Information-Regularized Attention for Visual-Centric Reasoning

Guohao Sun[1,2⋆], Xiaofang Wang[1], Yash Patel[1], Mengchen Liu[1], Zhiqiang Tao[2], Praveen Krishnan[1]

[1]FAIR at Meta [2]Rochester Institute of Technology

Abstract. Vision–language models (VLMs) have become a paradigm for multimodal learning, yet remain unstable due to object hallucination, weak visual grounding, and catastrophic forgetting after full-parameter instruction tuning. We claim these failures result from a lack of explicit control over visual representation learning during the standard next-token prediction objective. As a result, visual embeddings thus become passively optimized and prone to injecting redundant or spurious signals. To counter this, we introduce Information-Regularized Attention (IRA), a stochastic attention mechanism that explicitly regulates the amount of visual information injected into the hidden states of intermediate transformer layers. This local reparameterization translates uncertainty about visual representations into local noise that is independent across data points. Beyond evaluating model performance, we also quantify embedding properties, where IRA produces smoother curvature trajectories and suppresses attention-sink across all layers, indicating a more stable transformation of the visual signal. Our results suggest that stochastic attention is not merely a regularizer but a key contributor to representation learning in a generative architecture, offering a new direction for building more reliable VLMs.

Keywords: Representation learning · Information regularization · Vision-Language understanding

1 Introduction

Vision-language models (VLMs) have emerged as a general-purpose framework for multimodal understanding, achieving strong performance across tasks such as visual question answering, image captioning, and multimodal dialogue [2, 17, 28, 35]. Despite this progress, modern VLMs remain limited by reliability issues, including object hallucination and unreliable grounding, where generated content is not supported by the visual input [31, 48]. These failures suggest that standard next-text-token prediction objectives do not sufficiently align vision and language information in latent space.

⋆ Part of this work is done as intern at Meta.
Correspondence: Zhiqiang.Tao@rit.edu, pkrishnan@meta.com

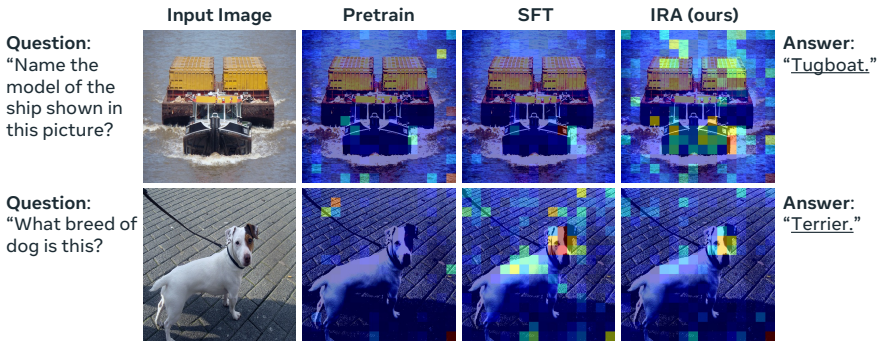


Fig. 1: Cross-attention of a VLM. We look at the normalized attention, i.e., $Attn$(answer $\rightarrow$ visual tokens) of the last layer. The model indicates a bad visual dependency without visual instruction tuning. **SFT** improves alignment between visual cues and output text, but biased information is introduced due to unregularized visual embedding. **IRA** restricts biased knowledge and only encodes necessary information.

Current VLM training paradigms are largely data-centric, relying on increasingly diverse forms of post-training supervision, including visual instruction tuning [33, 59], preference optimization [41, 42, 45, 57, 58], and policy optimization [49, 50, 56]. While effective, these approaches primarily improve model behavior by expanding the supervision signals, rather than directly regularizing the feature representations. Under the standard next-token prediction objective, all the embeddings are optimized only indirectly through language supervision. As a result, task-irrelevant or noisy visual signals can propagate through attention layers and interfere with cross-modal reasoning.

This issue is reflected in recent observations of attention sinks [5,18,25,34] and spike values in attention heads [61,62,71], where attention collapses onto semantically uninformative tokens and produces noisy cross-modal interactions [24,36, 48]. Our study in Fig. 1 provides further evidence: the pretrained VLM fails to attend to the relevant visual regions, whereas standard supervised instructional fine-tuning improves alignment only partially and still exhibits biased, noisy attention. These findings suggest that improving VLM reliability requires moving beyond data-centric post-training alone. Instead, we aim to improve visual representation learning within the generative architecture through end-to-end training, thereby achieving robust visual embeddings.

Prior work, including attention weight optimization [14,26] and gating mechanisms [44], has shown that controlling the information flow within the attention module can effectively reduce the attention sink. Despite their effectiveness, they operate at the level of attention weights or attention outputs rather than improving representation learning in the intermediate layers [6,24]. In contrast, we address the attention problem by introducing an information-theoretic mechanism that regularizes the attention's input embedding. We posit that hallucinations arise from a deficiency in visual mutual information [68]: the token

prediction is overwhelmed by *noisy* visual tokens [25, 60] that dilute the signal of query-relevant evidence. This occurs when the model fails to differentiate between semantically coherent and incoherent visual-textual pairs within its internal hidden representations [36, 52].

To this end, we propose Information-Regularized Attention (IRA), a mechanism that explicitly regulates the amount of visual information injected into the visual hidden states within the attention module (see Fig. 2). After evaluating embedding quality using geometric [22] measures (e.g., curvature), we find that representations that exhibit highly curved trajectories across transformer layers tend to produce unstable attention patterns and higher sink ratios, whereas smoother, more linear trajectories are associated with improved grounding and prediction. This suggests that embedding linearity is a key structural property underlying reliable representation learning.

Our contributions are summarized as follows:

- We propose a novel Information-Regularized Attention (IRA) mechanism that regulates visual representations before they enter cross-modal attention, explicitly controlling visual information at the representation level during full-parameter instruction tuning.
- We demonstrate that IRA improves the representation geometry, leading to greater robustness across tasks and stable training dynamics.
- We identify a correlation between the attention sink phenomenon and representational curvature, suggesting an avenue for future work to use representational geometry to design more robust multimodal architectures.

2 Related Works

2.1 Representation learning and neural dynamics.

The evolution of general-purpose VLMs [4, 10, 27, 33] has focused on bridging vision encoders with LLMs. However, understanding how these models encode and organize information requires deeper analysis. Early research utilized linear probes [1] and SVCCA [46] to examine feature dynamics, though these were largely confined to vision-only or shallow networks. Recent literature has extended to layer-wise analysis in large-scale LLMs, revealing that linguistic features and semantic roles are most effectively encoded in the intermediate Transformer layers [52]. Specifically, the middle layers contain surprisingly robust features [66], challenging the traditional emphasis on final-layer representations. Our work builds on this by regularizing the representations within intermediate layers to achieve robust embeddings.

2.2 Information regularization.

Regularization is a common technique to improve the generalization performance of deep neural networks, and various implementations are available depending

on the network architecture and target application. A well-known example is dropout [54], which randomly turns off a subset of hidden units in neural networks by multiplying a Bernoulli-distributed noise term. On the other hand, information bottleneck [3, 63] formalizes representation learning as a principled trade-off between task sufficiency and information compression [21, 34]. Recent work has begun applying these principles to VLM robustness. For example, VIB-Probe [76] utilizes VIB to detect and mitigate hallucinations by filtering noise from internal attention heads. However, most of the works treat the latent variable as external knowledge to support the downstream task. Instead, this work uses the information bottleneck as a regularization tool, aiming to directly influence representation learning within the transformer layers from a generative model.

2.3 Mitigating attention sinks.

Recent studies have highlighted phenomena in transformer architectures that hinder reliability, namely attention sinks [18, 72] and massive activations [61]. In causal generative models (e.g., LLM and VLM), attention sink [25] occurs when the model attends to irrelevant tokens, often driven by extreme outliers in the hidden state. Current mitigation strategies include gating mechanisms [44] and attention distribution optimization [24,26,78]. While these methods optimize the attention weights themselves, they often treat the underlying representations untouched. Our proposed IRA goes beyond attention distribution optimization by regularizing the inner representations prior to attention computation.

3 Method

3.1 Information-Theoretic View of VLM

A vision-language model (VLM) consists of a vision encoder, a projector, and a large language model (LLM). The input to the LLM is a concatenated sequence of visual x_{img} and language tokens x_{txt}. We use x to denote the set of all img and txt tokens concatenated together. Within the LLM, each transformer layer $f(\cdot)$ applies a deterministic mapping to produce hidden states $h^{(\ell+1)} = f(h^{(\ell)})$, where $h^{(0)} = x$. The final hidden states $h^{(L)}$ are decoded into output tokens y. Therefore, the entire model defines a Markov chain:

$$x \to h^{(1)} \to \cdots \to h^{(L)} \to y, \tag{1}$$

and $p_\theta(y \mid h^{(L)})$ is the probability of final answer y given output layer hidden states $h^{(L)}$, predicted by a model parameterized by θ.

After simplifying equation 1 into $x \to h \to y$, we can potentially formulate a mutual information between h, and (x, y) as:

$$\mathbb{I}(h; x, y) = \mathbb{I}(h; x) + \mathbb{I}(h; y \mid x) \tag{2}$$

$$= \mathbb{I}(h; y) + \mathbb{I}(h; x \mid y)\,, \tag{3}$$

where $\mathbb{I}(h;x) = \mathbb{I}(h;y) + \mathbb{I}(h;x \mid y) - \mathbb{I}(h;y \mid x)$, and the IB objective becomes:

$$\max \mathbb{I}(h;y) - \beta\, \mathbb{I}(h;x \mid y)\,. \tag{4}$$

Given this objective, we can learn a representation h that preserves useful predictive information and suppresses nuisance input information. However, h is a deterministic mapping of x in standard LLM, and the next-text-token prediction using supervised finetuning (SFT) optimizes

$$\max_{\theta} \mathbb{E}\left[\log p_\theta(y \mid h)\right] \approx \max_{\theta} \mathbb{I}(h;y)\,, \tag{5}$$

without discarding irrelevant information in the condition feature h. We hypothesize that using SFT alone may encode noisy information into visual embeddings, diluting attention and reducing discrimination among visual tokens.

To address this challenge, we propose that information regularization can be formulated as injecting random noise into h, yielding stochastic hidden units z. By doing this, we can leverage well-defined probabilistic formulations to analyze the conventional training procedure and propose better optimization approaches.

3.2 Information Regularization

In this work, we formulate information regularization as a variational inference problem [40] in latent space. Given the representations $h^{(\ell)}$ of input from each layer, we aim to sample a latent variable $z^{(\ell)}$ from a posterior distribution as:

$$z^{(\ell)} \sim p(z^{(\ell)} \mid h^{(\ell)})\,, \tag{6}$$

but approximating posterior is intractable through Bayes' rule $p(z \mid h) = \frac{p(h|z)p(z)}{p(h)}$. This work involves variational inference to approximate the intractable posterior. In this formulation, inference is cast as an optimization problem in which we optimize the model parameters ϕ of $q_\phi(z^{(\ell)} \mid h^{(\ell)})$ to approximate $p(z^{(\ell)} \mid h^{(\ell)})$.

Such a variational principle provides a tractable lower bound as

$$Ł(\theta,\phi|y) = \mathbb{E}\left[\log p_\theta(y \mid z^{(L)})\right] - \sum_{\ell=1}^{\mathcal{L}} D_{\mathrm{KL}}(q_\phi(z^{(\ell)} \mid h^{(\ell)}) \,\|\, p(z^{(\ell)}))\,. \tag{7}$$

This work reinterprets each layer as performing a data-dependent amortized variational inference. Specifically, instead of introducing a separate inference network $q_\phi(\cdot)$, we attach a lightweight parametric head to each transformer layer to produce the posterior as $q_{\theta,\phi}(z^{(\ell)} \mid h^{(\ell)})$, thereby tightly coupling latent variable encoding with the forward pass of the model. This design enables efficient layer-wise stochasticity without additional encoding overhead, and allows the variational parameters (i.e., ϕ) to co-evolve with the backbone representations (i.e., θ) during end-to-end training.

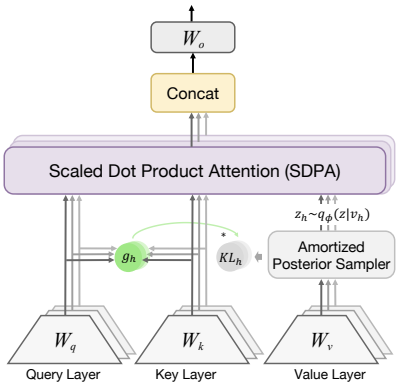


Fig. 2: Architecture of the proposed Information-Regularized Attention (IRA). We introduce a lightweight posterior sampler that incorporates stochastic representations prior to the attention computation.

3.3 Regularizing Visual Representation in Attention Module

Multi-head attention (MHA) plays a central role in embedding updates by enabling the model to process information through multiple parallel pathways [64]. Each attention head applies independent projections of **queries**, **keys**, and **values**, allowing the model to attend to different representation subspaces and relational structures simultaneously. From an information-theoretic perspective, the value states v are the primary channel through which information enters the output representation, where $v = hW_v$ with $v \in \mathbb{R}^{\mathcal{H}*d}$, $\mathcal{H}$ is the number of attention heads, and d is the dimension for each. Therefore, instead of regularizing each transformer layer's output representation, we apply it before value states enter the attention computation, as shown in Fig. 2. This mechanism enables the subsequent MLP layer to adapt smoothly to the stochastic attention output. This work mainly focuses on regularizing the visual representation by first extracting the visual components of the input value states.

Conceptually, the causal LLM in our VLM system can be viewed as a flexible architecture that shares parameters between the generative and inference processes, using the same hidden representations to produce outputs and to parameterize the variational posterior. This parameter sharing is reminiscent of hierarchical variational models such as Ladder-VAE [53], although our formulation does not explicitly implement separate bottom-up and top-down inference pathways. Therefore, to alleviate the same early collapse and local-optimal issues [53], we introduce a data-dependent prior and a prior-dependent posterior.

Prior. This work defines the prior distribution of z as data-dependent [13, 32], enabling adaptive regularization of latent variables conditioned on intermediate representations. However, such a design makes KL collapse to zero easily (i.e., $D_{\mathrm{KL}}(q_\theta(z \mid v) \,\|\, p(z \mid v)) \approx 0$), since the posterior and prior both condition

on the same input. To address this problem, we propose $p(z \mid \lfloor v \rfloor)$, where the input v is detached from the gradient graph. This stop-gradient design prevents trivial posterior–prior collapse and ensures that the KL term acts as a stable regularization signal. Overall, the prior distribution of z is a perturbation of v parameterized by a Gaussian:

$$p_\lambda(z \mid \lfloor v \rfloor) = \mathcal{N}(\lfloor v \rfloor, \sigma_p^2 I), \tag{8}$$

where $\sigma_p^2 = \exp(\log \sigma_\lambda^2) \in \mathbb{R}^{\mathcal{H}*d}$, and $\log \sigma_\lambda^2$ is a learnable parameter shared by all the latent variables from the same layer. This formulation can be interpreted as learning a zero-mean Gaussian noise from the entire training data and regularizing the intermediate representations of each data point, thereby enabling adaptive regularization while preserving the underlying semantic content.

Posterior. The posterior is parameterized as an Isotropic Gaussian as:

$$q_{\theta,\phi}(z|v) = \mathcal{N}(v + \Delta_\phi(v), \sigma_q^2 I), \tag{9}$$

where $\sigma_q^2 = \exp(\log \sigma_\phi^2(v))$, $\Delta_\phi(\cdot)$ and $\log \sigma_\phi^2(\cdot)$ are predicted by a linearly layer parameterized by ϕ condition on the current value state. The mean residual $\Delta_\phi(v) \in \mathbb{R}^{\mathcal{H}*d}$ captures an adaptive perturbation to the pretrained representation v, allowing the model to refine visual features when beneficial for prediction. Meanwhile, $\sigma_q^2 \in \mathbb{R}^{\mathcal{H}*1}$ controls the uncertainty of this refinement, determining how much flexibility is permitted for each attention head.

KL regularization. The independent KL divergence between posterior and prior in each transformer layer becomes:

$$D_{\mathrm{KL}}(q_{\theta,\phi} \,\|\, p_\lambda) = \frac{1}{2}\left[\frac{\|\Delta_\phi(v)\|^2 + \sigma_q^2}{\sigma_p^2} - 1 + \log \sigma_p^2 - \log \sigma_q^2\right]. \tag{10}$$

The KL regularization constrains the posterior to remain close to a detached, reference distribution centered at the original value state. The mean-shift penalty limits excessive deviation from the pretrained representation, while the variance-matching terms regulate the magnitude of injected stochasticity. To propagate gradients back through q_ϕ, we use the reparameterization trick as:

$$z = v + \Delta_\phi(v) + \sigma_q * \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)\,. \tag{11}$$

This denotes as injecting learnable feature-level Gaussian noise (i.e., $\Delta_\phi(v) + \sigma_q * \epsilon$) in the deterministic value states v.

3.4 Uncertainty-Conditioned Information Regularization

The KL regularization in equation 10 applies a uniform penalty across all visual tokens and attention heads, regardless of their semantic importance. However,

in multi-head attention, different heads specialize in different types of visual relationships. Some heads may focus on spatial alignment, others on semantic attributes, and others on fine-grained textures. This observation suggests that regularization should be applied selectively rather than uniformly. Specifically, more should be applied to visual tokens that are both: i) **Unimportant**, where attention assigns low relevance, and ii) **Uncertain**, where cross-modal attention distribution has high entropy. We therefore introduce a weighting mechanism that adaptively adjusts the regularization magnitude for each visual token. Let q_{txt} and k_{img} denote the text and visual parts of the query and key states, respectively. We then compute the normalized attention distribution as:

$$p = \text{softmax}(\frac{\lfloor q_{txt} \rfloor^\top \lfloor k_{img} \rfloor}{\sqrt{d}}) \,, \tag{12}$$

where $\lfloor \cdot \rfloor$ is gradient detachment, preventing the KL regularization term from directly influencing q and k representations, which could otherwise lead to unstable optimization.

Given $p \in \mathbb{R}^{\mathcal{T}*\mathcal{S}*\mathcal{H}}$, where $\mathcal{T}$ and $\mathcal{S}$ are the number of text and visual tokens, we compute the un-normalized importance score of the i-th visual token as:

$$a_i = \frac{1}{\mathcal{T}} \sum_{j=1}^{\mathcal{T}} p_{j,i} \,, \tag{13}$$

where $a_i \in \mathbb{R}^{\mathcal{H}}$ indicates the average attention mass assigned to the i-th visual token by the language contexts in each head. However, relying solely on high or low importance scores is unreliable, as the softmax function's normalization often leads to an attention sink [5,18]. In particular, high attention weights do not necessarily indicate semantically meaningful information, as they may arise from biased distributions. To address this limitation, we propose reweighting token importance based on visual attention entropy. Entropy quantifies the dispersion of attention and provides a signal indicating the confidence of attention allocation. Intuitively, a low-entropy distribution reflects confidence and requires less regularization, whereas a high-entropy distribution suggests a potentially noisy representation and thus requires more regularization. To achieve this goal, we compute normalized entropy as:

$$\mathbb{H} = \frac{1}{\mathcal{H}} \sum_{n=1}^{\mathcal{H}} -\frac{\sum_{j=1}^{\mathcal{T}} p_{j,n} \log p_{j,n}}{\log \mathcal{T}} \,. \tag{14}$$

To this end, we construct a token-wise weighting mechanism (i.e., $g_i = \mathbb{H} \cdot (1 - a_i)$) to adaptively weight the KL penalty defined in equation 10 for the i-th stochastic visual token as:

$$D_{\text{KL}}(q_{\theta,\phi}(z \mid v) \,\|\, p_\lambda(z \mid \lfloor v \rfloor)) = \mathbb{E}_{(i)\sim Unif} \left[g_i * D_{\text{KL}}(q_{\theta,\phi}(z_i \mid v_i) \,\|\, p_\lambda(z_i \mid \lfloor v_i \rfloor)) \right] .$$

Moreover, we control the magnitude of noise by $z_i = v_i + \Delta_\phi(v_i) + g_i * \sigma_i * \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$, preventing excessive noise injection into relatively important

visual representations and encourage stronger regularization on the others. Notably, the weighting estimate g is only required during training, and we use $z = v + \Delta_\phi(v)$ at inference time.

Overall, the final objective function becomes:

$$Ł(\theta, \phi, \lambda|y) = \mathbb{E}\left[\log p_\theta(y \mid z^{(\mathcal{L})})\right] - \beta \sum_{\ell=1}^{\mathcal{L}} D_{\mathrm{KL}}(q_{\theta,\phi}(z^{(\ell)} \mid v^{(\ell)}) \,\|\, p_\lambda(z^{(\ell)} \mid \left\lfloor v^{(\ell)} \right\rfloor)),$$

where β is scheduled using cosine interpolation, increasing from 0 to β_{max} during the first $k\%$ of the total training steps. We conduct experiments with β_{max} for a better regularization-performance trade-off in the Supplementary. Typically, we find $\beta_{max} = 1 \times 10^{-4}$ and $k = 50$ achieves the best performance. By initializing training with only reconstruction loss and gradually introducing the variational regularization term, we avoid local minima at $D_{\mathrm{KL}}(q \,\|\, p) \approx 0$ during optimization. This idea has previously been considered in [7, 53].

Table 1: Hyper-parameters for training the models in different stages. We denote Stage 1/1.5 as pre-training (**Pt**) and Stage 2 as full-parameter supervised fine-tuning (**SFT**). The proposed **IRA** is an improvement of **SFT** in **Stage 2** training.

	Model	Trainable	#Data	Batch size	Context len	LR (Backbone)	LR (IRA)	W Decay	Epochs	IRA
Stage 1	InternVL2.5-8B	MLP	-	512	16384	2×10^{-4}	-	0.05	-	No
	LLaVA-OneVision-8B		558k	512	8182	2×10^{-6} / 1×10^{-5}	-	0	1	
Stage 1.5	InternVL2.5-8B	ViT + MLP	-	512	16384	1×10^{-5}	-	0.05	-	No
	LLaVA-OneVision-8B	Full Model	4M	256	32768	2×10^{-6}/1×10^{-5}	-	0	1	
Stage 2 (ours)	InternVL2-8B	Full Model	3.2M	512	16384	1×10^{-5}	1×10^{-4}	0.05/0.01	1	Yes
	InternVL2.5-8B			512	16384	1×10^{-5}	1×10^{-4}	0.05/0.01		
	LLaVA-OneVision-8B			256	32768	2×10^{-6}/1×10^{-5}	1×10^{-4}	0		

4 Experiments

4.1 Experimental Setup

Model architectures. We mainly conduct experiments on open-sourced VLMs, including InternVL2, InternVL2.5 [10] and LLaVA-Onevision1.5 [27]. In practice, we chose these models because they encode visual input into a token sequence that can be mapped back to the pixel level, enabling us to study visual grounding with greater interpretability.

Baseline method. Previous works have empirically shown that supervised fine-tuning (SFT) has poor generalization and severe catastrophic forgetting [9,69,75] when models are adapted to diverse tasks and domains after full-parameter instruction tuning. Please see the Supplementary for our empirical analysis of catastrophic forgetting. This tension raises a fundamental challenge for robust representation learning [11, 40, 43] and visual understanding.

Training details. Typically, training a VLM from scratch requires first *pre-training* on image-text pairs and then full-parameter *supervised fine-tuning* on

Table 2: Comparison of training methods for general-purpose VLMs. We report the summation of perception and cognition scores for MME. The best results are **bold**.

Method	STEM		General QA					Text & Chart		
	MMMU-Pro	MMMU	MME	MME-RW	MMStar	MMBench	OK-VQA	TextVQA	ChartQA	DocVQA
InternVL2-Pt	26.0	42.1	1598	31.2	43.7	70.9	**44.3**	64.8	70.3	76.2
+ SFT	30.1	45.1	1792	37.6	**60.2**	80.4	42.3	73.0	**80.1**	85.3
+ IRA	**31.2**	**45.7**	**1959**	**40.5**	58.8	**81.0**	43.0	**73.8**	79.9	**85.5**
InternVL2.5-Pt	26.7	41.3	1669	29.5	47.3	72.7	43.4	69.2	73.1	81.8
+ SFT	30.4	46.4	1981	40.4	**61.1**	80.6	39.8	74.5	81.0	**86.7**
+ IRA	**30.6**	**47.6**	**2038**	**40.8**	58.8	**81.8**	**43.6**	**74.7**	**81.8**	**86.7**
LLaVA-OV-Pt	20.5	36.3	1530	28.0	30.6	67.3	11.0	36.8	37.6	11.5
+ SFT	27.4	44.4	2030	**40.9**	57.9	79.5	41.7	75.9	79.9	88.3
+ IRA	**28.0**	**45.3**	**2109**	40.0	**58.1**	**79.6**	**47.4**	**77.1**	**80.7**	**88.5**

visual instructional data, denoted as **Pt** and **SFT**. The proposed IRA aims to improve full-parameter training during the SFT stage by initializing the model with the stage-1.5 checkpoint, as shown in Table 1. Note that we set a 10× higher learning rate for the IRA parameters, allowing the regularization term to quickly catch up with task optimization. At the model architecture level, IRA introduces negligible additional parameters to the attention module, including a $\texttt{Linear}(d, d{+}1)$ and an $\texttt{Embedding}(\mathcal{H}, d)$ per layer, where $d = 128$ and $\mathcal{H} = 8$ indicate head dimension and #heads. Such a modification transforms the deterministic propagation of visual embeddings into a stochastic process.

Training data. We mainly follow LLaVA-Onevision1.5 [27] to prepare our training data, but we consider only the single-image subset with 3.2M samples, as the goal is to evaluate the generalization of SFT and IRA to multi-image and video understanding after training. We aim to test how different training methods mitigate forgetting and evaluate their robustness in an OOD scenario.

4.2 General Visual Understanding

VLM is a general-purpose assistant in the wild. To validate the capabilities in real-world scenarios with open-form instructions, we use MMStar [8], MME [15], MME-Realworld [77], and RealWorldQA [70]. Beyond chat capability, visual perception assesses the model's reasoning ability, so we adopt MMMU [73] and MMMU-Pro [74]. Previous studies have shown that generalization usually improves semantic reasoning but harms fine-grained visual understanding(e.g., text parsing) [23, 55]. Therefore, we incorporate TextVQA [51], ChartQA [38], DocVQA [39], and EmbSpatial [12].

As shown in Table 2, the improvements of IRA focus on reasoning-intensive benchmarks (STEM and General QA), indicating that IRA primarily enhances compositional reasoning and the robustness of cross-modal grounding rather than merely strengthening surface-level alignment. This suggests that the method is not acting as a generic regularizer but rather specifically improving how visual evidence is injected and used during reasoning. We observe that IRA yields the largest gains on knowledge-intensive benchmarks, OK-VQA [37], suggesting

Table 3: Comparison of robustness and generalization.

Method	Multi-image and Video			Robustness				Spatial
	MuirBench	BLINK	MVBench	POPE	HallBench	VLM-Bias	VLM-Blind	EmbSpatial
InternVL2-Pt	31.1	41.1	35.6	83.9	33.6	17.6	**39.0**	52.3
+ SFT	**38.3**	45.1	48.7	86.4	35.7	18.1	33.4	63.9
+ IRA	36.8	**45.3**	**48.9**	**86.5**	**36.0**	**18.4**	33.9	**64.9**
InternVL2.5-Pt	31.6	39.5	47.4	86.6	32.5	18.3	36.4	51.7
+ SFT	35.2	**47.1**	51.5	87.0	37.6	17.8	33.9	**64.5**
+ IRA	**35.4**	45.5	**52.0**	**87.5**	**37.9**	**18.3**	**37.3**	64.2
LLaVA-OV-Pt	37.4	24.6	41.9	85.7	21.7	16.9	5.8	51.4
+ SFT	40.9	42.9	53.0	88.6	**38.1**	**20.9**	19.0	**63.5**
+ IRA	**41.4**	**43.2**	**54.0**	**88.7**	37.0	19.7	**19.1**	**63.5**

that noisy Attention improves semantic abstraction and cross-modal reasoning. Text-intensive tasks require high-fidelity visual transmission, where nearly all signals are informative. The improvements on TextVQA, DocVQA, and ChartQA indicate that the proposed uncertainty-conditioned noise successfully prevents over-regularization of critical visual details.

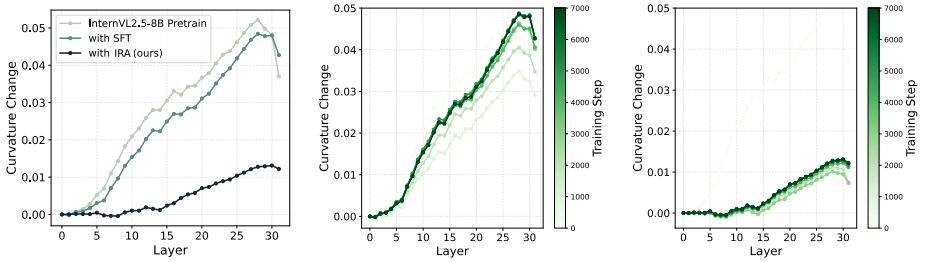
4.3 Robustness and Generalization

We evaluate the model's robustness using four benchmarks. POPE [30] formulates hallucination detection as a binary object-existence verification task, enabling precise measurement of false positives induced by language priors. HallusionBench [19] further probes fine-grained visual reasoning and illusion-induced errors through paired yes/no questions that require consistent, evidence-based judgments. VLM-are-biased [65] and VLM-are-blind [47] evaluate the model's ability to actually see low-level vision rather than relying on biased prior knowledge or reasoning. As shown in Table 3, IRA achieves improved robustness on the InternVL family model.

To further assess the generalizability of each method, we evaluate its performance on OOD tasks. For multi-image understanding, we report MuirBench [67] and BLINK [16], and for video understanding, we use MVBench [29]. In Table 3, IRA consistently improves video understanding and competitive performance on the multi-image tasks.

4.4 Representation Evaluation

The brain transforms the incoming visual input to make it more predictable [20]. Predictive models operate by extrapolating future states from current representations, a process that is well-conditioned when internal representations evolve approximately linearly across model depth [22]. Inspired by these works, we developed a curvature metric based on neural trajectories of image tokens and used it in all analyses. Considering at layer ℓ, we first extract the visual embeddings $x_1^{(\ell)}, x_2^{(\ell)} \cdots x_I^{(\ell)}$ and compute $v_1^{(\ell)}, v_2^{(\ell)} \cdots v_{I-1}^{(\ell)}$ as the difference between two adjacent states $v_k^{(\ell)} = x_{k+1}^{(\ell)} - x_k^{(\ell)}$. We then compute curvature as the angle between



(a) Curvature Change **(b)** InternVL2.5-SFT **(c)** InternVL2.5-IRA

Fig. 3: Analysis of the representation straightening. The y-axis shows the ΔC of each layer. A straighter trajectory indicates a smoother update to the representation. A model with better embedding quality exhibits a straighter curvature trajectory.

these vectors as $c_k^{(\ell)}=\arccos(\frac{v_{k+1}^{(\ell)}\cdot v_k^{(\ell)}}{\|v_{k+1}^{(\ell)}\|\|v_k^{(\ell)}\|})$. Then the average curvature across the visual token sequence is computed as $C^{(\ell)}=\frac{1}{K}\sum_{i=1}^{K}c_i^{(\ell)}$. Finally, we compute a change in curvature between each layer and the first layer as $\Delta C^{(\ell)}=C^{(\ell)}-C^{(0)}$. Our key insight is that curvature serves as a proxy for predictability: a straighter curvature trajectory across layers corresponds to smoother, more stable representational updates. As shown in Fig. 3, our method suppresses irregular variations and enforces straighter trajectories. Empirically, we show that reducing curvature is associated with improved training stability and attention distribution.

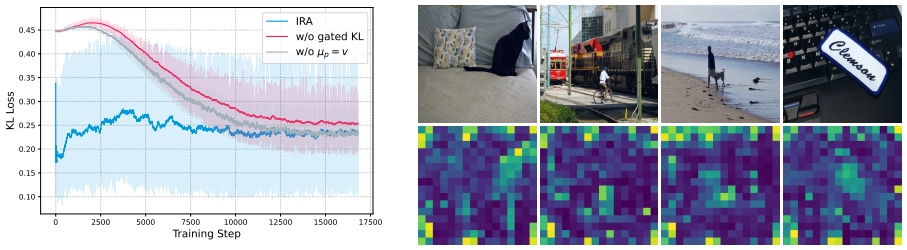
4.5 Ablation Study

The proposed IRA contains two key components: an uncertainty-guided weighting mechanism that adaptively controls the regularization, and a data-dependent prior centered on the pretrained visual representation. We conduct ablation studies in Table 4 and observe that removing the weighting gate consistently degrades performance across most benchmarks, with particularly large drops on MMMU and MuirBench. This suggests that applying a uniform noise injection is too aggressive, while the weighting mechanism effectively balances the regularization of uncertain visual representations. We study the effect of the prior design by replacing the data-dependent prior with a global learnable embedding initialized to zero. This modification consistently reduces performance, especially on text-intensive benchmarks, indicating that anchoring the prior to pretrained visual representations can save fine-grained visual information.

Moreover, we analyze the KL loss during training in Fig. 4a, where IRA maintains stable KL regularization throughout training, whereas removing the weighted KL results in higher KL values. This indicates that the weighting mechanism effectively prevents excessive regularization during the early stages of training. In contrast, replacing the pretrained-centered prior with a learnable prior also leads to greater KL fluctuations, suggesting that anchoring the prior to pretrained visual representations stabilizes the regularization process. After

Table 4: Ablation study of our main designs. w/o IRA indicates vanilla SFT.

Method	MMMU	RealWorldQA	VQA^{text}	ChartQA	EmbSpatial	MuirBench	CVBench	Avg.
IRA	46.1	58.3	70.3	79.8	64.7	38.4	71.7	**61.3**
w/o weighted KL	45.1	57.0	69.5	80.4	65.6	35.9	71.8	60.7
w/o $\mu_p = v$	45.6	58.2	69.4	79.6	64.6	34.6	72.5	60.6
w/o IRA	44.9	56.6	69.6	76.5	63.5	34.7	70.7	59.5



(a) KL Loss During Training **(b)** Visual Token Information Map

Fig. 4: (a) Adaptive KL enables fewer regularization at the start of training, and gradually regularizes the prior knowledge. (b) Token-wise KL divergence across visual tokens, where brighter regions indicate stronger deviation from the prior and higher information capacity allocated to the corresponding visual patches.

training, we visualize the KL divergence between the posterior distribution of each visual token and the global prior in Fig. 4b, where brighter means higher KL, indicating more information. We have observed that the main objects in the image tend to be more important, while the background is less important. We find that some border tokens exhibit unexpectedly high importance, which may be due to their low frequency in the entire training data.

4.6 Impact of IRA in Different Transformer Layers

Previous studies [1, 24, 52] have shown that different transformer layers play distinct roles in representation learning. In particular, mid-depth layers progressively compress embeddings and remove noise, while later layers focus more on language-aligned reasoning and output decoding. Prior analyses further indicate that entropy decreases across layers during training, suggesting a gradual consolidation of information.

Motivated by these observations, we investigate where the proposed noisy Attention should be applied within the LLM decode of the VLM system. For InternVL2.5-8B, whose language model has 32 layers, we apply IRA to different layer ranges and train the model for 10k steps.

As shown in Table 5, applying IRA to a contiguous mid-to-late subset of layers (60%–80%) achieves the best overall performance in our study, yielding improvements of +0.58%, respectively. These configurations consistently improve performance on multimodal reasoning benchmarks such as MME, VQA^{text}, and

Table 5: Study of applying IRA in different layers. Assuming an LLM with 32 layers, $20\% - 80\%$ means applying IRA in transformer layers 6-26. We utilize the $60\% - 80\%$ setting in all experiments.

Layer Depth	β_{max}	MMMU	MME	VQA^{text}	ChartQA	EmbSpatial	MuirBench	CVBench	Improve(↑ %)
baseline	-	45.1	1839	72.0	78.4	63.6	37.2	72.1	+0.00%
$0\% - 100\%$	4×10^{-4}	47.3	1940	70.0	76.0	61.2	34.4	72.3	-0.93%
$20\% - 80\%$	4×10^{-4}	45.1	1956	70.0	75.7	61.0	35.0	68.7	-2.08%
$20\% - 40\%$	8×10^{-5}	43.4	1791	70.6	77.4	62.9	38.2	72.7	-1.30%
$40\% - 60\%$	8×10^{-5}	43.6	1956	71.7	78.4	62.8	33.6	70.2	-1.56%
$50\% - 70\%$	8×10^{-5}	45.2	1930	71.7	79.4	63.2	35.8	73.5	+0.51%
$60\% - 80\%$	8×10^{-5}	44.8	1981	71.9	78.4	63.9	35.3	73.4	**+0.58%**

CVBench, while maintaining competitive results on spatial tasks like EmbSpatial. This suggests that mid-to-late layers are the most effective stage for regulating visual information, where representations are sufficiently abstract to benefit from compression but still retain critical visual grounding signals. In contrast, applying IRA across all layers (0%–100%) or over a broad range (20%–80%) degrades performance. This indicates that over-regularization across the entire network can excessively regularize visual information, hindering both reasoning and generalization. Similarly, applying IRA to earlier layers (20%–40%) yields only limited gains, since these layers primarily encode low-level visual features and lack strong cross-modal interactions.

Overall, these results suggest that mid-to-late transformer layers are the critical stage for integrating and refining visual representations, and that applying regularization in this region achieves the best trade-off between visual compression and reasoning capability.

4.7 IRA Reduces Visual Attention Sink

To quantify the model's dependence on visual inputs, we provide a layer-wise analysis of the visual attention distribution and sink ratio. Specifically, we first extract the attention weights at layer l as $\mathcal{A}^{(\ell)} \in \mathbb{R}^{\mathcal{H} \times \mathcal{T} \times \mathcal{S}}$ assigned by teacher-forced output tokens to each input visual tokens, where $\mathcal{T}$ and $\mathcal{S}$ are the number of output and visual tokens, and $\mathcal{H}$ is the number of heads. For each transformer layer, the aggregated visual attention map is computed as $A^{(\ell)} = \frac{1}{\mathcal{T} \cdot \mathcal{H}} \sum_{i=1}^{\mathcal{T}} \sum_{n=1}^{\mathcal{H}} \mathcal{A}_{i,n}^{(\ell)}$. For all the visual tokens in the sequence, an attention sink [62] exists when there are tokens that receive more than ϵ average attention, i.e., $s_\epsilon^{(\ell)} = \max_{1 \le k \le \mathcal{S}} A_k^{(\ell)} > \epsilon$. Finally, we report the model-level sink ratio by averaging $s_\epsilon^{(\ell)}$ over all the layers. In our experiments, we use $\epsilon = 0.15$ consistently. As shown in Table 6, IRA suppresses the sink ratio and improves overall model performance.

Table 6: Performance and attention sink ratio analysis.

Method	Acc. ↑	Sink Ratio ↓
InternVL2.5-Pt	48.8	96.9%
+ SFT	53.8	46.9%
+ IRA	**54.4**	**40.6%**

5 Conclusion

In this work, we address the problem of visual representation learning in vision–language models and identify the lack of explicit control over visual embeddings as a central cause of low interpretability, robustness, and catastrophic forgetting. We introduce Information-Regularized Attention (IRA), a stochastic attention mechanism that injects data-dependent noise into hidden states, enabling adaptive and principled regularization during end-to-end training. Through extensive analysis, we show that IRA not only improves downstream performance but also fundamentally reshapes the geometry of learned representations, yielding smoother curvature trajectories and mitigating attention sink behavior. These findings suggest that adaptively regularizing internal representations is critical for stabilizing multimodal reasoning. More broadly, our results highlight that attention should be viewed not merely as a weighting mechanism but as a metric for evaluating the internal representation. We hope this perspective opens new avenues for designing robust and interpretable multimodal systems.

Limitation Due to resource constraints, we apply the proposed methods to models up to 8B parameters, but we expect the conclusions to hold for larger models with more parameters, such as 13B, 26B, and 73B. Additionally, we believe that an IRA can serve as a general architecture for robust representation learning during the pre-training stage of LLMs and VLMs. We leave this to future study.

Acknowledgements

This research was supported in part by the Meta AI, the National Science Foundation under Grant 2502050, and the National Institutes of Health under Award R16GM159146. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems (2022)
3. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2017)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Barbero, F., Arroyo, A., Gu, X., Perivolaropoulos, C., Bronstein, M., Veličković, P., Pascanu, R.: Why do llms attend to the first token? arXiv preprint arXiv:2504.02732 (2025)

6. Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., Steinhardt, J.: Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112 (2023)
7. Bowman, S., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of the 20th SIGNLL conference on computational natural language learning. pp. 10–21 (2016)
8. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., Zhao, F.: Are we on the right way for evaluating large vision-language models? In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
9. Chen, S., Hou, Y., Cui, Y., Che, W., Liu, T., Yu, X.: Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
10. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences (2024)
11. Dong, X., Luu, A.T., Lin, M., Yan, S., Zhang, H.: How should pre-trained language models be fine-tuned towards adversarial robustness? Advances in Neural Information Processing Systems (2021)
12. Du, M., Wu, B., Li, Z., Huang, X.J., Wei, Z.: Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (2024)
13. Dziugaite, G.K., Roy, D.M.: Data-dependent pac-bayes priors via differential privacy. Advances in neural information processing systems **31** (2018)
14. Fan, X., Zhang, S., Chen, B., Zhou, M.: Bayesian attention modules. Advances in Neural Information Processing Systems (2020)
15. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. ArXiv (2023)
16. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision (2024)
17. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J., et al.: Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision (2022)
18. Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., Lin, M.: When attention sink emerges in language models: An empirical view. In: ICLR (2025)
19. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
20. Hénaff, O.J., Goris, R.L., Simoncelli, E.P.: Perceptual straightening of natural videos. Nature neuroscience **22**(6), 984–991 (2019)
21. Hong, J.H., Kim, H.J., Jeon, K.S., Lee, S.W.: Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)

22. Hosseini, E.A., Fedorenko, E.: Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=h3lTrt4Ftb
23. Jiang, J., Liu, Z., Zheng, N.: Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering. International Journal of Computer Vision (2024)
24. Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., Yang, X.: Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
25. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. arXiv preprint arXiv:2503.03321 (2025)
26. Li, B., Ni, J., Qu, C., Miao, I., Yang, L., Fu, X., Chen, M., Cheng, D.Z.: Reinforced attention learning. arXiv preprint arXiv:2602.04884 (2026)
27. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. PMLR (2023)
29. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
30. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
31. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023)
32. Li, Z., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, Z., Zhao, H.: Data-dependent gaussian prior objective for language generation. In: International Conference on Learning Representations (2020)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems (2023)
34. de Llano, E.Q., Arroyo, A., Barbero, F., Dong, X., Bronstein, M.M., LeCun, Y., Shwartz-Ziv, R.: Attention sinks and compression valleys in LLMs are two sides of the same coin. In: The Fourteenth International Conference on Learning Representations (2026)
35. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems (2019)
36. Mahajan, S., Le, H., Park, H., Farhadzadeh, F., Hayat, M., Porikli, F.: Attention guided alignment in efficient vision-language models. arXiv preprint arXiv:2511.17793 (2025)
37. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (2019)
38. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2022 (2022)

39. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (2021)
40. Noh, H., You, T., Mun, J., Han, B.: Regularizing deep neural networks by noise: Its interpretation and optimization. Advances in neural information processing systems **30** (2017)
41. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems (2022)
42. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2025)
43. Poole, B., Sohl-Dickstein, J., Ganguli, S.: Analyzing noise in autoencoders and deep networks. arXiv preprint arXiv:1406.1831 (2014)
44. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., Lin, J.: Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
45. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems (2023)
46. Raghu, M., Gilmer, J., Yosinski, J., Sohl-Dickstein, J.: Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. Advances in neural information processing systems **30** (2017)
47. Rahmanzadehgervi, P., Bolton, L., Taesiri, M.R., Nguyen, A.T.: Vision language models are blind: Failing to translate detailed visual features into words. arXiv preprint arXiv:2407.06581 (2024)
48. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018)
49. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
50. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J.M., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. ArXiv (2024)
51. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
52. Skean, O., Arefin, M.R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., Shwartz-Ziv, R.: Layer by layer: Uncovering hidden representations in language models. arXiv preprint arXiv:2502.02013 (2025)
53. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. Advances in neural information processing systems (2016)
54. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. (2014)
55. Steinberg, J., Gal, O.: Where vision becomes text: Locating the ocr routing bottleneck in vision-language models. arXiv preprint arXiv:2602.22918 (2026)

56. Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., RABBANI, M., Rao, R., Tao, Z.: Latent chain-of-thought for visual reasoning. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems (2025)
57. Sun, G., Qin, C., Feng, Y., Chen, Z., Xu, R., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Structured policy optimization: Enhance large vision-language model via self-referenced dialogue. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
58. Sun, G., Qin, C., Fu, H., Wang, L., Tao, Z.: Self-training large language and vision assistant for medical question answering. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Nov 2024)
59. Sun, G., Qin, C., Wang, J., Chen, Z., Xu, R., Tao, Z.: Sq-llava: Self-questioning for large vision-language assistant. In: European Conference on Computer Vision. pp. 156–172 (2024)
60. Sun, G., Wang, Y., Ma, S., Xie, Y., Cheng, Y., Tao, Z., Wang, J.: If-prune: Information-flow guided token pruning for efficient vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3522–3531 (June 2026)
61. Sun, M., Chen, X., Kolter, J.Z., Liu, Z.: Massive activations in large language models. arXiv preprint arXiv:2402.17762 (2024)
62. Sun, S., Canziani, A., LeCun, Y., Zhu, J.: The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. arXiv preprint arXiv:2603.05498 (2026)
63. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
64. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc. (2017)
65. Vo, A., Nguyen, K.N., Taesiri, M.R., Dang, V.T., Nguyen, A.T., Kim, D.: Vision language models are biased. arXiv preprint arXiv:2505.23941 (2025)
66. Voita, E., Sennrich, R., Titov, I.: The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)
67. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., Yan, T., Mo, W.J., Liu, H.H., Lu, P., Li, C., Xiao, C., Chang, K.W., Roth, D., Zhang, S., Poon, H., Chen, M.: Muirbench: A comprehensive benchmark for robust multi-image understanding. ArXiv (2024)
68. Witten, E.: A mini-introduction to information theory. La Rivista del Nuovo Cimento **43**(4), 187–227 (2020)
69. Wu, J., Xiong, Y., Li, X., Xia, Y., Wang, R., Wang, Y., Yu, T., Kim, S., Rossi, R.A., Yao, L., Shang, J., McAuley, J.: Mitigating visual knowledge forgetting in MLLM instruction-tuning via modality-decoupled gradient descent. In: Findings of the Association for Computational Linguistics: EMNLP 2025 (Nov 2025)
70. x.ai: Grok-1.5 vision preview. Tech. rep., x.ai (2024)
71. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and efficient post-training quantization for large language models. In: International conference on machine learning. pp. 38087–38099. PMLR (2023)

72. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453 (2023)
73. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
74. Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., Tong, S., Sun, Y., Yu, B., Zhang, G., Sun, H., et al.: Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15134–15186 (2025)
75. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In: Conference on Parsimony and Learning (Proceedings Track) (2023)
76. Zhang, F., Wu, Y., Wang, Z., Wang, X., Lv, C., Huang, X., Zheng, X.: Vib-probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck. arXiv preprint arXiv:2601.05547 (2026)
77. Zhang, Y.F., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? arXiv preprint arXiv:2408.13257 (2024)
78. Zhao, J., Zhang, F., Sun, X., Feng, C.: Mitigating hallucination in large vision-language models through aligning attention distribution to information flow. In: Findings of the Association for Computational Linguistics: EMNLP 2025 (2025)