

When Images Speak Louder: Mitigating Language Bias-induced Hallucinations in VLMs through Cross-Modal Guidance

Jinjin Cao Zhiyang Chen* Zijun Wang Liyuan Ma Weijian Luo Guojun Qi*

MAPLE Lab, Westlake University

caojinjin@westlake.edu.cn, chenzyang@westlake.edu.cn, guojunq@gmail.com

ABSTRACT

Vision-Language Models (VLMs) have shown solid ability for multimodal understanding of both visual and language contexts. However, existing VLMs often face severe challenges of hallucinations, meaning that VLMs tend to generate responses that are only fluent in the language but irrelevant to images in previous contexts. To address this issue, we analyze how language bias contributes to hallucinations and then introduce Cross-Modal Guidance (CMG), a training-free decoding method that addresses the hallucinations by leveraging the difference between the output distributions of the original model and the one with degraded visual-language attention. In practice, we adaptively mask the attention weight of the most influential image tokens in selected transformer layers to corrupt the visual-language perception as a concrete type of degradation. Such a degradation-induced decoding emphasizes the perception of visual contexts and therefore significantly reduces language bias without harming the ability of VLMs. In experiment sections, we conduct comprehensive studies. All results demonstrate the superior advantages of CMG with neither additional conditions nor training costs. We also quantitatively show CMG can improve different VLM’s performance on hallucination-specific benchmarks and generalize effectively.

1 INTRODUCTION

Vision-Language Models (VLMs) like GPT-4o[24], LLaVA-Series[11, 20, 35], QwenVL-Series[1, 2, 30], and others [3, 5, 7, 9, 17, 29, 31, 33], have shown solid abilities in multi-modal information perception and reasoning, sparking a new wave of applications of modern artificial intelligence. Despite powerful capacities, many recent researchers have found that VLMs sometimes suffer from hallucinations: *VLMs often tend to generate incorrect responses that are irrelevant to image inputs in previous contexts*. For instance, as Figure 1(a) shows, when we ask the VLM to count the apples in the image, the answer seems to depend more on whether the world *images* is singular or plural, rather than the factual visual content. Even though we provide no images, VLMs can still generate a plausible answer. This phenomenon may be denoted as language bias, in that the model generates responses following the learned language pattern and ignoring visual information.

Such a property brings uncontrollable risks for users when using VLMs, preventing the broader use of VLMs. To mitigate this issue, some existing works have proposed various solutions, such as prompt-engineering-based methods [14], post-training using human feedback data [25], and developing different inference strategies [16, 26].

In this paper, we focus on the inference of VLMs. We introduce *Cross-Modal Guidance (CMG)*, a training-free inference algorithm

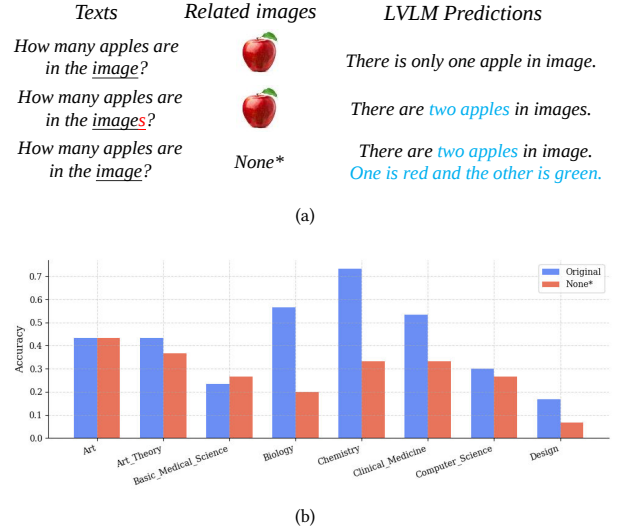


Figure 1: An illustration of hallucinations induced by language bias. (a) Examples of hallucinations induced by language bias in VLMs. The blue words are hallucination contents. (b) Accuracy in MMMU[32] Benchmark on LLaVA-v1.5-7B. 'None*' represents images that are removed from the visual question input.

for vision-language models that can significantly reduce hallucinations. CMG first corrupts the visual-language attention by randomly masking attention weights in certain transformer-based VLM decoder layers. Then CMG computes the inference logit values by adding a scaled difference term of the output logits using original and masked attention weights. Such an attention-corruption mechanism enhances the visual-language perceptions inside the neural networks, distinguishing CMG from previous methods such as VCD[16] that directly add Gaussian noises on input images. Besides, CMG is different from other previous methods such as ConVis[26] that call expensive additional models to enhance the visual information. From a high level of view, the CMG *make images speak louder* inside the neural network, therefore can address the insufficient visual perception of VLMs that potentially cause hallucinations.

In section 4, we find that CMG can improve the generation performances of VLM with a significant effect of reducing hallucinations. CMG also outperforming its counterparts with VCD and ConVis in POPE and HallusionBench benchmark. On the MME benchmark, CMG surpass VCD by a large margin, reaching 13.54%

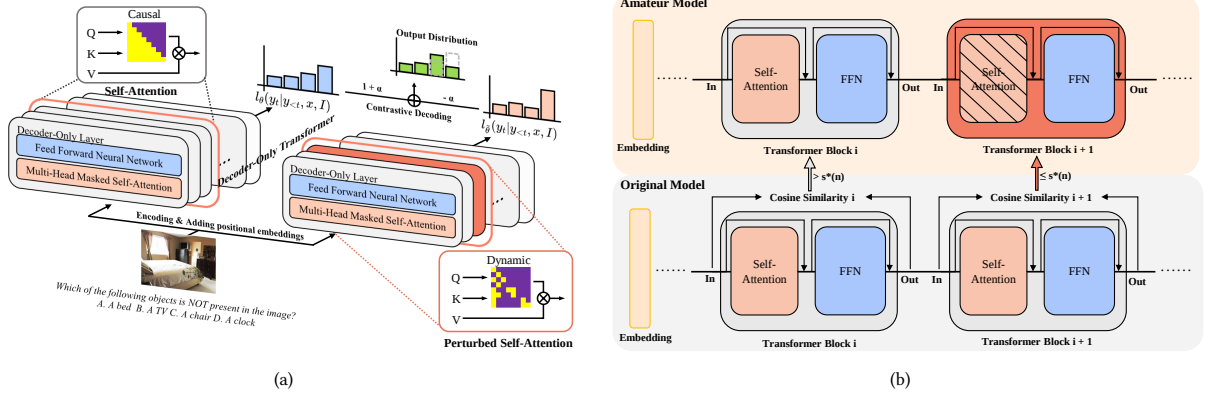


Figure 2: Architecture of Cross-Modal Guidance. CMG utilizes a perturbed self-attention map to amplify language priors in the underlying decoder-only transformer backbone. The original self-attention uses a causal mask, while the perturbed self-attention map replaces it with a dynamic mask, which varies from different samples. Perturbed self-attention is applied to several dynamically selected decoder-only layers. CMG contrasts the two distributions to correct hallucinations from the original outputs.

performance gain and also exceeds ConVis by **+8.0**. On the POPE benchmark, LLaVA-v1.5-7B with CMG achieves an overall accuracy value of **85.48**, outperforming its counterparts with VCD and ConVis. On HallusionBench, CMG exceeds VCD by **+7.1** and ConVis by **+6.3** in accuracy with no additional training, marking a leading performance among training-free inference approaches.

Our contributions are summarized as follows:

- We introduce CMG, a training-free inference method that can effectively reduce models’ hallucinations by enhancing the visual-language perceptions through random attention masks;
- We identify the insufficient visual-attention connections as one of the causes of hallucinations with rigorous evidence;
- We quantitatively evaluate CMG and show its solid performances on multiple benchmarks.

2 RELATED WORK

2.1 Hallucinations in Vision-Language Models

Vision-Language Models (VLMs) [2, 3, 12, 20, 23, 24, 34, 36] have revolutionized based on the development of Large Language Models (LLMs). VLMs can receive both visual and textual input, generating text responses iteratively. Specifically, to process image inputs, VLMs use an image encoder and linear projections to align text and image embeddings; for instance, LLaVA-v1.5 [20] uses CLIP [27] as its image encoder. However, despite the powerful abilities, misalignment emerges in VLMs. Hallucination generally refers to cases where generated responses include information unrelated to image content. Some benchmarks [6, 10, 19] are collected to evaluate hallucinations.

2.2 Content-Aware Decoding

Proper decoding (inference) methods are essential for both large language models and vision-language models to get optimal performances. From a high level of view, decoding methods can generally be categorized into search and sampling algorithms [18].

Search methods, like greedy and beam search, produce accurate results but often lead to tedious and repetitive outputs. In contrast, sampling methods, such as nucleus sampling [15], generate more diverse text but can suffer from unnatural topic shifts. To address these issues, content-aware decoding [18, 28] was proposed for large language models, leveraging the difference between two output probabilities to construct a new and potentially enhanced output distribution. Similar ideas arose in the literature of vision-language models in recent years. VCD [16] contrasts distribution with original and distorted image inputs to reduce statistic bias and language priors in LVMs. Also focusing on image input, ConVis [26] utilizes an additional text-to-image model to regenerate the caption of the original image, and then uses the difference in details between the new image and the original image to guide the generation. However, these methods only distort image input, leaving in-depth research on the black-box nature of VLMs. In this paper, we focus on the transformer attention mechanism, which has not been studied in previous research. We found language bias induces the emergence of hallucinations. By destroying the modal attention connection between image and text to contrast with the original distribution, we strengthen the reliance on visual context and eliminate the influence of language bias.

3 THE PROPOSED METHOD

3.1 Preliminaries

A Vision-Language Model (VLM) parameterized by θ with an autoregressive, autoencoding or encoder-decoder architecture pretrained on a large corpus of millions to trillions of tokens. VLMs are usually

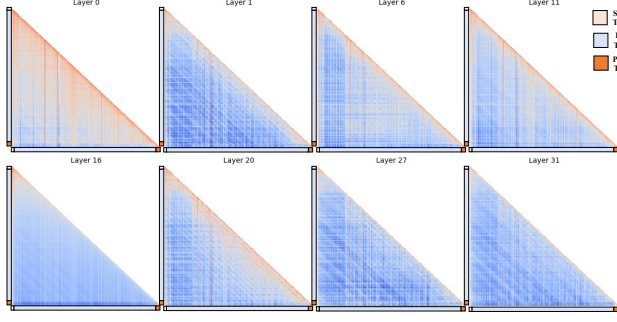


Figure 3: Visualization of Attention Weights Changes Across Transformer Layers. The overall trend of image token weight ratio is getting lower as the number of transformer layers increases.

adapted for a specific task, for example, image captioning or vision-question answering, by fine-tuning in a relatively small dataset compared to pretraining.

VLMs receive interleaved images(or videos) and texts as input, generates coherent and fluent texts as answers. Specifically, we consider text input of length n , denoted as $x = \{x_1, x_2, \dots, x_n\}$ and visual input of length m , denoted as $I = \{I_1, I_2, \dots, I_m\}$. After decoding, we acquire a text sequence of length k denoted as $y = \{y_1, y_2, \dots, y_k\}$.

The text output y is generated auto-regressively by the underlying language model p_θ . During decoding, tokens are generated iteratively, each conditioned on the preceding context:

$$p_\theta(y|x, I) = \prod_{t=1}^k p_\theta(y_t|y_{<t}, x, I) \quad (1)$$

where $p_\theta(y_t|y_{<t}, x, I)$ denotes the next token distribution. We use different subscripts to denote different weights of language model: p_θ is the original VLM, $\tilde{p}_\theta$ is the amateur VLM where the model result is less accurate than the original.

In detail, the relationship between the output distribution and the direct output of logits is:

$$p_\theta(y|x, I) = \text{softmax}[l_\theta(y|x, I)], \quad (2)$$

where $l_\theta(y|x, I)$ denotes as the logits output of language model.

3.2 Language Bias Raises Hallucinations in VLMs

Language bias refers to answers being strongly biased towards textual part of input questions while the importance of visual part is overlooked. This bias strongly influences the responses of VLMs, leading to a preference for content closely related to the language pretraining data, while being weakly or even completely unrelated to the current visual input. In fig. 1(a), we show two typical cases of hallucination in VLMs. By simply replacing “image” in the question with “images”, VLM outputs two completely different answers under the same image input conditions. The other case is VLM responds to text prompt when there is no image input at all. In fig. 1(b), compared to the baseline, when completely deleting image input, the performance on the MMMU Benchmark does not degenerate to completely inaccurate. Different degrading scores at subsets implies

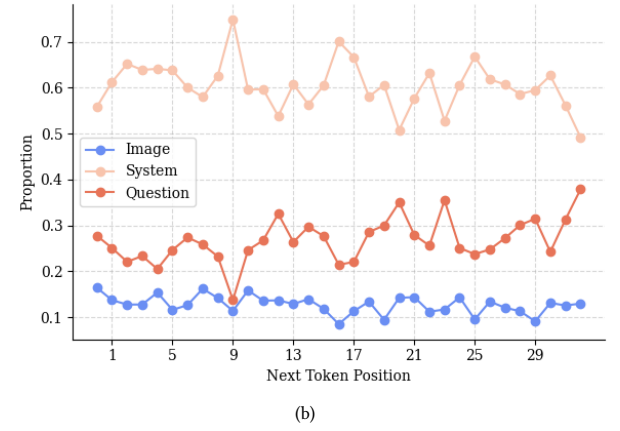
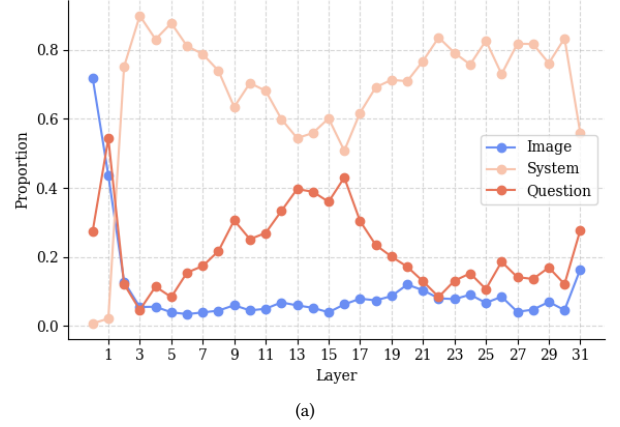


Figure 4: Variation in Attention Weight Proportions Across Token Sequence Parts. (a) The proportion of image attention weights changes with transformer layer. (b) The proportion of image attention weights changes with generated token sequence lengths.

that samples are affected by language bias in separate degrees. The scores on some subtasks like Art are greater than random selection probability, which indicates that the baseline’s answers are affected by inherent language bias.

However, the image tokens usually outnumber the text tokens in input sequence of VLMs. Taking LLaVA-v1.5-7B as an example, the image tokens are typically encoded into a sequence of 576 tokens, while the text token length is about one-tenth of the image length. So why the impact of language bias on the output can sometimes outweigh the influence of the images? As shown in fig. 3, we discover image attention weights drop sharply in shallow layers, maintaining low weights in deeper layers. fig. 4(a) clearly shows the image attention weights decay sharply in the first few shallow layers, only rising slightly in the last transformer layers. In contrast, in the shallow layer, the ratio of text attention weights increases significantly, especially the ratio of system tokens even exceeds that of question tokens containing key information to answers. This implies the role of image token is largely overlooked

than text tokens as transformer layer goes deeper, which induces hallucination in VLMs.

We further investigate the change in attention weights as the generated token sequence grows. As shown in fig. 4(b), the proportion of image attention weights also decreases gradually, while text tokens maintain a high proportion. This phenomenon can explain why hallucinations are more likely to occur when generating long contexts.

Findings in language bias in VLMs highlight that language bias contributes to hallucinations in VLMs, and thus we ought to mitigate it by enhancing model’s attention on images.

3.3 Cross-Modal Guidance

3.3.1 Constructing Amateur Model with Attention Mask. As shown in fig. 5, the self attention A in transformer blocks consists of three types: inter-visual attention A_{iv} , inter-textual attention A_{it} , and cross-modal attention A_{cr} .

$$A = A_{iv} \cup A_{it} \cup A_{cr} \quad (3)$$

As we discussed above, in order to mitigate language bias, we need to enhance both inter-visual attention and cross-modal attention to make better use of visual contents.

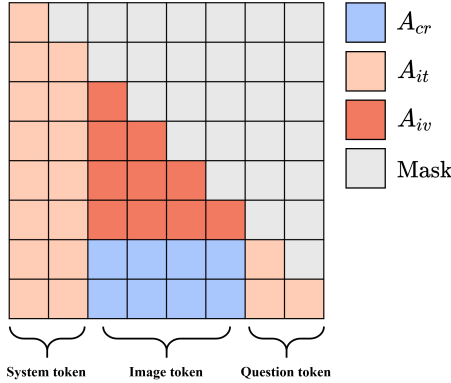


Figure 5: Self Attention Weights with Causal Mask

If we retain only inter-textual attention A_{it} , VLMs degrade to a model that is forced to generate the output distributions solely biased towards language questions. By comparing this biased output distributions with original ones, we form the pointwise mutual information (PMI) between the final output y and visual-related attentions A_{cr} and A_{iv} similar to those in CFG and contrastive decoding [28], and it can be used to adjust the original output distributions of VLMs,

$$\tilde{p}_\theta(y|x, I) \propto p_\theta(y|x, I) \left(\frac{p_\theta(y|x, I; A_{cr}, A_{iv}, A_{it})}{p_\theta(y|x, I; \emptyset, \emptyset, A_{it})} \right)^\alpha \quad (4)$$

where $p_\theta(y|x, I; \emptyset, \emptyset, A_{it})$ denotes both cross-modal attention and inter-visual attention are completely masked out. This yields a preference over an output y that is more likely to be generated with these visual-related attentions rather than without them.

However, as shown in table 6, we find simply removing all visual-related attentions results in poor p_θ that fails to generate correct answers in VLMs. We hypothesize that it would cause the collapse

of the underlying VLM network as its attention structure was over-disturbed. Usually, a constraint shall be imposed on how much disturbance ought to be allowed. This inspires us to only remove part of cross-modal and inter-visual attention weights using masks, and we set a maximum size for these attention masks to control how many attention weights can be removed.

By removing part of cross-modal and inter-visual attention, we strengthen the role of inter-textual attention that would enhance the language bias in VLMs. In this case, if the introduction of cross-modal and inter-visual attention leads to an increase in the probability of a responding word, we may deduce that this word should be highly related to the visual content. Thus we should favor these words during sampling, which can be achieved by adjusting the output distribution with a similar PMI ratio as in Eq. 4.

Formally, we adjust the original VLM’s output distribution $p_\theta(y|x, I; A_{cr}, A_{iv}, A_{it})$ to obtain a new one $\tilde{p}_\theta(y|x, I)$ as

$$\tilde{p}_\theta(y|x, I) \propto q_\theta(y) \left(\frac{q_\theta(y)}{q_\theta(y; M)} \right)^\alpha \quad (5)$$

where M denotes the mask imposed on the attention map, that is

$$M := M_{cr} \cup M_{iv} \quad (6)$$

$$q_\theta(y) := p_\theta(y|x, I; A_{cr}, A_{iv}, A_{it}) \quad (7)$$

$$q_\theta(y; M) := \tilde{p}_\theta(y|x, I; A_{cr} \odot M_{cr}, A_{iv} \odot M_{iv}, A_{it}) \quad (8)$$

Here M_{cr} and M_{iv} are masks on cross-modal attention and inter-visual attention respectively, and we denote the masked VLM model by *Amateur Model*. These masks are applied to self-attentions in Eq. 10 below

$$SA(Q, K, V; M) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} \odot M\right)V \quad (9)$$

$$= \text{Softmax}(A \odot M)V. \quad (10)$$

3.3.2 Dynamically Masking Amateur Models. Finding an optimal mask M subject to a maximum size n_0 can be formulated by maximizing the divergence between $q_\theta(y)$ and $q_\theta(y; M)$,

$$\max_M \text{KL}[q_\theta(y; M), q_\theta(y)] \quad (11)$$

$$\text{s.t. } \|1 - M_{cr}\|_0 + \|1 - M_{iv}\|_0 \leq n_0 \quad (12)$$

$$M_{cr} \cup M_{iv} \in \{0, 1\}^N \quad (13)$$

where KL is the KL divergence, $\|\cdot\|_0$ is the ℓ_0 -norm that accounts the number of non-zero elements, and N is the total number of candidate positions to mask in a VLM model. By maximizing the divergence, we will obtain a masked model that maximizes the contrast with the original model. In this way, the ratio of $q_\theta(y)$ and $q_\theta(y; M)$ is more likely to increase¹ if the likelihood of generating an output y is more likely after the masked inter-visual and cross-modal attentions are filled back to the model. This strengthens the role of these visual-related attentions, which could mitigate the language bias often related with text-only attentions.

Unfortunately, directly optimizing the above constrained objective is intractable as it is a NP-hard problem. We provide two dynamic strategies to determine which attention weights to mask in some selected layers. The basic idea is to find the part of attention weights making the largest contribution to the output of the

¹We cannot directly maximize the ratio of $q_\theta(y)$ and $q_\theta(y; M)$ in Eq. 5 since the groundtruth output y is unknown beforehand in the inference.

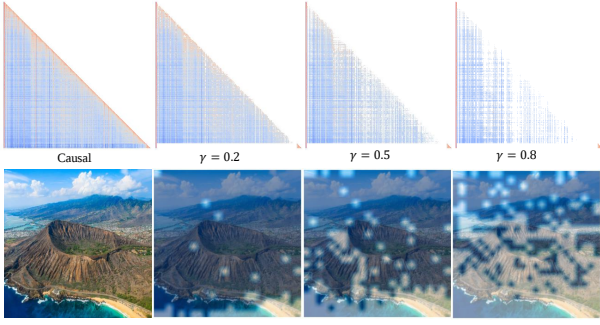


Figure 6: Inter-visual attention masks with different γ . The masks are visualized by an average of mask values associated with a pixel patch. The asked question for this example in the VLM model is *Describe this photo in detail*.

original VLM model, by removing which its output distribution $q_\theta(y)$ could be greatly changed.



Figure 7: Attention Mask When $\gamma=0.5$. The whiter part is the masked part.

Dynamic attention masking. Attention weights determine the level of importance each element contributes to the model’s output. We partially mask the largest γ -portion of attention weights in A_{iv} and A_{cr} , resulting in

$$\widetilde{SA}(Q, K, V; M) = \text{Softmax}(A \odot M(\gamma))V \quad (14)$$

$$M(\gamma) = A_{cr}(\gamma) \cup A_{iv}(\gamma). \quad (15)$$

As shown in fig. 6, masking positions are usually related to objects in an image, which would make the answer to the asked question of “Describe this photo in detail” more related to the relevant visual information in the image. As γ increases, the masks would become more selective in retaining visual parts in answering the question. fig. 7 provides more examples to illustrate the relationship between the masking position and the question.

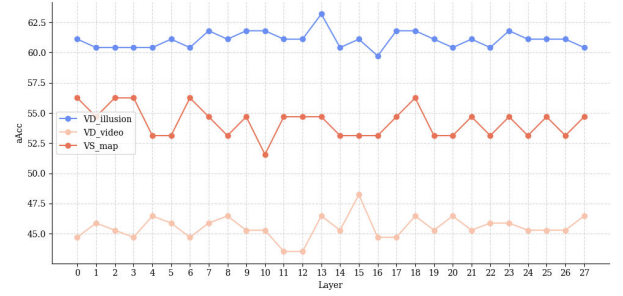


Figure 8: All accuracy score(aAcc) in HallusionBench when masking different transformer layers in Qwen-2-VL-2B-Instruct[30]. Each time only one layer is masked.

Dynamic layer selection. We also observe that different transformer blocks contribute unequally to the output distribution. As shown in fig. 8, when only a single layer is selected to apply the above dynamic attention masking approach, we find that the optimal layer index varies across subsets in terms of accuracy scores.

This suggests that different layers have different effects on the result. We must carefully decide which layers to mask the attention weights by adopting a dynamic layer selection strategy. Formally, we determine which layers need to be selected by calculating the cosine similarities between the layer input X and output distribution Y . We only choose those layers whose cosine similarity is sufficiently small, i.e., the layer output changes a lot from its input. In other word, if a layer changes its inputs a lot to give its outputs, it is considered playing an important role in the model. So selecting to mask its attentions could more significantly disturb the original VLM model.

Formally, given the number of n layers, the dynamic layer selection can be defined as:

$$s(i) = \cos(X_i, Y_i) = \frac{X_i \cdot Y_i}{\|X_i\|_2 \|Y_i\|_2} \quad (16)$$

$$s^* = \text{AscentSort}(\{s(i)|i = 1, \dots, n\})[\tau \cdot n] \quad (17)$$

$$\mathbb{Z} = \{i|s(i) \leq s^*\} \quad (18)$$

where τ denotes the proportion of layers that shall be selected, s^* is the cosine similarity at the smallest τ percentile, and $\mathbb{Z}$ denotes the index set of selected layers.

After a layer is selected, the dynamic attention masking is applied to this layer. fig. 2 shows the full Cross-Modal Guidance method to construct an amateur model.

Why we choose cosine similarity to determine layer importance? Cosine similarity is a commonly used way to measure similarity between two vectors. Dot product and Euclidean distance also can be used to measure similarity between the output and input of a layer. [4, 21] suggest that the magnitude of hidden states in transformers tend to grow as the layer becomes deeper. The dot product and Euclidean distance are both influenced by the vector magnitude, which means that they also changes with the layer. Consequently, we adopt cosine similarity, which only depends on the vector direction.

4 EXPERIMENTS

4.1 Experimental Setup

Benchmarks. To prove effectiveness of our method on mitigating language bias in LVLMs, we conduct experiments on three benchmarks. They are:

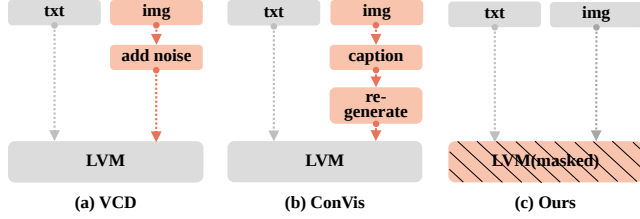


Figure 9: Comparison chart of VCD, ConVis and Ours

- Hallucination-related: HallusionBench[13], POPE[19]
- Comprehensive: MME[10]

Compared Methods. Current researches have designed various methods to construct an amateur model. As fig. 1(b) shows, VCD[16] adds Gaussian noise to original images as the amateur’s visual input. ConVis[26] captions the original image transforms the original image input to caption prompts, and utilize text-to-image model to generate a new image based on captions, and utilize the difference between the original and re-generated image input. However, in these methods only the input of the model is focused, without in-depth research on the model decoding mechanism. The amateur input obtained by distorting the visual input not only changes the entire image distribution, but also becomes uncontrollable in the deep network of the model. Our method use the attention mask to dynamically corrupt amateur model’s performance, fully considering features of different types of samples.

Implementation Details. We employ LLaVA-v1.5-7B [20], Instructblip-7B[8], Qwen2-VL-7b[30] and InternVL2.5-8b[22] as our backbone model, using publicly available checkpoint weights. We set the top parameter to 0.9, beam search parameter to 5, temperature to 0.7 in our baseline. We set $\alpha=0.3$, $\gamma=0.5$, $\tau=0.5$ for hallucination-specific benchmarks, and $\alpha=0.1$, $\gamma=0.5$, $\tau=0.1$ for general benchmark MME. For both VCD and ConVis, the parameters employed are consistent with the optimal parameters provided in their respective papers.

4.2 Results

Results on POPE benchmark. The POPE benchmark evaluates object hallucinations by prompting VLMs to answer "yes" or "no" to questions regarding the existence of objects. The results of the POPE experiments, including recall, accuracy, precision, and overall. As demonstrated in table 1, while conventional methods struggle with architecture upgrades (ConVis and VCD both dropping Overall Score on InternVL-2.5 versus its 89.0 baseline), our approach shows positive scaling (89.0→89.3 Overall) across model generations. This upward-compatible performance confirms our method’s effectiveness as a universal solution for visual-language alignment. In contrast to VCD and ConVis, which perform well only on older versions, our method achieves state-of-the-art performance across multiple backbones.

Results on HallusionBench benchmark. HallusionBench is specific on hallucination in VLMs. It divides vision-question pairs into two types. The visual dependent questions relies on heavily on provided images, while the visual supplement questions can be answered without images referenced. The 1129 questions consists of diverse topics and formats, with binary choices of yes or no. HallusionBench encompasses a broader range of hallucination types compared to the POPE benchmark, that is limited to assessing object hallucinations. In table 1, our method outperforms all listed baseline in both figure accuracy and overall accuracy. Question pair accuracy calculates the proportion of correctly answered questions when the picture is missing, figure accuracy measures the proportion of correctly answered images within the dataset, while overall accuracy represents the percentage of correctly answered questions. Specifically, we achieve a significant improvement in figure accuracy, demonstrating our enhanced capability to mitigate hallucinations in image understanding.

Results on MME benchmark. MME is a comprehensive benchmark measuring both the perception and reasoning abilities of VLMs. It consists of 14 subtasks which are divided into perception and reasoning categories. In perception category, MME evaluates coarse-grained recognition, fine-grained recognition and ocr abilities. The score for each subtask is the sum of the accuracy and accuracy+, where the latter refers to the score based on each image where all questions need to be answered correctly.

As shown in table 2, our method surpasses other decoding approaches in perception-related subtasks, which reflects fine-grained image understanding abilities. The scores presented in the table are the sum of accuracy and accuracy+. The accuracy+ metric is calculated at the image level, in other words, when all questions related to a single image are answered correctly, the accuracy+ for that image equals 100.

Our total score in the perception domain outperforms baseline, exceeding VCD by +62.08 and ConVis by +7.30 on average. Notably, our method achieves the highest scores in the "color," "scene", "landmark," subsets, where hallucinations caused by language bias are particularly prevalent. In other subsets, our method also demonstrates competitive performance. The results on the MME benchmark demonstrate that our model continues to exhibit outstanding performance on general tasks.

Results on Different Size of Models. table 3 and table 4 both show our model achieves robust performance across varying model scales, including 2B, 7B, 13B, and 26B parameter configurations, highlighting its scalability and architectural adaptability.

Results on the Method’s Cumulative Effects. table 5 shows our method can be combined with both VCD and ConVis, achieving superior performance compared to using VCD or ConVis individually. However, this synergistic effect is not consistently observed, likely due to divergent optimization objectives among the different methods.

Table 1: Evaluation results on datasets designed for hallucinations

Model	Method	HallusionBench			POPE			
		Question Pair	Acc(qAcc)	Figure Acc(fAcc)	All Acc(aAcc)	Recall	Accuracy	Precision
LLaVA-v1.5 7b	Baseline	11.4	16.2	46.1	78.9	85.9	91.8	84.9
	VCD	11.6	16.8	45.8	78.7	85.7	91.5	84.6
	ConVis	11.2	15.9	45.3	78.8	85.9	91.8	84.8
	Ours	11.9	16.5	46.2	79.3	86.3	91.7	85.4
InstructBlip 7b	Baseline	17.1	20.9	51.1	77.1	85.2	92.1	83.9
	VCD	17.1	20.9	51.0	76.67	84.93	91.98	83.63
	ConVis	19.1	22.6	53.7	82.5	85.4	87.6	85.0
	Ours	19.4	23.7	54.0	82.9	86.1	88.6	85.7
Qwen2-VL 7b	Baseline	43.7	39.0	68.5	78.3	87.4	95.7	86.1
	VCD	40.7	34.7	65.6	81.7	88.6	94.7	87.7
	ConVis	43.5	37.9	67.8	78.5	87.5	95.7	86.2
	Ours	45.5	39.0	68.5	80.2	88.2	95.4	87.1
InternVL2.5 8b	Baseline	41.1	43.1	68.1	84.2	89.6	94.3	89.0
	VCD	39.3	37.6	65.2	83.5	88.7	93.1	88.1
	ConVis	39.3	38.0	65.4	83.1	88.6	93.4	88.0
	Ours	41.1	36.7	68.8	85.3	89.7	93.7	89.3

Table 2: Evaluation results on the MME benchmark.

Model	Method	Perception										Total
		existence	count	position	color	ocr	poster	celebrity	scene	landmark	artwork	
LLaVA-v1.5 7b	Baseline	185.00	93.33	113.33	160.00	117.50	90.81	90.29	142.00	136.50	114.00	1242.78
	VCD	195.00	153.33	116.67	160.00	140.00	137.76	133.24	156.75	157.00	118.75	1468.49
	ConVis	195.00	158.30	133.30	155.00	132.50	143.20	139.70	153.80	155.30	121.50	1487.60
	Ours	190.00	158.33	126.67	160.00	147.50	142.52	134.12	154.50	158.00	124.00	1495.63
InstructBlip 7b	Baseline	180	65.00	55.00	138.33	95.00	129.59	153.53	158.25	101.00	130.00	1205.70
	VCD	180.00	70.00	60.00	143.33	95.00	134.01	158.24	143.5	104.75	128.00	1216.83
	ConVis	180.00	65.00	55.00	143.33	95.00	128.57	153.83	159.75	114.50	130.00	1224.98
	Ours	180.00	60.00	53.33	143.33	95.00	126.87	161.47	158.00	99.50	132.25	1209.76
Qwen2-VL 7b	Baseline	185.00	128.33	161.67	178.33	177.50	164.29	125.00	144.75	154.25	131.75	1550.86
	VCD	195.00	138.33	163.33	185.00	125.00	182.99	136.77	168.25	171.50	139.75	1605.92
	ConVis	195.00	160.00	160.00	180.00	155.00	182.65	151.18	162.75	185.75	148.25	1680.58
	Ours	190.00	155.00	165.00	185.00	170.00	185.37	149.41	160.75	184.25	149.75	1694.54
InternVL2.5 8b	Baseline	200.00	175.00	170.00	183.33	177.50	162.93	138.24	153.00	172.00	157.00	1688.99
	VCD	200.00	175.00	151.67	175.00	177.50	167.01	136.74	155.50	169.00	159.75	1666.89
	ConVis	200.00	175.00	158.33	175.00	177.50	165.65	140.88	154.75	172.00	160.50	1679.61
	Ours	195.00	160.00	165.00	185.00	170.00	185.37	148.53	160.75	185.00	147.50	1702.00

5 DISCUSSIONS

5.1 Case Study

fig. 10(a) showcases a painting comprehension task, highlighting the necessity for precise instruction adherence, accurate image interpretation, and pre-trained knowledge. The original model mistakenly associates “hat” with characters’ attire due to skewed co-occurrence biases, a misjudgment exacerbated by the attention-masked model. CMG effectively rectifies this, diminishing the “hat” confidence and accurately identifying “bandana” as the correct choice.

In fig. 10(b), a query about players’ T-shirt colors presents a challenge, with black caps potentially misleading VLMs. The original model incorrectly favors option A, failing to discern pertinent visual details. CMG intervention adjusts the PMI ratio, elevating option C’s confidence and steering the model towards the accurate response.

Table 3: Evaluation results on the POPE with Different Size of Models.

Model	Size	Method	Recall	Accuracy	Precision	Overall
InstructBlip	13B	Baseline	35.3	60.5	95.7	51.6
		Ours	40.2	62.7	94.3	56.4
Internvl2.5	26B	Baseline	88.1	90.6	92.6	90.3
		Ours	85.7	90.6	94.9	90.1
Qwen2VL	2B	Baseline	82.3	89.0	95.0	88.2
		Ours	83.7	89.2	94.0	88.5
Qwen2.5VL	7B	Baseline	77.5	87.6	97.0	86.2
		Ours	80.3	88.5	96.0	87.4

Table 4: Evaluation results on the HallusionBench with Different Size of Models.

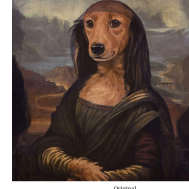
Model	Size	Method	qAcc	fAcc	aAcc
InstructBlip	13B	Baseline	22.9	16.2	49.9
		Ours	24.4	18.2	53.6
Internvl2.5	26B	Baseline	46.6	47.7	71.5
		Ours	47.9	45.1	72.1
Qwen2VL	2B	Baseline	32.5	28.9	60.7
		Ours	33.4	30.1	61.4
Qwen2.5VL	7B	Baseline	40.4	35.8	65.7
		Ours	47.0	45.7	70.5

Table 5: Additive effect evaluation of methods

Method		HallusionBench			POPE			
		fAcc	qAcc	aAcc	Recall	Accuracy	Precision	Overall
A	Baseline	43.7	39.0	68.5	78.3	87.4	95.7	86.1
B	Ours	45.5	39.0	68.5	80.2	88.2	95.4	87.1
C ₁	A + VCD	40.7	34.7	65.6	81.7	88.6	94.7	87.7
C ₂	B + VCD	47.2	47.0	68.1	82.4	88.8	94.5	88.0
D ₁	A + ConVis	43.5	37.9	67.8	78.5	87.5	95.7	86.2
D ₂	B + ConVis	39.0	45.3	67.4	81.3	88.3	94.5	87.4

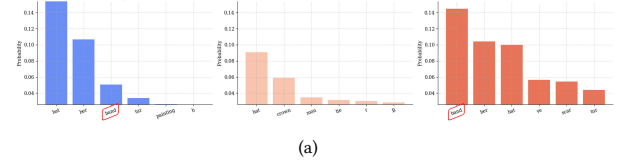
Table 6: Results on HallusionBench for Qwen-2-VL-2B. ‘None*’ refers to the ablated model that removes all vision-related attention mechanisms. ‘Noise*’ refers to the ablated model that replaces images with random noise. ‘Text-only*’ denotes the ablated model that converts image inputs into textual captions, substituting the original image with its description.

Method	fAcc	qAcc	aAcc
None*	12.14	16.70	53.84
Noise*	29.48	31.21	59.93
Text-only*	29.77	32.75	61.09
Ours	30.06	33.41	61.41



Prompt: Discuss how this creative twist on a classic work of art might be interpreted differently by various audiences.

Answer: The creative twist on a classic work of art, featuring a dog wearing a Renaissance-style dress and a bandana* best fits the image content)



What color is the shirt one of the players is wearing? Answer with the option's letter from the given choices directly.

- A. Black and white
 - B. Blue and white
 - C. Red and white
 - D. Yellow and white
- (Correct Answer: C)



Figure 10: Case Study. Red boxes circle the correct options.

5.2 Unrestrained Amplification of Language Bias in Amateur Models

In Table 6, language bias is amplified by directly removing inter-visual attention and cross-modal attention, but this approach fails to mitigate hallucinations in VLMs; instead, it degrades performance. While removing inter-visual attention and cross-modal attention from inputs, as outlined in eq. (4), appears to be a straightforward and mathematically consistent solution, our experiments reveal that this method is ineffective. We also explore an alternative amateur model by transforming images into captions to replace the original visual input. This ‘Text-only*’ method is expected to amplify language bias even compared to CMG, as it introduces accumulated bias through image captioning. However, this approach is also invalid. These findings indicate that the amateur model cannot be constructed arbitrarily; it must occupy a position that is weaker within the original distribution but cannot too weak to answer questions.

6 CONCLUSION AND LIMITATION

This paper delves into the role of language bias in inducing hallucinations within Vision-Language Models (VLMs) and introduces *Cross-Modal Guidance (CMG)*, an innovative inference strategy designed to counteract such biases. CMG enriches visual context by contrasting outputs from the original model against those from a modified version with disrupted attention maps. Extensive experimentation across various benchmarks has demonstrated CMG’s efficacy in bolstering VLM performance.

Despite its advantages, CMG is not without its challenges. It necessitates careful selection of hyper-parameters, like the mask ratio linked to n_0 in Eq. 12, suggesting a need for tailored adjustments across different scenarios. Optimal results currently require dynamic hyper-parameter tuning, a complexity we aim to explore further in subsequent research.

REFERENCES

- [1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv:2309.16609* (2023).
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *arXiv preprint arXiv:2308.12966* (2023).
- [3] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793* (2023).
- [4] Xiaodong Chen, Yuxuan Hu, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. 2024. Streamlining Redundant Layers to Compress Large Language Models. *arXiv:2403.19135* [cs.CL] <https://arxiv.org/abs/2403.19135>
- [5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.
- [6] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479* (2023).
- [7] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.
- [8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500* [cs.CV] <https://arxiv.org/abs/2305.06500>
- [9] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv:2303.03378* (2023).
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiwu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arxiv:2306.13394* (2024).
- [11] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010* (2023).
- [12] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv preprint arXiv:2305.04790* (2023).
- [13] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *arXiv preprint arXiv:2310.14566* (2023).
- [14] Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and Preventing Hallucinations in Large Vision Language Models. *arXiv:2308.06394* [cs.CV] <https://arxiv.org/abs/2308.06394>
- [15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. *arXiv:1904.09751* [cs.CL] <https://arxiv.org/abs/1904.09751>
- [16] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13872–13882.
- [17] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS* (2021).
- [18] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive Decoding: Open-ended Text Generation as Optimization. *arXiv:2210.15097* [cs.CL] <https://arxiv.org/abs/2210.15097>
- [19] Yifan Li, Du Yifan, Zhou Kun, Wang Jinpeng, Wayne, Zhao Xin, and Wen Ji-Rong. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=xozJw0kZXF>
- [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv:2304.08485* (2023).
- [21] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, and Beidi Chen. 2023. Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time. *arXiv:2310.17157* [cs.LG] <https://arxiv.org/abs/2310.17157>
- [22] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training. *arXiv:2410.08202* [cs.CV] <https://arxiv.org/abs/2410.08202>
- [23] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424* (2023).

- [24] OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774* (2023).
- [25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155* [cs.CL] <https://arxiv.org/abs/2203.02155>
- [26] Yeji Park, Deokyeon Lee, Junsuk Choe, and Buru Chang. 2024. ConVis: Contrastive Decoding with Hallucination Visualization for Mitigating Hallucinations in Multimodal Large Language Models. *arXiv preprint arXiv:2408.13906* (2024).
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- [28] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. *arXiv:2305.14739* [cs.CL] <https://arxiv.org/abs/2305.14739>
- [29] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. PandaGPT: One Model To Instruction-Follow Them All. *arXiv:2305.16355* (2023).
- [30] 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* (2024).
- [31] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv:2108.10904* (2021).
- [32] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9556–9567.
- [33] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Griffon: Spelling out all object locations at any granularity with large language models. In *European Conference on Computer Vision*. Springer, 405–422.
- [34] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* (2023).
- [35] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199* (2023).
- [36] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592* (2023).