# Mitigating Vision-Text Order Bias in Vision-Language Models

Weilin Gan[1,2] Yifan Song[1] Zhuocheng Yu[1] Sujian Li[1]*
[1] Key Laboratory of Computational Linguistics, MOE, School of CS, Peking University
[2] School of Software and Microelectronics, Peking University
ganwl@stu.pku.edu.cn lisujian@pku.edu.cn

## Abstract

*Vision-Language Models (VLMs) suffer from a significant vision-text order bias where suffix order (visual tokens at the end) overwhelmingly underperforms prefix order (visual tokens at the beginning). To explore the impact of visual content order on Vision-Language Models (VLMs), we systematically investigate vision-text order bias. Our preliminary experiments reveal that placing visual tokens at the beginning and end of a sequence yields superior performance compared to placing them in the middle, resulting in a U-shaped performance curve. To address the performance bias brought by non-prefixed input in real-world scenarios, we propose Dual-Order Contrastive Decoding (DOCD), a training-free and lightweight inference scheme designed to enhance non-prefix understanding in VLMs. DOCD parallelly infers on both prefix and suffix orders and contrastively compensates the suffix logits with the prefix logits, utilizing the visual comprehension of prefix order while maintaining close attachment to the visual content of suffix order. Experimental results show that suffix inputs with DOCD can match or even outperform the prefix order in a wide range of difficult benchmarks, including Muirbench, Vlmsareblind, and MMMU-Pro.*

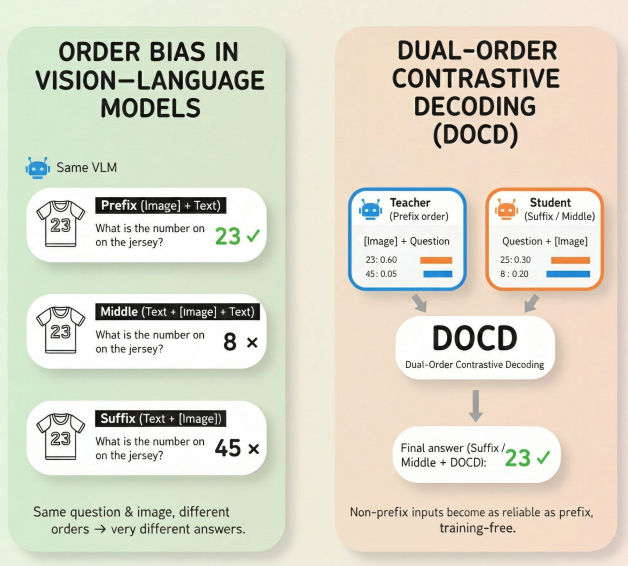



Figure 1. Overview of vision–text order bias in VLMs and the proposed Dual-Order Contrastive Decoding (DOCD). The left panel illustrates how different image–text orders (prefix, middle, suffix) yield drastically different predictions under the same VLM. The right panel shows how DOCD uses a prefix-order teacher and a suffix-order student to re-rank student logits and obtain a robust final answer for non-prefix inputs.

## 1. Introduction

The rapid progress of Vision-Language Models (VLMs) is transforming human–AI interaction beyond unimodal boundaries [1, 19, 20], enabling complex tasks from mathematical reasoning [12, 24, 29] to nuanced spatial reasoning [17]. Despite their impressive capabilities, current VLMs are often constrained by rigid input conventions established during pre-training. Specifically, most leading models—including LLaVA, Qwen2.5-VL, MiMo-VL, and InternVL [1, 9, 14, 27]—recommend a prefix-ordered input format, in which visual information precedes textual prompts, to achieve optimal performance. This structural preference largely stems from the composition of pre-training corpora, which remain dominated by prefix-ordered sequences. However, this structural bias diverges from natural multimodal communication, where visual and textual inputs can appear in arbitrary orders, as in multi-turn dialogue or document-grounded reasoning. Investigating such biases and developing strategies to mitigate them is therefore essential for the robust development and broader application of VLMs.

To the best of our knowledge, no prior work has conducted a systematic investigation for this vision–text order bias. We bridge this gap with an extensive empirical study. By evaluating leading open-source models and varying only

* Sujian Li is the corresponding author.

the relative position of images, our results consistently reveal a U-shaped performance curve: performance peaks at the prefix and suffix positions, but collapses dramatically when images are positioned in the middle of the sequence (lost-in-the-middle), similar to the phenomenon observed in LLMs handling long contexts [10].

Mitigating this severe vision–text order bias is thus an urgent priority for robust VLM deployment. We find that relying on pre-training to solve this bias is both prohibitively expensive and demonstrably insufficient. Moreover, models like Qwen2.5-VL [1] and LLaVA 1.5 [9], despite incorporating significant interleaved vision-text data during pretraining, still exhibit a strong order bias, as our experiments confirm. This establishes the paramount importance of a low-cost, effective, and training-free mitigation strategy.

Although our method improves non-prefix inputs broadly, regardless of where images are placed in the sequence, in this paper we mainly focus on the suffix-order configuration for two reasons. First, it reflects many natural interactions, such as multi-turn dialogues, where users provide a textual query before supplying an image, i.e., a suffix order. Second, complex interleaved documents can be conceptualized as multiple local suffix segments. Enhancing suffix-order performance is therefore the most critical lever for mitigating the overall bias.

This focus reveals a central trade-off. Intuitively, the suffix order should perform poorly, as our attention maps confirm it suffers from weak vision–text fusion. However, our U-shaped curve shows it markedly outperforms the lost-in-the-middle configurations. We attribute this unexpected resilience to its structural proximity advantage: in the suffix order, image tokens are closest to the auto-regressive decoder, minimizing attention decay. This reveals the core tension: the prefix order possesses superior vision–text fusion (an artifact of training), while the suffix order benefits from a proximity advantage (an artifact of architecture).

This trade-off motivates our central question: Can we devise a method that simultaneously leverages the proximity advantage of the suffix order and the superior fusion capabilities of the prefix order? To this end, we introduce Dual-Order Contrastive Decoding (DOCD), a novel, training-free inference scheme that contrasts both input orders within a single model to robustly enhance suffix-order performance. Beyond suffix-order inputs, we further show in our experiments that DOCD also significantly improves mid-sequence configurations (e.g., r = 0.5), where images are placed in the middle of the text. This indicates that DOCD is a general remedy for non-prefix inputs rather than a suffix-specific trick. Moreover, a direct supervised fine-tuning (SFT) baseline with additional interleaved training data yields much smaller gains and sometimes even harms prefix performance, whereas DOCD delivers larger improvements without touching the training pipeline. Together, these results highlight that inference-level dual-order contrast is necessary to robustly mitigate vision–text order bias.

The main contributions of this paper are threefold:

1. Systematic characterization of vision–text order bias. We are the first to systematically investigate and quantify vision–text order bias—i.e., sensitivity to the relative position of images and text, rather than intra-image or intra-text ordering. Across leading VLMs and diverse benchmarks, we reveal a robust U-shaped image-position effect and show that inserting images in the middle of the sequence induces a "lost-in-the-middle" failure mode with attention collapse on preceding tokens, as evidenced by layer-wise attention heatmaps.
2. Dual-Order Contrastive Decoding (DOCD). We propose DOCD, a training-free inference scheme that requires no extra model and no change to the training pipeline. DOCD reuses a single VLM under two input orders (prefix and suffix), and contrastively corrects suffix logits using the prefix distribution. This design strategically leverages both the model's inherent prefix-order bias and the architectural proximity advantage of suffix inputs, and yields consistent gains across non-prefix orders, not only for suffix but also for mid-sequence image positions.
3. Strong and broad empirical gains over both suffix and training baselines. On challenging benchmarks such as MMStar, VLMs-are-Blind and MMMU-Pro, DOCD matches or even surpasses the strong prefix baseline. At the same time, ablation studies show that straightforward supervised fine-tuning on interleaved data yields much smaller gains than DOCD, underscoring that DOCD is a necessary inference-level remedy rather than an engineering tweak.

## 2. Related Work

### 2.1. Vision-Language Models (VLMs)

**Vision–Language Models (VLMs).** Vision–Language Models (VLMs) have flourished recently, changing the way people communicate with AI systems. As early as CLIP, based on contrastive learning to align image and text embeddings, showed enormous potential for zero-shot fine-grained visual classification [15]. Subsequent works such as BLIP/BLIP-2 and SigLIP further improved text–vision embedding alignment [5, 6, 31], but they were still not systems capable of answering user questions based on images and text. The emergence of LLaVA bridged a visual encoder with an LLM via a projection layer and large-scale image–text pretraining, thereby constructing VLMs truly capable of seeing and speaking [8]. Later models, e.g., LLaVA-1.5, InternVL, and Qwen-VL strengthen capabilities via resolution adaptation, visual-token organization and higher-quality data [1, 9, 14].

However, a limitation persists: although some models incorporate interleaved data during training, pretraining and inference are still conducted in a prefix vision–text order (image tokens before text). For example, the Hugging Face multimodal chat template explicitly concatenates image tokens before the text for LLaVA-style models. Some research also models the visual sequence as a prefix to the language model, e.g., Object Recognition as Next-Token Prediction mentions that VLMs treat image tokens as a prefix. And DeepStack demonstrates that VLMs uses visual tokens as a prefix. [13, 28]. Official documents for InternVL and Qwen-VL likewise present examples that place image tokens before text to achieve the best results [14]. In a word, these works indicate a significant vision–text input order bias in current VLMs.

**Order bias in VLMs.** Order biases in sequence models have been widely discussed, but the phenomenon we study is qualitatively different. Prior work on VLMs mostly targets intra-visual ordering—e.g., position bias among multiple images or the temporal order of image sequences—rather than the relative order between vision and text. Tian et al. [22] identify and mitigate position bias across multiple images, while Burn After Reading [18] probes whether VLMs preserve the order of events in image sequences, both revealing biases within the visual stream. Other work, such as Wardle and Teo Sušnjak [26], explores how different interleaved prompting templates (image–text–image, text–image–text, etc.) affect performance, but primarily from a prompt engineering perspective, without isolating or analyzing vision–text order bias as a systematic property of the model. In contrast, we explicitly vary only the relative position of image tokens within a fixed text sequence and evaluate leading open-source VLM (Qwen2.5-VL) on diverse benchmarks [11, 16, 21, 23, 27, 30], showing that prefix order consistently outperforms suffix and that middle placements lead to a pronounced lost-in-the-middle degradation. This establishes vision–text input order bias as a distinct and pervasive failure mode in current VLMs.

## 3. Vision-Text Order Bias

**Task and notation.** Given one or multiple images $\mathcal{I} = \{I_m\}_{m=1}^{M}$ and a textual prompt $\mathbf{x} = [x_1, \ldots, x_L]$, a vision-language model (VLM) with a causal decoder generates an answer sequence $\mathbf{y} = [y_1, \ldots, y_T]$ by maximizing

$$\mathbf{y}^\star = \arg\max_{\mathbf{y}} \prod_{t=1}^{T} p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{z}), \tag{1}$$

where $\mathbf{z}$ denotes the encoded multimodal context consisting of image tokens $\mathbf{v}$ and text tokens $\mathbf{e}$.

**Image position along the sequence.** We study a structural bias induced by the position of the visual tokens within the multimodal context. Let the text sequence $\mathbf{e}$ have length $L$, and let $r \in [0, 1]$ denote a relative position at which we insert the image tokens:

$$\mathbf{z}^{(r)} = [\, x_1, \ldots, x_{\lfloor rL \rfloor},\ \mathbf{v},\ x_{\lfloor rL \rfloor + 1}, \ldots, x_L \,]. \tag{2}$$

Two extreme cases correspond to common practice:

$$\textbf{Prefix (image-first):} \quad r = 0,\ \ \mathbf{z}^{\text{pre}} = [\mathbf{v}, \mathbf{e}], \tag{3}$$

$$\textbf{Suffix (image-last):} \quad r = 1,\ \ \mathbf{z}^{\text{suf}} = [\mathbf{e}, \mathbf{v}]. \tag{4}$$

Under causal attention, information must flow left-to-right within the decoder. Thus, changing $r$ alters the effective fusion pathway and the token-to-output distance of visual evidence, potentially leading to systematic position-dependent performance.

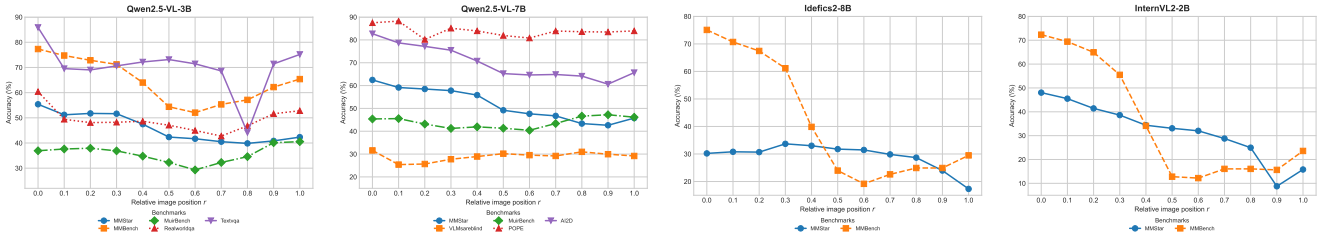
**Evaluation protocol.** To isolate the effect of image position, we adopt a controlled protocol:

1. **Same model and decoding:** identical backbone, prompt template, and decoding hyperparameters (temperature, top-$k$, max new tokens).
2. **Only image position changes:** for each instance, we keep the text and images fixed, and vary $r$ over $\{0.0, 0.1, \ldots, 1.0\}$ by inserting $\mathbf{v}$ at the corresponding relative position in the text.

**U-shaped image-position effect.** Figure 2 plots the accuracy of Qwen2.5-VL-3B/7B, InternVL2-2B and Idefics2 as a function of the relative image position $r$ on a diverse set of benchmarks, including general VQA (MMStar, MMBench, RealWorldQA, TextVQA, AI2D), hallucination tests (POPE, VLMs-are-Blind), and multi-image reasoning (MuirBench).

Across models and benchmarks, we consistently observe a U-shaped curve: placing visual tokens as a prefix ($r = 0$) or suffix ($r = 1$) of the sequence yields substantially higher accuracy than inserting them in the middle ($0 < r < 1$). In other words, visual information in vision–text sequences suffers from a lost-in-the-middle effect, with prefix still outperforming suffix overall. This mirrors similar observations on long contexts in purely textual LLMs [10]. Additional curves and benchmarks are provided in our experimental analysis and supplementary material.

**Layer-wise attention heatmap.** To better understand this position sensitivity, we visualize average attention heatmaps across representative layers under different image positions. As shown in Fig. 3, placing the images in the prefix position allows all subsequent text tokens to attend to




(a) Qwen2.5-VL-3B (b) Qwen2.5-VL-7B (c) Idefics2 (d) InternVL2

Figure 2. **U-shaped image-position effect across models.** Accuracy as a function of the relative image position $r$ in the vision–text sequence for four VLMs: Qwen2.5-VL-3B, Qwen2.5-VL-7B, Idefics2, and InternVL2. Placing the image at the beginning ($r = 0$) or the end ($r = 1$) yields higher performance than placing it in the middle ($0 < r < 1$), revealing a U-shaped position curve. Prefix and Suffix are the two endpoints of this curve.

all image tokens, achieving the most comprehensive vision-text fusion. The heatmaps exhibit fewer low-attention regions and appear more uniform. In contrast, when images are placed in the middle or suffix positions, the model's middle layers exhibit a phenomenon where image tokens pay low attention to the text tokens that precede them. This is visually manifested as distinct dark blocks in the left portion of the heatmaps for the middle layers (as seen in the second and third rows). This incomplete information fusion, compounded by the model's unfamiliarity with these sequential orders, consequently leads to the observed performance degradation.

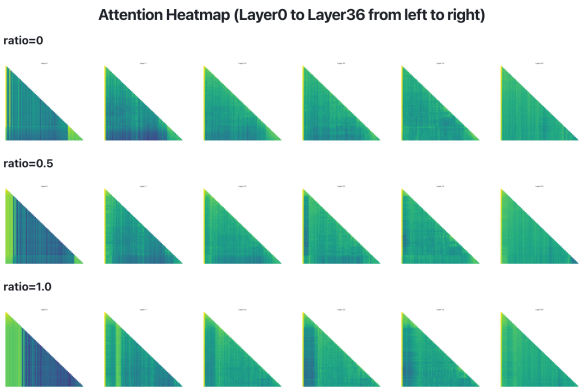



Figure 3. **Layer-wise attention under different image positions.** The three rows are attention heatmaps of Qwen2.5-VL-3B when the insert ratio equals 0, 0.5, and 1, representing prefix, in-the-middle and suffix order, respectively. Deeper layers in Prefix show stronger cross-modal fusion, whereas middle and suffix positions exhibit more localized attention and under-attend distant visual tokens, indicating a weaker fusion pathway.

# 4. Dual-Order Contrastive Decoding

## 4.1. Problem setup

Given one or more images and a textual prompt, we aim to decode an answer sequence under a *non-prefix* vision–text order (in particular the suffix/text-first order), while still exploiting the strong visual fusion that a causal VLM exhibits under the *prefix* (image-first) order.

Let $\mathcal{I} = \{I_m\}_{m=1}^{M}$ be the input images and $\mathbf{x} = [x_1, \ldots, x_L]$ a textual prompt. A vision–language model with parameters $\theta$ defines a conditional distribution over answer sequences $\mathbf{y} = [y_1, \ldots, y_T]$ given a multimodal context $\mathbf{z}$:

$$p_\theta(\mathbf{y} \mid \mathbf{z}) = \prod_{t=1}^{T} p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{z}), \tag{5}$$

where $\mathbf{z}$ is a sequence of image tokens $\mathbf{v}$ and text tokens $\mathbf{e}$ after vision and text encoders.

We use a *teacher* context with prefix order (image-first),

$$\mathbf{z}^{\text{pre}} = [\mathbf{v}, \mathbf{e}] \quad \text{(Prefix / image-first)}, \tag{6}$$

and a *student* context with a generic position-controlled order parameterized by $r \in [0, 1]$. We split the text into a "front" and "back" segment:

$$\mathbf{e}^{\text{front}}(r) = [x_1, \ldots, x_{\lfloor rL \rfloor}], \tag{7}$$

$$\mathbf{e}^{\text{back}}(r) = [x_{\lfloor rL \rfloor + 1}, \ldots, x_L], \tag{8}$$

and define

$$\mathbf{z}^{(r)} = \left[\mathbf{e}^{\text{front}}(r), \mathbf{v}, \mathbf{e}^{\text{back}}(r)\right]. \tag{9}$$

Here $r$=0 recovers the prefix order $\mathbf{z}^{(0)} = [\mathbf{v}, \mathbf{e}]$, $r$=1 gives the *suffix* (text-first) order $\mathbf{z}^{(1)} = [\mathbf{e}, \mathbf{v}]$, and $r \approx 0.5$ inserts images in the middle of the sequence (our "lost-in-the-middle" setting).

Our goal is to decode under the student order $\mathbf{z}^{(r)}$—especially the suffix case $r$=1—while contrastively exploiting the stronger visual fusion available under the prefix teacher $\mathbf{z}^{\text{pre}}$, using a *single* VLM without any additional training.

### 4.2. Dual-order decoding states

At decoding step $t$, we maintain two autoregressive paths that share parameters $\theta$ but differ in multimodal context:

$$p_\theta^{\text{pre}}(y_t \mid \mathbf{y}_{<t}) = p_\theta(y_t \mid \mathbf{y}_{<t}, \mathbf{z}^{\text{pre}}), \tag{10}$$

$$p_\theta^{(r)}(y_t \mid \mathbf{y}_{<t}) = p_\theta\Big(y_t \mid \mathbf{y}_{<t}, \mathbf{z}^{(r)}\Big). \tag{11}$$

In implementation, we realize these as a batch of size two: the teacher sequence with context $\mathbf{z}^{\text{pre}}$ (batch index 0) and the student sequence with context $\mathbf{z}^{(r)}$ (batch index 1), sharing the same transformer and key–value caches.

Let $\boldsymbol{\ell}_t^{\text{pre}} \in \mathbb{R}^V$ and $\boldsymbol{\ell}_t^{(r)} \in \mathbb{R}^V$ be the unnormalized logits over the vocabulary $\mathcal{V}$ for teacher and student at step $t$, obtained from a single forward pass. The corresponding log-probability vectors are

$$\mathbf{s}_t^{\text{pre}} = \log \operatorname{softmax}\big(\boldsymbol{\ell}_t^{\text{pre}}\big), \tag{12}$$

$$\mathbf{s}_t^{(r)} = \log \operatorname{softmax}\big(\boldsymbol{\ell}_t^{(r)}\big). \tag{13}$$

In Dual-Order Contrastive Decoding (DOCD), the student path $p_\theta^{(r)}$ produces the final output sequence, while the teacher path $p_\theta^{\text{pre}}$ provides a vision-attentive distribution that guides token selection.

### 4.3. Top-$k$ dual-order contrastive scoring

To keep DOCD efficient, we restrict contrastive scoring to a small candidate set drawn from the student distribution. At step $t$, we choose

$$C_t = \operatorname{TopK}_k\big(\mathbf{s}_t^{(r)}\big), \tag{14}$$

the $k$ tokens with highest log-probability under the student path (typically $k$=10 or 16).

For each $v \in C_t$, we compute a *dual-order contrastive score* as a convex combination of teacher and student log-probabilities:

$$\operatorname{score}_t(v) = (1-\beta)\,\mathbf{s}_t^{(r)}(v) + \beta\,\mathbf{s}_t^{\text{pre}}(v), \tag{15}$$

where $\beta \in [0,1]$ controls the strength of teacher guidance. When $\beta$=0, DOCD reduces to standard decoding under $\mathbf{z}^{(r)}$; when $\beta$>0, tokens supported by both paths receive higher scores, while tokens favored only by the student are downweighted.

The next token is chosen greedily from the candidate set:

$$y_t = \arg\max_{v \in C_t} \operatorname{score}_t(v), \tag{16}$$

and appended to the partial sequence $\mathbf{y}_{<t}$.

### 4.4. Position-controlled vision–text order

The order parameter $r$ directly controls where the images are inserted relative to the text in the student context $\mathbf{z}^{(r)}$. For suffix inputs, users typically type their textual query first and then provide an image, which corresponds to $r$=1 where all text precedes the images. For middle-position inputs, we choose $r \in (0,1)$ so that a prefix of the text appears before the images and the rest after, matching the "lost-in-the-middle" configuration from our position sweep. When $r$=0, the student order coincides with the prefix order and DOCD degenerates to standard prefix decoding.

This single parameter thus unifies suffix, middle, and prefix configurations, and we reuse the same decoding algorithm for any $r \in [0,1]$. In experiments, we highlight $r$=1 (suffix), $r \approx 0.5$ (middle), and $r$=0 (prefix) as representative cases.

### 4.5. Decoding algorithm and complexity

For a fixed order $r$, DOCD runs both paths in parallel. Given $\mathcal{I}$ and $\mathbf{x}$, we construct $\mathbf{z}^{\text{pre}}$ and $\mathbf{z}^{(r)}$ and feed them as a batch of size two into the same VLM. At each step $t$, a single batched forward pass produces $\boldsymbol{\ell}_t^{\text{pre}}$ and $\boldsymbol{\ell}_t^{(r)}$ together with shared key–value caches. We then compute $\mathbf{s}_t^{\text{pre}}$ and $\mathbf{s}_t^{(r)}$, form $C_t = \operatorname{TopK}_k(\mathbf{s}_t^{(r)})$, evaluate $\operatorname{score}_t(v)$ for $v \in C_t$ using Eq. (15), and select $y_t = \arg\max_{v \in C_t} \operatorname{score}_t(v)$. The chosen token $y_t$ is appended to both paths and used as input for the next step.

DOCD is *training-free*: it reuses a single pre-trained VLM and only increases inference cost by running two contexts in a batched forward pass plus a small top-$k$ reweighting overhead.

### 4.6. Relation to prior contrastive decoding methods

**General contrastive decoding.** Classical contrastive decoding (CD) contrasts a strong "expert" model with a weaker "amateur" model and reweights next-token scores using their disagreement to avoid both dull and hallucinated continuations [7], yielding gains on open-ended generation and reasoning benchmarks.

**Layer-based contrastive decoding.** DoLa (Decoding by Contrasting Layers) [2] removes the need for a separate amateur model by contrasting logits from earlier and later layers of the *same* network, treating intermediate-layer logits as conservative scores and final-layer logits as more adventurous scores to reduce hallucinations.

**Vision-aware contrastive decoding.** Several methods adapt contrastive decoding to vision–language models. Visual Contrastive Decoding (VCD) [4] contrasts predictions conditioned on the original image versus a distorted image to down-weight tokens that ignore visual changes. Multi-Modal Mutual-Information Decoding (M3ID) [3] reweights sampling toward tokens with higher mutual information between text and image. Instruction Contrastive Decoding

(ICD) [25] perturbs the instruction and contrasts predictions under the original and disturbed prompts to suppress visually ungrounded responses driven by instruction bias.

**Our dual-order contrastive decoding.** Our Dual-Order Contrastive Decoding (DOCD) differs in both *what is contrasted* and *what failure mode is targeted*. Instead of contrasting different models, layers, or perturbed inputs, DOCD contrasts two input *orders of the same VLM* that share identical visual and textual content: the image-first (prefix) order and the text-first (suffix) order. The prefix path is used as a teacher to provide a more vision-grounded distribution, while the suffix path matches the user's non-prefixed input and produces the output. By contrastively nudging suffix logits toward the prefix distribution, DOCD directly addresses the vision–text input order bias, rather than generic hallucination, and requires neither extra models nor modified training.

# 5. Experiments

## 5.1. Experimental Setup

**Models.** We mainly evaluate our method on the open-source Qwen2.5-VL models with 3B and 7B parameters, denoted as Qwen2.5-VL-3B and Qwen2.5-VL-7B, respectively. Unless otherwise stated, we use the official checkpoints and follow their recommended preprocessing pipelines for image resolution, tokenization, and chat templating. No additional training or fine-tuning is performed for our main results; all DOCD numbers are obtained at inference time only. For the SFT baseline in Sec. 5.3, we additionally fine-tune the same backbones on extra interleaved vision–text data using standard supervised learning. To assess the generality of our findings beyond the Qwen2.5-VL family, we further apply DOCD to the open-source Idefics2 model on MMStar and MMBench (see Table 2).

**Benchmarks.** We test on eight widely used multimodal benchmarks covering diverse skills: MMStar for comprehensive visual reasoning, MMBench for general multi-choice VQA, MuirBench for multi-image reasoning, RealWorldQA for real-world photo understanding, VLMs-are-Blind for fine-grained perception and text-to-image faithfulness, ScienceQA-IMG for science exams with diagrams, MMMU-Pro for expert-level multi-discipline reasoning, and AI2D for diagram-based question answering. All benchmarks are evaluated using the official splits and scoring protocols provided by `lmms-eval`.

**Practical instantiation of DOCD.** We use the dual-order contrastive decoding scheme described in Sec. 4 with a simple instantiation that closely matches our implementation.

Concretely, for each instance we run the VLM on a batch of size two: (1) a *teacher* sequence in prefix (image-first) order, $\mathbf{z}^{\text{pre}} = [\mathbf{v}, \mathbf{e}]$, and (2) a *student* sequence in order $\mathbf{z}^{(r)}$ where images are inserted at a relative text position $r \in [0, 1]$. At each decoding step $t$, we obtain teacher and student logits $\boldsymbol{\ell}_t^{\text{pre}}$ and $\boldsymbol{\ell}_t^{(r)}$ from a single forward pass, compute log-probabilities $\mathbf{s}_t^{\text{pre}}$ and $\mathbf{s}_t^{(r)}$ via log-softmax, and select a small candidate set $C_t$ of size $k$ by taking the top-$k$ tokens under the student distribution.

For each candidate $v \in C_t$, we compute a dual-order contrastive score

$$\text{score}_t(v) = (1 - \beta_0)\, \mathbf{s}_t^{(r)}(v) + \beta_0\, \mathbf{s}_t^{\text{pre}}(v), \qquad (17)$$

where $\beta_0$ is a scalar hyperparameter controlling the relative weight of teacher guidance. We then choose $y_t = \arg\max_{v \in C_t} \text{score}_t(v)$, append $y_t$ to the output, and feed the same token back into both teacher and student sequences to advance their key–value caches. If an EOS token is generated (we collect all model- and tokenizer-specific EOS IDs), decoding stops.

In our main experiments, we set $k = 16$ and $\beta_0 = 0.5$ by default and decode under the suffix order $r = 1.0$ (text-first, image-last). For the middle-position ablation, we set $r = 0.5$ and slightly increase the candidate size and teacher weight (*e.g.*, `cd_topk`$= 32$, `cd_beta0`$= 1.0$) to test robustness. All these hyperparameters are exposed via `gen_kwargs` in the `lmms-eval` interface (flags `use_cd`, `cd_topk`, `cd_beta0`, `cd_r`), and are kept fixed across datasets within each experimental setting.

## 5.2. Main Results

We first focus on the two practically important endpoints of the position sweep: *Prefix* ($r = 0$, image-first), *Suffix* ($r = 1$, text-first), and our decoding method applied to suffix, denoted as *Suffix+DOCD* (teacher in prefix, student in suffix). Several consistent trends emerge from Table 1:

- **Vision–text order bias.** On almost all benchmarks and both model sizes, *Suffix* (image-last) performs significantly worse than *Prefix* (image-first). For example, on MMStar and AI2D, the gap between Prefix and Suffix reaches 15–18 points for both 3B and 7B. This complements the U-shaped position analysis in Sec. 3 and confirms a strong *vision–text order bias* in current VLMs.
- **Effect of DOCD on suffix order.** *Suffix+DOCD* consistently improves over plain Suffix across all datasets. The gains are especially pronounced on benchmarks that rely heavily on visual grounding and long-context reasoning, such as MMStar, VLMs-are-Blind, AI2D, and MMMU-Pro, where DOCD recovers $\sim$6–18 points relative to the Suffix baseline.
- **Reaching or surpassing Prefix.** In many cases, *Suffix+DOCD* not only closes the gap but also matches or

| Benchmark | Qwen2.5-VL-3B | | | | | Qwen2.5-VL-7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Prefix↑ | Suffix↑ | Suffix+DOCD↑ | $\Delta_{\text{DOCD-Suf}}$ | $\Delta_{\text{DOCD-Pre}}$ | Prefix↑ | Suffix↑ | Suffix+DOCD↑ | $\Delta_{\text{DOCD-Suf}}$ | $\Delta_{\text{DOCD-Pre}}$ |
| MMStar | 56.36 | 41.01 | 56.20 | 15.19 | -0.16 | 62.42 | 45.77 | 62.59 | 16.82 | 0.17 |
| MMBench | 77.66 | 64.86 | 78.09 | 13.23 | 0.43 | 83.16 | 72.94 | 82.82 | 9.88 | -0.34 |
| MuirBench | 36.88 | 40.35 | 41.12 | 0.77 | 4.24 | 45.38 | 45.23 | 47.27 | 2.04 | 1.89 |
| RealWorldQA | 58.56 | 51.63 | 63.40 | 11.77 | 4.84 | 69.02 | 61.70 | 68.63 | 6.93 | -0.39 |
| VLMs-are-Blind | 32.39 | 20.36 | 34.71 | 14.35 | 2.32 | 31.35 | 28.74 | 33.51 | 4.77 | 2.16 |
| ScienceQA-IMG | 80.07 | 78.88 | 80.61 | 1.73 | 0.54 | 87.31 | 75.51 | 87.41 | 11.90 | 0.10 |
| MMMU-Pro | 32.25 | 26.42 | 32.72 | 6.30 | 0.47 | 35.84 | 31.79 | 37.80 | 6.01 | 1.96 |
| AI2D | 78.72 | 64.15 | 78.30 | 14.15 | -0.42 | 82.84 | 65.22 | 83.03 | 17.81 | 0.19 |

Table 1. **Order sensitivity and DOCD gains (suffix order).** Results on eight benchmarks with Qwen2.5-VL-3B and Qwen2.5-VL-7B under different vision–text orders. Suffix (image-last, $r = 1$) markedly underperforms Prefix (image-first, $r = 0$), confirming a strong vision–text order bias. Our decoding method (Suffix+DOCD) consistently improves upon Suffix (blue numbers) and often matches or exceeds Prefix (positive $\Delta_{\text{DOCD-Pre}}$, marked in red). Numbers are accuracies (%) as provided by our runs.

| Benchmark | Prefix↑ | Suffix↑ | Suffix+DOCD↑ | $\Delta_{\text{DOCD-Suf}}$ | $\Delta_{\text{DOCD-Pre}}$ |
|---|---|---|---|---|---|
| MMStar | 30.21 | 17.31 | 29.97 | 12.66 | -0.24 |
| MMBench | 75.09 | 29.50 | 72.59 | 43.10 | -2.49 |

Table 2. **DOCD on Idefics2 (suffix order).** Results on MMStar and MMBench with the Idefics2 VLM. Suffix (image-last) strongly underperforms Prefix, reproducing the vision–text order bias on a different backbone. Applying DOCD under the suffix order substantially improves performance, recovering most of the prefix–suffix gap.

slightly exceeds *Prefix*. For instance, on MMMU-Pro and several other benchmarks, Suffix+DOCD attains performance on par with or higher than Prefix for both 3B and 7B, suggesting that combining dual-order signals can sometimes be strictly better than relying on Prefix alone.

Beyond Qwen2.5-VL, Table 2 shows that Idefics2 exhibits a similar order sensitivity: moving from prefix to suffix leads to large drops on both MMStar and MMBench, while Suffix+DOCD recovers 12.66 and 43.10 points respectively, nearly matching the prefix baseline. This supports that the U-shaped position effect and the benefit of DOCD are not specific to a single architecture.

Our main results focus on suffix inputs ($r = 1$), which capture many natural user workflows. However, the position sweep in Sec. 3 reveals that placing images in the *middle* of the sequence (roughly $r \approx 0.5$) often induces the strongest degradation ("lost-in-the-middle"). To test whether DOCD generalizes beyond suffix order, we evaluate a student order with $r = 0.5$ and apply the same dual-order contrastive decoding (teacher in prefix order).

Table 3 reports results for representative benchmarks under three configurations: Prefix, Middle (images inserted at $r = 0.5$), and Middle+DOCD (teacher in prefix, student in middle). We observe that:

- Middle-position baselines are weaker than Prefix, echoing the lost-in-the-middle effect revealed in Sec. 3.
- DOCD with $r = 0.5$ yields large improvements over the middle baseline, often recovering 10–18 points on MMStar and AI2D and 6–12 points on other benchmarks.
- The resulting Middle+DOCD performance is typically very close to Prefix, and in some cases slightly higher, demonstrating that our dual-order contrastive scheme transfers prefix visual fusion not only to suffix inputs but to general non-prefix configurations.

### 5.3. Ablation: DOCD vs. Supervised Fine-Tuning

A natural question is whether the vision–text order bias can be removed simply by fine-tuning on additional interleaved data. To this end, we compare DOCD with a supervised fine-tuning (SFT) baseline where the same Qwen2.5-VL backbones are trained on extra interleaved vision–text examples using standard cross-entropy loss, and then decoded under the suffix order without any contrastive scheme. Table 4 reports the comparison between plain suffix, Suffix+SFT, and Suffix+DOCD. The comparison shows that:

- SFT provides only modest improvements for suffix inputs and is sometimes *harmful*, especially on more challenging reasoning benchmarks such as MMStar, MuirBench, MMMU-Pro, and VLMs-are-Blind (negative $\Delta_{\text{SFT-Suf}}$).
- In contrast, Suffix+DOCD consistently outperforms both plain suffix and Suffix+SFT by a large margin, often adding 7–16 points on top of the SFT model.
- These results indicate that merely exposing the model to more interleaved training data is insufficient to eliminate vision–text order bias. Inference-time dual-order contrast is a *necessary* mechanism to transfer the strong prefix visual fusion to user-preferred non-prefix configurations.

### 5.4. Discussion and Limitations

Overall, our experiments demonstrate that Dual-Order Contrastive Decoding is:

| Benchmark | Qwen2.5-VL-3B | | | | | Qwen2.5-VL-7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Prefix↑ | Middle↑ | Middle+DOCD↑ | $\Delta_{\text{DOCD-Mid}}$ | $\Delta_{\text{DOCD-Pre}}$ | Prefix↑ | Middle↑ | Middle+DOCD↑ | $\Delta_{\text{DOCD-Mid}}$ | $\Delta_{\text{DOCD-Pre}}$ |
| MMStar | 56.36 | 41.01 | 56.20 | 15.19 | -0.16 | 62.42 | 45.77 | 62.08 | 16.31 | -0.34 |
| MMBench | 77.66 | 64.86 | 77.67 | 12.81 | 0.01 | 83.16 | 72.94 | 82.90 | 9.97 | -0.26 |
| MuirBench | 36.88 | 40.35 | 40.42 | 0.07 | 3.54 | 45.38 | 45.23 | 47.73 | 2.50 | 2.35 |
| RealWorldQA | 58.56 | 51.63 | 63.40 | 11.77 | 4.84 | 69.02 | 61.70 | 69.15 | 7.45 | 0.13 |
| ScienceQA-IMG | 80.07 | 78.88 | 80.32 | 1.44 | 0.25 | 87.31 | 75.51 | 87.06 | 11.55 | -0.25 |
| MMMU-Pro | 32.25 | 26.42 | 32.20 | 5.78 | -0.05 | 35.84 | 31.79 | 37.98 | 6.19 | 2.14 |
| AI2D | 78.72 | 64.15 | 78.53 | 14.38 | -0.19 | 82.84 | 65.22 | 83.06 | 17.84 | 0.22 |

Table 3. **DOCD on middle-position inputs ($r$=0.5).** We insert images in the middle of the text (student order $r = 0.5$) and run DOCD with a prefix teacher. Middle+DOCD consistently improves over the middle baseline (blue numbers), often recovering most of the gap to Prefix. In several settings, Middle+DOCD matches or slightly exceeds Prefix (positive $\Delta_{\text{DOCD-Pre}}$, marked in red), showing that DOCD is a general remedy for non-prefix orders, not only for suffix.

| Benchmark | Qwen2.5-VL-3B | | | | | Qwen2.5-VL-7B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Suffix | Suffix+SFT | Suffix+DOCD | $\Delta_{\text{SFT-Suf}}$ | $\Delta_{\text{DOCD-SFT}}$ | Suffix | Suffix+SFT | Suffix+DOCD | $\Delta_{\text{SFT-Suf}}$ | $\Delta_{\text{DOCD-SFT}}$ |
| MMStar | 41.01 | 40.62 | 56.17 | -0.39 | 15.55 | 45.77 | 47.39 | 62.59 | 1.62 | 15.20 |
| MMBench | 64.86 | 65.80 | 77.84 | 0.94 | 12.04 | 72.94 | 71.13 | 82.82 | -1.80 | 11.68 |
| MuirBench | 40.35 | 34.65 | 40.88 | -5.70 | 6.23 | 45.23 | 44.16 | 47.27 | -1.07 | 3.11 |
| RealWorldQA | 51.63 | 57.52 | 61.05 | 5.89 | 3.53 | 61.70 | 63.17 | 68.63 | 1.47 | 5.46 |
| VLMs-are-Blind | 20.36 | 22.06 | 35.23 | 1.70 | 13.17 | 28.74 | 24.30 | 33.51 | -4.44 | 9.21 |
| ScienceQA-IMG | 78.88 | 73.33 | 80.81 | -5.55 | 7.48 | 75.51 | 78.38 | 87.41 | 2.87 | 9.03 |
| MMMU-Pro | 26.42 | 23.41 | 32.66 | -3.01 | 9.25 | 31.79 | 30.29 | 37.80 | -1.50 | 7.51 |
| AI2D | 64.15 | 64.35 | 78.43 | 0.20 | 14.08 | 65.22 | 66.74 | 83.03 | 1.52 | 16.29 |

Table 4. **DOCD vs. supervised fine-tuning (suffix order).** Comparison between plain suffix decoding, a supervised fine-tuning (SFT) baseline on interleaved data, and our Dual-Order Contrastive Decoding (DOCD), all evaluated under the suffix order. $\Delta_{\text{SFT-Suf}}$ measures the direct benefit of SFT over the original model, while $\Delta_{\text{DOCD-SFT}}$ (blue numbers when positive) measures the additional benefit of DOCD on top of (or instead of) SFT. DOCD typically yields much larger gains than SFT and can even correct cases where SFT hurts performance.

- *Effective*: it recovers up to ∼17 points under suffix or middle orders and often matches or slightly exceeds prefix performance across a wide range of benchmarks and both 3B/7B model scales, and yields large gains for Idefics2 under the suffix order.
- *General*: it improves not only suffix inputs but also mid-sequence image placements ($r$=0.5), mitigating the lost-in-the-middle failure mode and showing broad applicability to non-prefix orders.
- *Training-free*: it requires no architecture change or additional model, and outperforms SFT baselines that consume extra interleaved training data.

There are, however, several limitations. First, DOCD doubles the number of autoregressive streams at inference time (teacher + student), leading to non-negligible latency overhead, although shared vision encoding and batched forward passes mitigate this cost. Deploying DOCD in strict real-time settings may require further engineering optimization or partial application (e.g., only for difficult instances). Second, while DOCD substantially narrows the gap between prefix and non-prefix orders, it does not completely eliminate all order effects on every dataset, especially for highly specialized tasks. Finally, our present study focuses on image inputs (single- and multi-image cases) and short video clips; extending DOCD to longer videos and more complex interleaved document scenarios is an interesting direction for future work.

## 6. Conclusion

We empirically identify a strong vision–text order bias in modern vision–language models: the relative position of image tokens alone can cause large performance fluctuations, with prefix inputs generally outperforming suffix and middle placements. To address this, we propose Dual-Order Contrastive Decoding (DOCD), a simple, training-free inference scheme that reliably improves non-prefix orders and often closes most of the gap to prefix performance. Future work includes reducing the computational overhead of dual-order decoding, extending the approach to longer videos and document-level interleaving, and exploring tighter integration with training objectives to further mitigate order sensitivity.

## Acknowledgement

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Natural Science Foundation of China projects (No. 62476010 and 92470205).

## References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. 1, 2

[2] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 5

[3] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14303–14312. IEEE, 2024. 5

[4] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE, 2024. 5

[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. 2

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 2

[7] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12286–12312, Toronto, Canada, 2023. 5

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE, 2024. 1, 2

[10] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 2, 3

[11] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, pages 216–233. Springer, 2024. 3

[12] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1

[13] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 3

[14] OpenGVLab. Internvl docs: Chat data format. `https://internvl.readthedocs.io/en/latest/get_started/chat_data_format.html`, 2025. 1, 2, 3

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 2

[16] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind 128374. In *Computer Vision - ACCV 2024 - 17th Asian Conference on Computer Vision, Hanoi, Vietnam, December 8-12, 2024, Proceedings, Part V*, pages 293–309. Springer, 2024. 3

[17] Navid Rajabi and Jana Kosecka. GSR-BENCH: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *CoRR*, abs/2406.13246, 2024. 1

[18] Yingjin Song, Yupei Du, Denis Paperno, and Albert Gatt. Burn after reading: Do multimodal large language models truly capture order of events in image sequences? In *Findings of the Association for Computational Linguistics,*

*ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24316–24342. Association for Computational Linguistics, 2025. 3

[19] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 1

[20] OpenAI team. Gpt-4o system card, 2024. 1

[21] Qwen Team. Qwen2.5-vl: Model card and usage examples. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct, 2025. 3

[22] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. pages 10599–10609, 2025. 3

[23] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28*, 2025. 3

[24] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15*, 2024. 1

[25] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15840–15853. Association for Computational Linguistics, 2024. 6

[26] Grant Wardle and Teo Susnjak. Image first or text first? optimising the sequencing of modalities in large language model prompting and reasoning tasks. *Big Data Cogn. Comput.*, 9 (6):149, 2025. 3

[27] Xiaomi LLM-Core Team et al. Mimo-vl technical report. *CoRR*, abs/2506.03569, 2025. 1, 3

[28] Kaiyu Yue, Bor-Chun Chen, Jonas Geiping, Hengduo Li, Tom Goldstein, and Ser-Nam Lim. Object recognition as next token prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16645–16656. IEEE, 2024. 3

[29] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE, 2024. 1

[30] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15134–15186. Association for Computational Linguistics, 2025. 3

[31] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE, 2023. 2