a reduced capacity for selectively activating key visual tokens.

This observation is further spatially corroborated by the visualizations in Figure 3: while the model

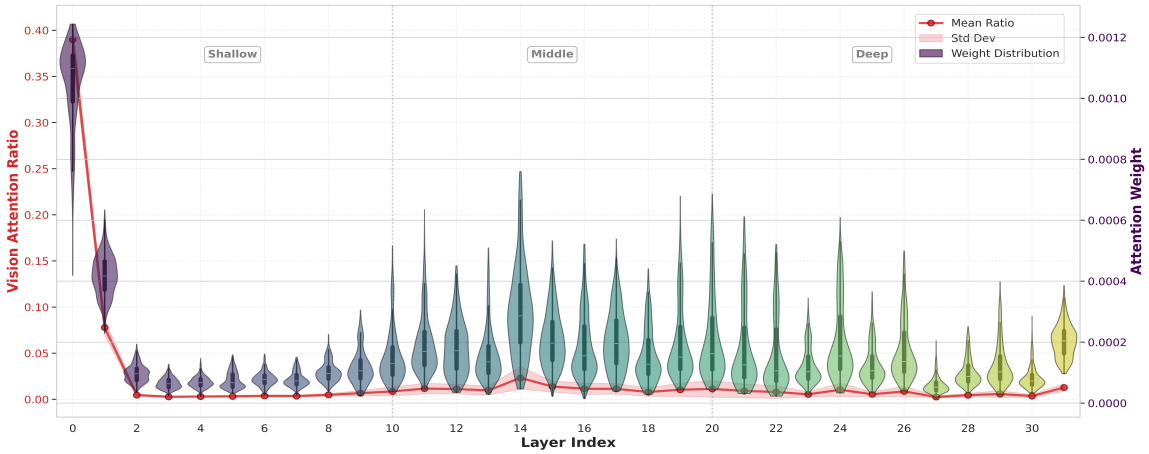

Figure 2: **Layer-wise visual attention distribution.** The red curve denotes the mean vision attention ratio across layers, which decreases sharply with depth. The violin plots illustrate the distribution of visual attention weights at each layer, with the long-tailed distribution shrinking in the deep layers.

Figure 3: **Visualization of layer-wise visual attention. (Left)** Shallow layers (L0-L1) capture dense contextual information with broad coverage. **(Middle)** Middle layers (L15-L16) successfully converge on key semantic regions (the board), exhibiting sparse and focused attention. **(Right)** In deep layers (L30-L31), this focus deteriorates into a diffuse and disordered state.

successfully converges on key semantic regions (e.g., the screen) in the middle layers (L15-L16), this focused state rapidly deteriorates into a diffuse and disordered distribution in the deep layers (L30-L31). Thus, beyond the mere neglect of visual signals in deep layers, Visual Attenuation also manifests as a spatially structural collapse.

This dual degradation in preference magnitude and spatial structure indicates that MLLMs not only **"see less"** but also **"see inaccurately"**, rendering them unable to sustain fixation on critical visual cues. Collectively, these factors precipitate the loss of fine-grained perception capabilities, thereby motivating our proposed solution.

3 Related Work

3.1 Multimodal Large Language Models

With the success of large language models (LLMs), extensive efforts have focused on building multimodal large language models (MLLMs) for unified vision-language perception. Most MLLMs adopt a modular architecture, connecting a pre-trained visual encoder (e.g., CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023)) to the LLM via an intermediate projection module. Early models such as BLIP-2 (Li et al., 2023) employ Q-Former based samplers, while later works including Flamingo (Alayrac et al., 2022) integrate visual information through cross-attention within LLM layers. LLaVA (Liu et al., 2023b) and its variants (Zhang et al., 2025b; Huang et al., 2025) popularize a simpler design using lightweight MLPs for visual-text alignment. Despite these advances, existing MLLMs still struggle with fine-grained visual scenarios.

3.2 Visual Input Optimization in MLLMs

To alleviate fine-grained visual information loss caused by limited input resolution, many works enhance visual representations via higher-resolution inputs and multi-scale encoding. Early MLLMs process images at fixed low resolutions, constraining fine-grained perception, while later methods increase resolution (e.g., Qwen2.5-VL (Bai et al., 2025b), MMFuser (Cao et al., 2024)) or introduce high-resolution encoders (Guo et al., 2024). Another line of work adopts dynamic cropping or patch-based strategies, such as HiRes-LLaVA (Huang et al., 2025), and ViCrop (Zhang et al., 2025a), to preserve local details through coarse-to-fine re-encoding. While effective at in-

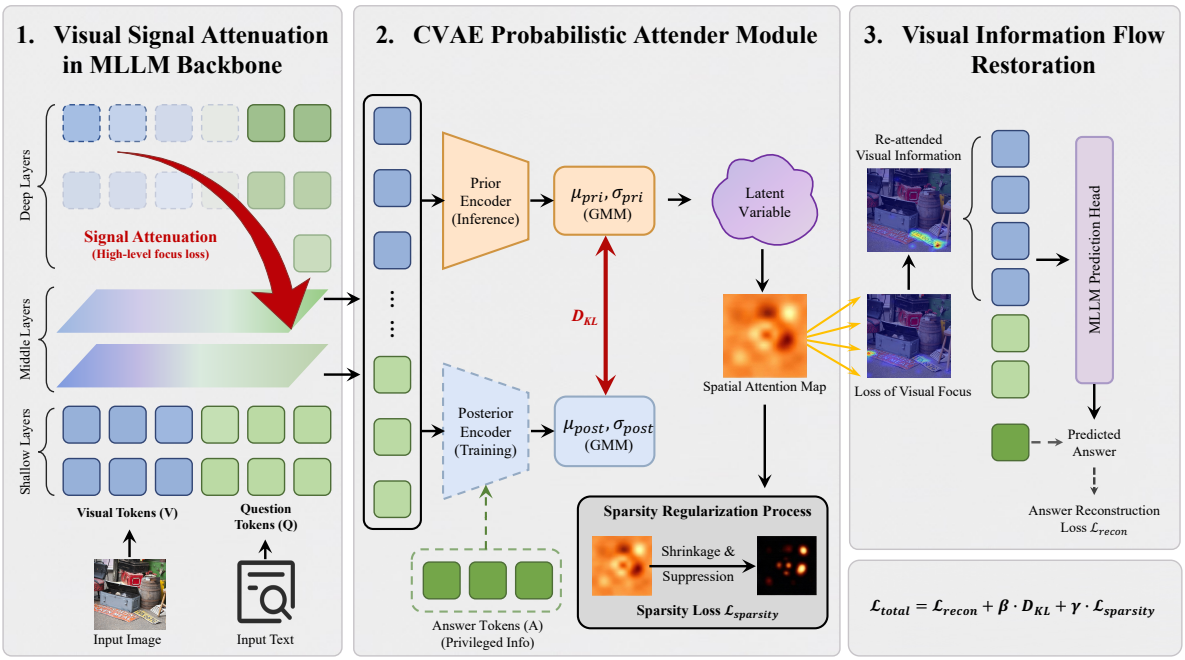

Figure 4: **Overview of the Variational Information Flow (VIF) Framework.** The framework consists of three stages: **(1) Visual Signal Attenuation Analysis.** The visual signal is out of focus in the deep layer. The model recovers rich visual cues from intermediate layers. **(2) CVAE Probabilistic Attender Module.** This module utilizes a GMM-based prior and posterior learning to reconstruct a sparse, task-relevant Spatial Attention Map from middle layers. **(3) Visual Information Flow Restoration.** This stage recovery involves injecting the learned visual focus into deeper layers to restore fine-grained visual cues to enhance the final prediction.

creasing pixel density, these approaches primarily enhance visual inputs and often incur substantial computational overhead, without explicitly addressing how visual information is preserved and exploited during downstream reasoning.

3.3 Instruction-Aware Visual Guidance

Beyond input-level improvements, recent works explore guiding visual modeling with textual semantics. Instruction-aware methods argue that different tasks require features from different semantic levels. IGVA (Li et al., 2026) re-weights multi-layer visual features conditioned on instructions, while TG-LLaVA (Yan et al., 2025) employs text-guided modules to suppress irrelevant regions and enhance fine-grained perception.

However, these methods typically rely on deterministic mappings conditioned solely on image and question. For complex fine-grained tasks or weakly specified queries, such guidance can be ambiguous, as multiple regions may satisfy the same instruction. In contrast, our approach models task-relevant visual importance as a probabilistic latent representation and leverages answer supervision during training to resolve ambiguity, enabling more robust visual guidance for multimodal reasoning.

4 Method

4.1 Problem Definition and Overview

In MLLMs, visual features often suffer severe attenuation through deep layers, as visual tokens (blue) gradually lose influence and are overshadowed by text tokens (green) with higher semantic density (Fig. 4, left). This text-dominated model bias undermines reasoning in the deeper layers, causing the model to lose not only a holistic understanding of the visual scene but also, more critically, its ability to capture fine-grained visual cues.

To mitigate this bottleneck, we propose the **Variational Information Flow** manipulation framework. The core idea is to model task-relevant visual saliency as a latent probability distribution and to introduce a plug-and-play CVAE-based probabilistic attention module. This module is designed to assist the backbone model in reconstructing the visual attention distribution within deep layers, reinjecting fine-grained cues into the decision flow to achieve a refocusing on critical visual cues.

4.2 Probabilistic Modeling via CVAE

Our goal is to learn a question-conditioned prior $p_\psi(z \mid V, Q)$ that approximates an answer-aware posterior $q_\phi(z \mid V, Q, A)$. After decoding, the la-

Figure 5: **Examples of Vision-Language Task Samples.** As illustrated, relying solely on questions often results in semantic ambiguity. Integrating answer information is thus critical to serve as a semantic anchor for robust task-driven visual modeling.

tent variable z characterizes the distribution of visual regions that are critical to answering the question under the context of image V, question Q, and answer A.

Motivation for Answer-Aware Modeling. As shown in Figure 5, relying only on the question Q can lead to semantic ambiguity, because the question alone may not uniquely specify the relevant visual evidence. To mitigate this issue, we introduce the ground-truth answer A during training as an additional semantic anchor that provides reliable supervision for task-relevant visual focus.

We model such uncertain visual importance with a latent variable z. Since fine-grained VQA often requires multi-hop reasoning over multiple sparse regions, the sampled latent variables are later decoded into a **spatial Gaussian mixture model (GMM)** with $K = 16$ components, allowing the model to capture multiple candidate visual foci within a unified probabilistic framework.

Posterior Learning with Privileged Information (Training Phase). During training, the ground-truth answer A is treated as privileged information. We construct the posterior distribution $q_\phi(z \mid V, Q, A)$ to capture the visual cues required to generate the answer A. This posterior acts as a teacher that encodes the intrinsic visual dependencies of the task.

Prior Inference for Reasoning (Inference Phase). During inference, as A is unavailable, we therefore learn a prior distribution $p_\psi(z \mid V, Q)$ to approximate the posterior. This compels the model to actively infer potential visual foci based solely on the image and question using its inherent reasoning capabilities, thereby achieving a transition from passive perception to active search.

4.3 Implementation of the Probabilistic Attender Module

The probabilistic attender consists of three components: latent encoding, spatial GMM decoding, and visual information flow restoration. Its central idea is to transform attenuated visual signals in deep layers into explicit focal cues, thereby turning attenuation into attention. In our formulation, the *information flow* ($\mathcal{IF}$) is explicitly defined as the attention probability distribution of an attention layer, rather than its hidden states.

Encoding Latent Variables via CVAE. At an middle layer l, we extract visual tokens V and question tokens Q. During training, answer tokens A are additionally available as privileged information. We build a prior branch and a posterior branch with the same architecture, each consisting of multi-head attention (MHA) and feed-forward networks (FFNs), but with separate parameters.

In each branch, visual and textual tokens first interact via bidirectional cross-attention, and the resulting features are further integrated by an additional MHA layer. The prior branch takes (V, Q) as input, whereas the posterior branch uses the full sequence $[Q, A]$ as textual input to encode answer-aware semantics. The fused representations are then fed into separate linear heads to predict the Gaussian parameters (μ_p, σ_p^2) and (μ_q, σ_q^2), respectively. The latent variables are sampled using the reparameterization trick:

$$\begin{aligned} z_p &= \mu_p + \sigma_p \odot \epsilon_p, \\ z_q &= \mu_q + \sigma_q \odot \epsilon_q, \\ \epsilon_p, \epsilon_q &\sim \mathcal{N}(0, \mathbf{I}). \end{aligned} \tag{1}$$

To encourage the prior distribution available at inference time to approximate the answer-aware posterior, we minimize the KL divergence between the two latent distributions:

$$\mathcal{L}_{KL} = \mathbb{E}\left[D_{KL}\left(\mathcal{N}(\mu_q, \sigma_q^2) \,\|\, \mathcal{N}(\mu_p, \sigma_p^2)\right)\right]. \tag{2}$$

Decoding to Spatial GMM and Information Flow Injection. To capture fine-grained visual focus, we decode the latent set $\{z_k\}_{k=1}^K$ into a Spatial GMM. Each component predicts a spatial center μ_k^{spa}, a spread σ_k^{spa}, and a mixture weight π_k. On the visual token grid $\{u_n\}_{n=1}^N$, the Gaussian response of the k-th component is rendered as

$$g_{k,n} = \exp\left(-\frac{\|u_n - \mu_k^{spa}\|_2^2}{2(\sigma_k^{spa})^2}\right). \tag{3}$$

Table 1: **Comparison on General Multimodal Capabilities. Bold** denotes the best result, and underline denotes the second best.

Model	LLM	Res.	Multimodal Understanding			Knowledge				OCR	Multi-Image
			SEED-I	LLaVAB	MME	GQA	VQAv2	AI2D	SQA	TextVQA	BLINK
LLaVA-v1.5	Vicuna-7B	336×336	67.2	61.8	1480.6	61.9	78.5	55.5	67.1	58.1	39.7
InstructBLIP	Vicuna-7B	224×224	58.8	59.8	1137.1	49.2	-	40.6	60.5	50.1	-
mPLUG-Owl2	LLaMA 2-7B	448×448	64.5	59.9	1450.2	56.1	-	55.7	68.7	54.3	-
MiniGPT-v2	LLaMA 2-7B	448×448	31.6	45.1	770.6	60.1	-	28.4	39.6	-	-
LLaMA-Adapter-v2	LLaMA 2-7B	-	32.7	-	972.7	-	-	-	-	-	-
IDEFICS-9B	LLaMA 2-7B	-	45.0	45.0	942.0	38.4	50.9	42.2	53.5	25.9	38.3
Mantis-8B-Fuyu	Fuyu-8B	1024×1024	59.3	46.8	1057.7	-	-	46.8	56.8	49.0	38.2
Qwen-VL	Qwen-7B	448×448	56.3	12.9	334.1	59.3	78.8	57.7	67.1	63.8	27.9
Qwen-VL-Chat	Qwen-7B	448×448	65.4	**67.7**	1487.6	57.5	78.2	63.0	68.2	61.5	28.2
LLaVA-NeXT	Mistral-7B	672×672	69.6	-	1519.3	**64.2**	-	**66.6**	68.5	64.9	-
LLaVA-1.5-HD	Vicuna-7B	672×1024	-	-	1414.0	54.1	-	63.8	79.3	64.0	-
Dragonfly	Vicuna-7B	2016×2016	-	-	1438.9	55.7	-	64.2	**79.7**	**66.5**	-
LLaVA-HR	Vicuna-7B	384×384	-	-	1522.3	-	**80.5**	-	59.6	-	-
DenseConnector	Vicuna-7B	-	-	67.4	-	63.8	-	-	69.5	59.2	-
MMFuser	Vicuna-7B	-	60.8	65.5	1479.7	62.8	-	-	68.7	58.8	-
LLaVA with ViCrop	Vicuna-7B	336×336	-	-	-	60.5	75.9	-	-	51.7	-
IGVA	Vicuna-7B	336×336	68.3	-	1519.8	63.1	-	57.0	70.2	59.4	-
TG-LLaVA	Vicuna-7B	336×336	65.0	-	-	63.4	-	-	-	-	-
VIF (Ours)	Vicuna-7B	336×336	**70.8**	66.4	**1547.2**	62.8	79.7	64.0	73.5	59.9	**40.5**

We then aggregate all components into a visual importance map and normalize it into a probability distribution:

$$V_{Map_n} = \sum_{k=1}^{K} \pi_k g_{k,n}, \hat{V} = \mathrm{Softmax}(V_{Map}). \quad (4)$$

Motivated by the analysis in introduction, where middle layers preserve stronger localization ability while deep layers often suffer from visual defocusing, we adopt a *selective layer patching* strategy. Specifically, we perform pair-wise injection from middle layers $l \in \{11, 13, 15, 17\}$ to deep layers $l' \in \{25, 27, 29, 31\}$. For each injection layer l', the original information flow is defined as the standard attention probability distribution:

$$\mathcal{IF}_{ori}^{(l')} = \mathrm{Softmax}\left(\frac{Q^{(l')}(K^{(l')})^\top}{\sqrt{d_h}} + M\right), \quad (5)$$

where M is the visibility mask. We then inject the decoded visual importance as an explicit focal bias:

$$\widetilde{\mathcal{IF}}_{inj}^{(l')} = \mathcal{IF}_{ori}^{(l')} + \alpha \cdot \hat{V}, \quad (6)$$

where α controls the injection strength.

To preserve the attention constraints on invisible regions, we apply masked re-normalization to obtain the final injected information flow:

$$\mathcal{IF}_{inj}^{(l')} = \mathrm{Norm}\Big(\widetilde{\mathcal{IF}}_{inj}^{(l')} \odot \mathbf{1}_{visible}\Big), \quad (7)$$

where Norm denotes row-wise normalization such that each attention distribution sums to 1. In the selected deep layer, $\mathcal{IF}_{inj}^{(l')}$ is used in place of the original attention probability matrix to compute the subsequent attention output, while all other operations remain unchanged.

4.4 Joint Optimization Objective

Our training objective follows the standard variational inference paradigm, aiming to maximize the Evidence Lower Bound (ELBO) of the conditional marginal likelihood $\log p_\theta(A|V, Q)$.

Derivation of the Variational Lower Bound. By introducing a variational posterior $q_\phi(z|V, Q, A)$ to approximate the true posterior, and applying Jensen's inequality, the ELBO is derived as follows:

$$\begin{aligned}\log p_\theta(A \mid V, Q) &\geq \mathbb{E}_{z\sim q_\phi}\left[\log \frac{p_\theta(A \mid V, Q, z)\, p_\psi(z \mid V, Q)}{q_\phi(z \mid V, Q, A)}\right] \\ &= \mathbb{E}_{z\sim q_\phi}[\log p_\theta(A \mid V, Q, z)] \\ &\quad - D_{\mathrm{KL}}(q_\phi(z \mid V, Q, A) \,\|\, p_\psi(z \mid V, Q))\end{aligned} \quad (8)$$

where the first term represents the response reconstruction objective aiming to maximize the likelihood of the ground-truth answer, optimized by the MLLM backbone parameters θ, while the second term enforces consistency between the variational posterior (parameterized by ϕ) and the conditional prior (parameterized by ψ).

Sparsity Regularization. To prevent the model from converging to trivial solutions that uniformly cover the entire image, we impose sparsity constraints on the Gaussian renderer. Specifically, we constrain the "volume" of each Gaussian compo-

Table 2: **Fine-Grained Perception and Referring Expression Comprehension. Bold** denotes the best result, and underline denotes the second best. REC results are reported with the CIDEr score.

Model	HR-Bench		Vstar	REC (RefCOCO)		
	4K	8K		val	testA	testB
LLaVA-v1.5	36.1	32.1	45.0	30.4	16.0	42.0
mPLUG-Owl2	36.9	33.8	35.6	-	-	-
MiniGPT-v2	25.5	26.1	-	-	-	-
IDEFICS-9B	30.6	28.8	-	-	-	-
Qwen-VL	31.8	28.4	-	-	-	-
IGVA	40.0	-	48.2	-	-	-
MMFuser	-	-	-	33.6	**17.7**	45.9
LLaVA with ViCrop	-	-	46.1	-	-	-
VIF (Ours)	**44.8**	**36.8**	**50.8**	**34.3**	16.6	**47.0**
Qwen2.5-VL-7B	68.6	64.9	76.4	-	-	-
CoVT	71.0	68.6	77.5	-	-	-
VIF w CoVT (Ours)	**71.5**	**68.8**	**79.0**	-	-	-

nent, defined as the product of its amplitude and variance, thereby encouraging the model to emphasize only the most critical regions. In addition, we incorporate an entropy regularizer to form a comprehensive sparsity objective, defined as:

$$\mathcal{L}_{\text{sparsity}} = \mathcal{H}(\boldsymbol{\pi}) + \frac{1}{K}\sum_{k=1}^{K}(\pi_k \times \boldsymbol{\sigma}_k^2), \quad (9)$$

where $\mathcal{H}(\boldsymbol{\pi}) = -\sum_{k=1}^{K}\pi_k \log \pi_k$ denotes the Shannon entropy, promoting sparse mixture distribution. Combined with this entropy regularization, the volume term simultaneously suppresses the variance $\boldsymbol{\sigma}_k$ (shrinking the spatial extent) and the amplitude π_k (dampening uncertain activations).

Total Loss Function. The final joint optimization objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{KL} + \gamma\mathcal{L}_{\text{sparse}} \quad (10)$$

where $\mathcal{L}_{\text{recon}} = -\log p_\theta(A|V, Q, z)$ denotes the standard cross-entropy loss for answer generation, and β and γ are hyper-parameters used to balance the three terms.

5 Experiments

To systematically validate the effectiveness of our proposed VIF framework in restoring fine-grained visual information, enhancing grounding capabilities, and maintaining general multimodal understanding, we conducted studies across 12 diverse benchmarks, spanning 15 distinct evaluation tasks. All experiments are conducted at the 7B scale, a controlled and widely adopted setting for MLLM. Under this setup, we train and validate VIF on a 7B-class backbone and benchmark it extensively across a broad range of multimodal tasks, with direct comparisons against multiple strong 7B-scale baselines. The results consistently demonstrate the effectiveness and robustness of our approach.

Table 3: **Ablation Study of the VIF Framework.**

Model	SEED-I	AI2D	HR-Bench		Vstar	REC (RefCOCO)		
			4K	8K		val	testA	testB
VIF (Ours)	**70.8**	**64.0**	**44.8**	**36.8**	**50.8**	**34.3**	**16.6**	**47.0**
- w/o AP	69.2	63.0	42.3	34.3	48.7	32.8	15.3	44.1
- w/o SP	68.2	62.5	36.8	32.9	49.2	29.6	13.9	40.7
Full-Seq	68.5	62.8	38.9	33.6	49.2	29.9	14.1	42.7
Deep-Only	68.2	63.1	39.1	33.0	49.7	30.7	15.6	43.1
Mid-Deep Feature	68.3	63.2	39.3	32.5	**50.8**	29.8	14.3	41.7

5.1 Datasets and Evaluation Metrics

To comprehensively evaluate model performance, we organize benchmarks into 4 dimensions.

(1) Multimodal Understanding. We evaluate general perception and cognition using MME (Fu et al., 2025) and SEED-Bench (Image) (Li et al., 2024), instruction-following with LLaVA-Bench (Liu et al., 2023b), and multi-image robustness with BLINK (Fu et al., 2024). **(2) Knowledge-Intensive and Logical Reasoning.** We adopt ScienceQA (Lu et al., 2022) and VQAv2 (Jia et al., 2024) as core VQA benchmarks, complemented by GQA (Ainslie et al., 2023) for compositional spatial reasoning and AI2D (Kembhavi et al., 2016) for scientific diagram understanding. **(3) OCR.** TextVQA (Singh et al., 2019) is used to assess optical character recognition on text-rich images. **(4) Fine-Grained Perception and Referring Expression Comprehension.** We use HR-Bench (Wang et al., 2025b) and VSTAR (Wu and Xie, 2023) to evaluate fine-grained visual perception, and RefCOCO (Yu et al., 2016) to assess visual grounding and referring expression comprehension.

5.2 Baselines

To evaluate our method, we compare it with representative MLLMs, grouped as follows: **General Baselines:** LLaVA-v1.5 (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), mPLUG-Owl2 (Ye et al., 2023), Qwen-VL (Bai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), LLaMA-Adapter-v2 (Gao et al., 2023), IDEFICS-9B (Laurençon et al., 2023), Mantis-8B-Fuyu (Jiang et al., 2024), and MiniGPT-v2 (Chen et al., 2023). **High-Resolution Models:** LLaVA-NeXT (Liu et al., 2024), LLaVA-1.5-HD (Liu et al., 2023a), Dragonfly (Thapa et al., 2024), and LLaVA-HR (Luo

et al., 2024). **Feature-Enhanced Models:** DenseConnector (Yao et al., 2024), MMFuser (Cao et al., 2024), LLaVA with ViCrop (Zhang et al., 2025a), IGVA (Li et al., 2026), and TG-LLaVA (Yan et al., 2025). We also include Qwen2.5-VL-7B (Bai et al., 2025b) and its Chain-of-Thought (CoT) fine-tuned variant, CoVT (Qin et al., 2025), and build our method on top of the CoVT backbone. For fair comparison, we strictly follow CoVT's original training protocol, using Qwen2.5-VL-7B with LoRA fine-tuning (rank 16, alpha 32), four training phases (4000/3000/3000/5000 steps), and the same vision-centric real-world and spatial perception data; because our fine-tuning dataset (LLaVA-LAION/CC/SBU) is smaller and lower quality than the closed-source data used by Qwen2.5-VL-7B, we adopt CoVT's CoT fine-tuning recipe instead of directly fine-tuning on our dataset.

5.3 Experimental Results

5.3.1 General Multimodal Capability

As shown in Table 1, VIF achieves competitive performance across a broad range of benchmarks. Compared with models of similar scale and settings, VIF shows consistent gains on most tasks, indicating its effectiveness and robustness.

Comparison with General Baselines. VIF performs favorably against standard 7B-scale models. Relative to its backbone LLaVA-v1.5, VIF achieves clear improvements on multiple benchmarks, including ScienceQA and LLaVA-Bench, and attains the best result on the multi-image benchmark BLINK, suggesting improved handling of fine-grained visual evidence. Compared with InstructBLIP and Qwen-VL-Chat, VIF exhibits more stable performance across multimodal tasks.

Comparison with High-Resolution Models. Despite using a standard input resolution of 336×336, VIF achieves results comparable to high-resolution approaches with substantially larger visual sequences. While ultra-high-resolution models retain advantages on pixel-level OCR, VIF demonstrates strong overall performance on comprehensive benchmarks such as MME and SEED-I, reflecting a favorable efficiency–performance trade-off.

Comparison with Feature Enhancement Models. VIF consistently outperforms existing feature enhancement methods such as IGVA and TG-LLaVA on comprehensive benchmarks. On SEED-I, VIF surpasses both baselines, suggesting that variational information flow is more effective than prior feature fusion or connector-based strategies for preserving task-relevant visual information.

5.3.2 Fine-Grained Perception and Referring Expression Comprehension

Table 2 summarizes results on fine-grained perception and referring expression comprehension. On **HR-Bench** (4K/8K) and **VSTAR**, VIF consistently outperforms the backbone LLaVA-v1.5 and feature-enhanced methods such as IGVA. For example, on VSTAR, VIF achieves 50.8, surpassing IGVA (48.2), demonstrating improved sensitivity to tiny objects and subtle visual cues. For visual grounding, VIF also outperforms MMFuser on RefCOCO (val/testB), indicating enhanced spatial reasoning and referring expression comprehension.

Generalization to Stronger Backbones. To examine the scalability of our approach, we further apply VIF to the advanced Qwen2.5-VL-7B backbone. Under the CoVT setting, although CoVT already exhibits strong reasoning performance, **VIF w/ CoVT** still yields consistent gains on fine-grained benchmarks. This suggests that VIF remains effective when paired with stronger backbones and advanced reasoning strategies. Moreover, this result helps position VIF with respect to recent "thinking with images" and visual chain-of-thought paradigms. While these approaches typically improve performance by enriching the input signal through iterative visual querying or external tools, VIF focuses on a complementary direction by strengthening how the MLLM backbone utilizes the already-provided visual inputs. The gains achieved by **VIF w/ CoVT** therefore indicate that VIF can serve as a complementary enhancement to stronger visual reasoning paradigms.

5.3.3 Ablation Study

Table 3 summarizes our ablation studies on the VIF framework. We evaluate the impact of: (1) removing answer posterior supervision; (2) omitting sparsity regularization; (3) applying importance scores across the full token sequence; (4) focusing on deep layers only ($l = l' \in 25, 27, 29, 31$); and (5) injecting middle-layer visual features into deep-layer representations.

Impact of Supervision and Regularization. Removing answer posterior supervision (*w/o AP*) consistently degrades performance (e.g., a 1.6% drop on SEED-I), confirming its role in guiding the model to focus on task-relevant visual regions. Omitting sparsity regularization (*w/o SP*) sharply

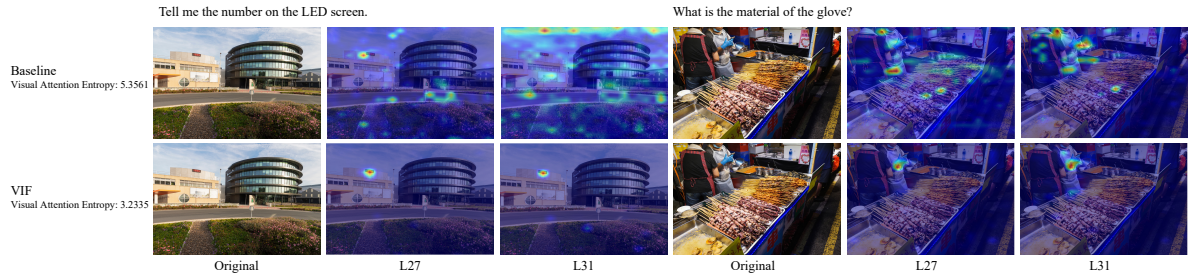

Figure 6: **visualization of attention maps comparing the baseline and our proposed model.** The top and bottom rows display the attention distributions of the baseline and our model, across different layers (L27 and L31).

Table 4: **Ablation study on extraction to injection layer configurations.**

Method / Configuration	Extraction Layers	Injection Layers	SEED-I	GQA	TextVQA	HR-Bench (4K)	Vstar
LLaVA-v1.5 (Baseline)	–	–	67.2	61.9	58.1	36.1	45.0
VIF (L-early, 4 pairs)	9,11,13,15	25,27,29,31	69.5	62.3	59.1	42.5	48.9
VIF (L-late, 4 pairs)	13,15,17,19	25,27,29,31	70.3	62.6	59.5	43.8	49.6
VIF (5 pairs)	9,11,13,15,17	23,25,27,29,31	**70.9**	**62.9**	59.8	**45.0**	**50.9**
VIF (Our Default)	**11,13,15,17**	**25,27,29,31**	70.8	62.8	**59.9**	44.8	50.8

reduces fine-grained performance: HR-Bench (4K) drops from 44.8 to 36.8, and REC (val) falls from 34.3 to 29.6. These results highlight the importance of sparsity in filtering noise and maintaining compact, discriminative information flow.

Scope and Layer Selection Analysis. Applying importance weighting to the full token sequence (*Full-Seq*) worsens results due to noise from indiscriminate enhancement. The *Deep-Only* variant underperforms, as deep-layer textual tokens already capture coarse visual summaries, making them less suitable for learning informative visual importance. The *Mid-Feature* strategy is less effective than our design, emphasizing the need to model attention flow rather than injecting visual features.

Impact of Extraction to Injection Layer Mapping. We compare different extraction to injection layer configurations, including earlier vs. later extraction layers and 4-pair vs. 5-pair mappings. Extracting from relatively early layers (*L-early, 4 pairs*) leads to slightly worse results across benchmarks. By contrast, using middle extraction layers and injecting into deeper layers performs more favorably, which validates our design choice of restoring visual information from intermediate representations into high-level reasoning layers. Although the 5-pair variant achieves slightly better results on several benchmarks, its gains over our default 4-pair setting are marginal. We therefore adopt the 4-pair mapping $[11, 13, 15, 17] \rightarrow [25, 27, 29, 31]$ as the default, as it provides a better balance between performance and computational cost.

5.3.4 Deep Layers Attention Analysis

Our approach was validated through both qualitative and quantitative analysis. As illustrated in Figure 6, unlike the baseline, which suffers from attention dispersion due to background noise (e.g., sky) or irrelevant objects (e.g., food), our model exhibits superior focus in deep layers (L27, L31). It effectively locates fine-grained targets, such as the small LED screen or the glove. Statistical analysis on 500 randomly sampled instances shows that our model reduces visual attention entropy from 5.3561 (baseline) to 3.2335, indicating improvements in both attention sparsity and certainty. These results further suggest that VIF helps the model form more concentrated and reliable task-relevant visual grounding in the deep layers.

6 Conclusion

In this work, we identified and addressed the critical bottleneck of **Visual Attenuation** in MLLMs, where fine-grained visual cues are progressively discarded and attention structures collapse in the deep layers. To address this challenge, we propose the **Variational Information Flow (VIF)** framework, a novel paradigm that shifts from passively augmenting input features to actively reconstructing deep-layer information flow. By leveraging a CVAE-based posterior learning mechanism together with a Gaussian Mixture Model (GMM) prior, VIF effectively reconstructs task-relevant visual signals while enforcing spatial sparsity.

Limitations

While our method effectively enhances fine-grained visual perception, it has two main limitations. First, although the proposed CVAE module is designed to be lightweight, it inevitably introduces additional parameters and computational overhead compared to the original backbone during inference. Future work that enables the model itself to maintain focused attention on critical visual cues could improve efficiency while enhancing performance. Second, since our approach operates at the level of information flow to restore attenuated features, its performance is partially constrained by the original input resolution of the frozen visual encoder (e.g., CLIP). For extremely small objects whose pixel-level information is lost during initial encoding, feature reconstruction remains challenging.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and suggestions, which helped improve this manuscript. This work was supported by the Beijing Natural Science Foundation under Grant No. L257006.

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu. 2024. Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. *arXiv preprint arXiv:2410.11829*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, and 1 others. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Runhui Huang, Xinpeng Ding, Chunwei Wang, Jianhua Han, Yulong Liu, Hengshuang Zhao, Hang Xu,

Lu Hou, Wei Zhang, and Xiaodan Liang. 2025. Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29814–29824.

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. 2024. Vqav2: Visual question answering for video quality assessment. *arXiv preprint arXiv:2411.03795*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

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

Xu Li, Yi Zheng, Haotian Chen, Xiaolei Chen, Yuxuan Liang, Chenghang Lai, Bin Li, and Xiangyang Xue. 2026. Instruction-guided fusion of multi-layer visual features in large vision-language models. *Pattern Recognition*, 170:111932.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2024. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Rahul Thapa, Kezhen Chen, Ian Covert, Rahul Chalamala, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2024. Dragonfly: Multi-resolution zoom-in encoding enhances vision-language models. *arXiv preprint arXiv:2406.00977*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025b. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7907–7915.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dawei Yan, Pengcheng Li, Yang Li, Hao Chen, Qingguo Chen, Weihua Luo, Wei Dong, Qingsen Yan, Haokui Zhang, and Chunhua Shen. 2025. Tg-llava: Text guided llava via learnable latent embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9076–9084.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. 2024. Dense connector for mllms. *Advances in Neural Information Processing Systems*, 37:33108–33140.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023). *arXiv preprint arXiv:2311.04257*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*, pages 69–85. Springer.

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

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025a. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025b. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.