

# Remember Me: Bridging the Long-Range Gap in LVLMs with Three-Step Inference-Only Decay Resilience Strategies

Peng Gao<sup>\*12</sup>, Yujian Lee<sup>\*12</sup>, Xiaofeng Zhang<sup>\*3†</sup>, Zailong Chen<sup>4</sup>, Hui Zhang<sup>2‡</sup>

<sup>1</sup>Hong Kong Baptist University, HongKong, China

<sup>2</sup>Beijing Normal-Kong Baptist University, ZhuHai, China

<sup>3</sup>Shanghai Jiao Tong University, ShangHai, China

<sup>4</sup>University of Wollongong, New South Wales, Australia

## Abstract

Large Vision-Language Models (LVLMs) have achieved impressive performance across a wide range of multimodal tasks. However, they still face critical challenges in modeling long-range dependencies under the usage of Rotary Positional Encoding (RoPE). Although it can facilitate precise modeling of token positions, it induces progressive attention decay as token distance increases, especially with progressive attention decay over distant token pairs, which severely impairs the model’s ability to remember global context. To alleviate this issue, we propose inference-only **Three-step Decay Resilience Strategies (T-DRS)**, comprising (1) **Semantic-Driven DRS (SD-DRS)**, amplifying semantically meaningful but distant signals via content-aware residuals, (2) **Distance-aware Control DRS (DC-DRS)**, which can purify attention by smoothly modulating weights based on positional distances, suppressing noise while preserving locality, and (3) **re-Reinforce Distant DRS (reRD-DRS)**, consolidating the remaining informative remote dependencies to maintain global coherence. Together, the T-DRS recover suppressed long-range token pairs without harming local inductive biases. Extensive experiments on Vision Question Answering (VQA) benchmarks demonstrate that T-DRS can consistently improve performance in an inference-only manner.

## Extended version —

<https://github.com/labixiaoq-qq/Remember-me>

## Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in perceiving and understanding complex multimodal information, enabling a broad spectrum of downstream tasks, referring expression comprehension, and multimodal dialogue (Liu et al. 2023; Dai et al. 2023; Zhu et al. 2023a; Gao et al. 2025; Lee et al. 2025a,b). These models are typically built upon large-scale

pretrained language models and extended to handle visual inputs through image encoders and cross-modal alignment modules. By effectively integrating interleaved sequences of images and texts, LVLMs can provide contextually grounded and semantically rich responses to user queries. At the architectural level, most LVLMs are based on the Transformer framework (Vaswani et al. 2017), which has become the de facto standard for modeling sequential and structured data. However, despite its representational power, the vanilla Transformer architecture is inherently permutation-invariant and lacks a built-in notion of order. That is, its self-attention mechanism computes token-to-token dependencies without considering their positions in the sequence, making it fundamentally blind to token order. To overcome this limitation, Positional Encoding (PE) schemes are introduced to inject order-sensitive information into the input representations, enabling the model to reason about structural dependencies and temporal relations across the input sequence.

PE strategies can be broadly classified into absolute and relative formulations. **Absolute PE**, which was proposed by (Vaswani et al. 2017), assigns each position a fixed embedding based on trigonometric functions. However, it cannot model the relative positions or distances between tokens explicitly, which limits its expressiveness, especially in scenarios requiring relational reasoning. **Relative PE**, on the other hand, captures the difference or distance between token positions, making it more suitable for encoding pairwise relations (Shaw, Uszkoreit, and Vaswani 2018; Dai et al. 2019). Among relative PE variants, two approaches have gained wide adoption in modern architectures: **learnable relative embeddings** and **Rotary Position Embedding (RoPE)** (Su et al. 2024). Learnable relative PE introduces trainable parameters to model positional offsets, allowing the attention mechanism to emphasize local patterns and adjacent dependencies. While effective for short sequences, such methods often exhibit poor extrapolation ability, failing to generalize beyond the maximum training length due to uninitialized or missing embeddings at test time. **RoPE**, by contrast, encodes relative positions continuously by applying complex-valued rotations. This approach maintains the distance-aware inductive bias directly

\*These authors contributed equally.

†project leader

‡Corresponding Author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

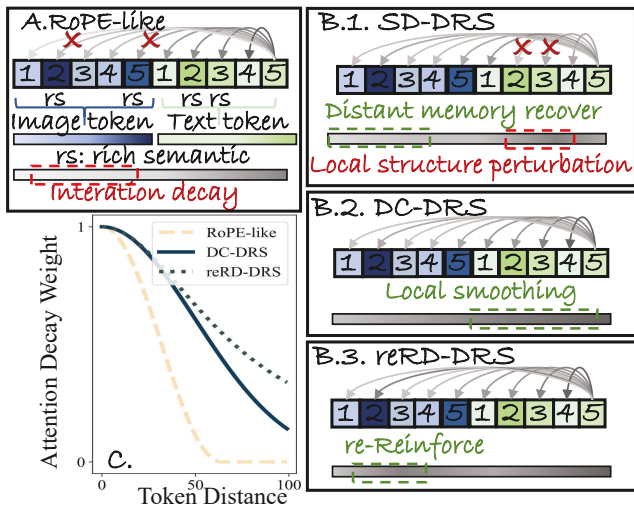


Figure 1: (A.) RoPE suffers from rapid attention decay over long token distances, weakening long-range dependencies between text and image tokens. (B.1, B.2, B.3) The proposed T-DRS framework alleviates this via three stages to collaboratively enhance and stabilize distant attention. (C.) The curves of the cross-attention weight decay in RoPE-like approaches, DC-DRS, and reRD-DRS, respectively.

within the dot-product between query and key vector, offering better generalization and compatibility with long sequences, which has been adopted in many large language and vision-language models (Touvron et al. 2023; Bai et al. 2025; Liu et al. 2023; Chen et al. 2024). However, as illustrated in Fig. 1(A), RoPE-like approaches can not be able to capture distant token information that is rich in semantics. The curve of RoPE-like in Fig. 1(C) exhibits a progressive decay of attention weights as the token distance increases, leading to the suppression of long-range interactions. Such decay severely limits the model’s ability to capture global context or long-range dependencies, both of which are essential in tasks requiring compositional reasoning. Several efforts explore position interpolation (Chen et al. 2023b; Su et al. 2024; Zhu et al. 2023b) or memory extension techniques to recover long-sequence behaviors (Xing et al. 2024; Zhao et al. 2025; Li et al. 2025b; Tang et al. 2025c; Barbero et al. 2024) to mitigate this limitation. But most of these models require retraining or finetuning, which may not be feasible in resource-constrained settings.

Motivated by the degradation of long-range dependencies in LVLMs and the limitations posed by resource constraints, we introduce an inference-only framework named **Three-stage Decay-Resilient Strategies (T-DRS)**. Without model re-training, T-DRS collectively enhance attention robustness over extended token sequences through 1) **Semantic-Driven DRS (SD-DRS)**, 2) **Distance-aware Control DRS (DC-DRS)**, and 3) **re-Reinforce Distant DRS (reRD-DRS)**. In Fig. 1(B.1), the SD-DRS initially introduces a content-aware residual term into the pre-softmax attention logits to strengthen semantically aligned but distant token interactions, ensuring that vital contextual information is not pre-

maturely suppressed. Having the potential of perturbed local structures, DC-DRS in Fig. 1(B.2) enforces a locality-aware prior, yielding a more structured and focused attention distribution, with its curved form in Fig. 1(C) functions as a smoothing and distance-aware mechanism that reduces attention weights. To further cope with the tokens that are distant but rich in semantics, we have reRD-DRS in Fig. 1(B.3). It introduces a reinforcement mechanism that selectively restores attention mass to under-attended yet semantically important long-range pairs, compensating for cumulative decay effects while preserving the local dependencies established by the previous DRS. Together, the T-DRS offers a principled and interpretable solution for preserving both global coherence and local precision in long-context reasoning. The entire pipeline operates at inference time, remains fully differentiable, and is compatible with existing transformer architectures.

**Our main contributions are summarized as follows:**

- We propose **T-DRS**, a training-free, inference-only framework that alleviates long-range attention decay through three complementary decay-resilient strategies.
- **Distant recovery and locality smoothing:** SD-DRS strengthens distant semantic links via content-aware residuals, while DC-DRS imposes a smooth locality bias to refine attention allocation.
- **Residual re-weighting:** reRD-DRS reinforces semantically important long-range dependencies by recovering suppressed attention weights, thereby enhancing global reasoning capabilities.
- Extensive experiments on benchmark vision-language answering tasks demonstrate that T-DRS consistently outperforms state-of-the-art baselines, especially under long-context and reasoning-heavy scenarios.

**Related Work**

**Multimodal Learning with LVLMs**

LVLMs have become a dominant paradigm for unified multimodal understanding and generation (Xue et al. 2025). By integrating high-capacity vision encoders with large-scale language models, LVLMs are capable of processing interleaved image and text inputs and generating free-form responses (Radford et al. 2021; Touvron et al. 2023; Fang et al. 2023; Tang et al. 2025a). Recent advances leverage instruction tuning (Ouyang et al. 2022; Zhang et al. 2022) to align vision and language modalities under unified prompts. Models such as Flamingo (Alayrac et al. 2022), BLIP-2 (Li et al. 2023a), and MiniGPT-4 (Zhu et al. 2023a) introduce lightweight adapters (i.e., Q-Formers) to efficiently inject vision into language decoders. LLaVA (Liu et al. 2023; An et al. 2025) integrates CLIP (Zhu et al. 2024b) visual features directly into a Vicuna-style decoder, demonstrating impressive performance on a wide range of multimodal benchmarks without explicit cross-attention layers. These models have significantly advanced downstream tasks such as visual question answering (Tang et al. 2025b; Zhu et al. 2024a), image-text retrieval, and multimodal dialogue, paving the way for unified multimodal agents (Yang et al. 2025).

## Position Embeddings in LVLMS

PE plays a crucial role in LVLMS, as it informs the model of token order and structural relationships across modalities. Early vision-language models typically adopt absolute (Dosovitskiy et al. 2020) or learnable (Li et al. 2021) positional embeddings, often applied independently to text and image tokens. While effective in short sequences, these methods generalize poorly to long or variable-length contexts. To address this, recent LVLMS adopt RoPE (Su et al. 2024), a relative encoding mechanism that applies sinusoidal rotations to query and key vectors. RoPE encodes distance implicitly, supports sequence length extrapolation, and avoids additional parameters, making it especially suitable for decoder-only architectures. However, RoPE inherently introduces a long-range decay effect: as the relative distance between tokens increases, their attention scores diminish due to the orthogonal nature of high-angle rotations.

While such decay aligns with the local inductive bias in language modeling, it can be problematic in multimodal tasks like VQA (Lu et al. 2022; Agrawal et al. 2018), where image and text tokens are often separated by large positional gaps. For example, a question word appearing early in the sequence may need to attend to a relevant visual region encoded much later. In these cases, RoPE’s distance-based suppression may weaken essential cross-modal interactions, impairing the model’s ability to align semantics across modalities. In this work, we propose a semantic-aware strategy to enhance RoPE’s long-range attention capabilities in multimodal contexts.

### Preliminary

In this section, we present pre-definitions of LVLMS, the baseline method RoPE, and our motivation to have T-DRS.

### Large Vision-Language Models

Large Vision-Language Models (LVLMS). Given a pre-trained vision encoder  $F_v$  and a lightweight projection head  $f$ , the visual content  $I_v$  is projected into the embedding space of the Large Language Model (LLM) as follows:

$$S_{\text{vision}} = f(F_v(I_v)) = \{w_0^v, w_1^v, \dots, w_{V-1}^v\}. \quad (1)$$

For the instruction prompt  $I_t$ , the LLM encodes it into a  $T$  textual tokens language embedding as

$$S_{\text{instr}} = F_t(I_t) = \{w_0^t, w_1^t, \dots, w_{T-1}^t\}. \quad (2)$$

$S_{\text{vision}}$  and  $S_{\text{instr}}$  are concatenated.

$$S = \{S_{\text{vision}}, S_{\text{instr}}\} \in \mathbf{R}^{(V+T) \times d}, \quad (3)$$

into a single multimodal token sequence, where  $d$  is the shared embedding dimension. The cross-modal attention layers interleave and attend to all the  $V + T$  tokens, seamlessly fusing visual and textual information to facilitate downstream generation.

### Rotary Positional Embedding

Rotary Positional Embedding (RoPE) encodes the position of the token through a rotation matrix applied to each embedded token. For a token  $w_m$  at position  $m \in [1, V + T]$ ,

the corresponding rotation matrix

$$R_{\theta, m}^d = \text{diag} \left( \begin{bmatrix} \cos m\theta_1 & -\sin m\theta_1 \\ \sin m\theta_1 & \cos m\theta_1 \end{bmatrix}, \dots, \begin{bmatrix} \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{bmatrix} \right) \quad (4)$$

is constructed as a block diagonal matrix composed of  $\frac{d}{2}$  two-dimensional rotations, where the rotation frequencies  $\{\theta_i\}_{i=1}^{d/2}$  follow a predefined sinusoidal schedule  $\theta_i = 10000^{-2(i-1)/d}$ . In practice, RoPE is applied to both query and key vectors across all layers of Transformer-based LVLMS. This design encodes the relative distance between tokens directly into the attention score. The attention logits  $A$  between query  $Q_i$  and key  $K_j$  are computed as:

$$A_{i,j} = \text{softmax} \left( \frac{Q_i^\top R_{j-i} K_j}{\sqrt{d}} \right), \quad (5)$$

where  $R_{j-i} = (R_{\theta, i}^d)^\top R_{\theta, j}^d$  is the relative rotation matrix based on the positional gap  $j - i$ .

### Motivation of T-DRS

While RoPE provides an elegant and efficient way to encode relative positions, its inherent design leads to a phenomenon known as *long-range decay* (Wei et al. 2025). As the relative distance  $|j - i|$  increases, the effective attention score  $A_{i,j}$  between tokens  $Q_i$  and  $K_j$  decreases due to the rotational orthogonality introduced by large-angle embeddings. This behavior aligns well with language modeling tasks, where distant tokens are typically less semantically related. However, in multimodal question answering tasks, the distant dependencies are less important often does not hold. In such settings, a question token may need to attend to visual features located far apart in the sequence, especially when the input consists of concatenated long textual instructions and high-dimensional visual tokens. RoPE’s tendency to suppress long-distance attention can thus hinder effective cross-modal alignment, as relevant visual evidence may be located dozens or even hundreds of tokens away from the query. This results in a mismatch between semantic importance and attention strength.

Motivated by this limitation, we aim to mitigate RoPE’s long-range decay, allowing the model to better preserve meaningful dependencies across long token spans. To this end, we introduce T-DRS, a simple yet effective strategy designed to reinforce long-range attention.

### The proposed T-DRS Framework

This section provides a detailed description of the T-DRS framework, containing three consecutive strategies: 1) Semantic-Driven DRS (SD-DRS), 2) Distance-aware Control DRS (DC-DRS), and 3) re-Reinforce Distant DRS (reRD-DRS). The framework is shown in Fig. 2.

#### Semantic-Driven DRS (SD-DRS)

Firstly, to address the visual regions and relevant text descriptions that reside at distant positions but are highly correlated, we have SD-DRS for recovery, which introduces a lightweight, training-free semantic bias to enhance  $A$  from Eq.(5). Under the influence of RoPE, the standard softmax

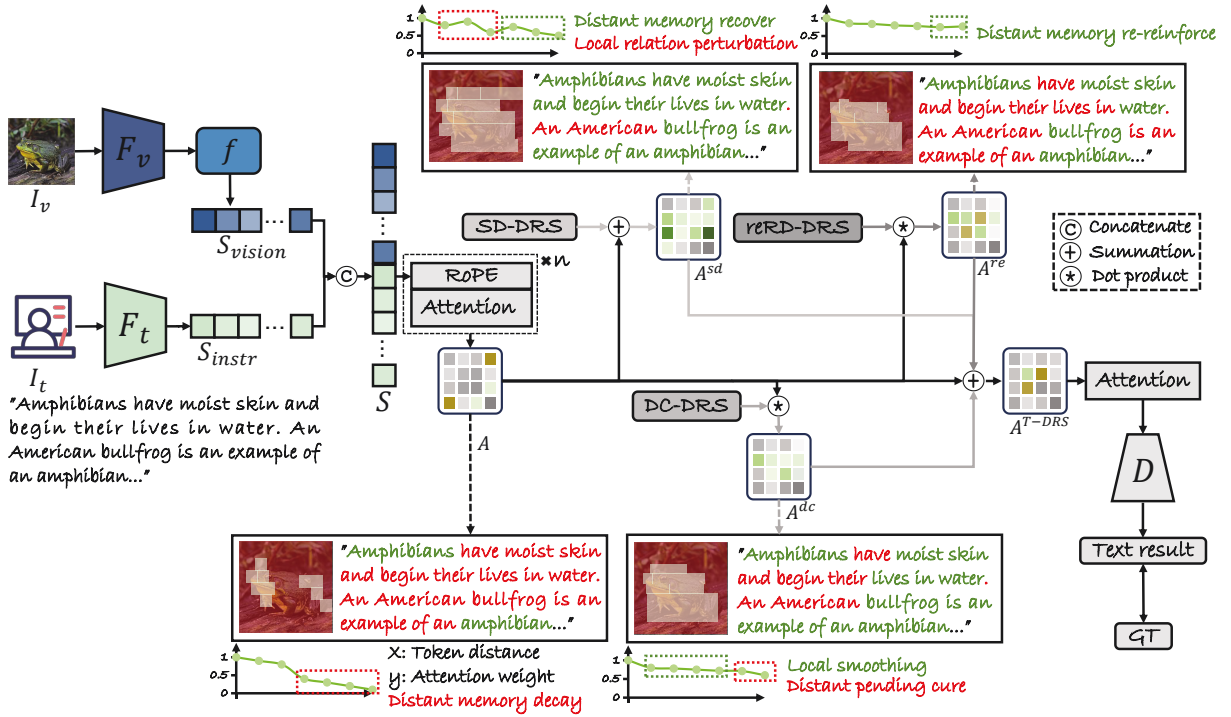


Figure 2: The proposed T-DRS is an inference-only framework. Given image and text inputs, visual and instruction features ( $S_{vision}$ ,  $S_{instr}$ ) can be extracted, then concatenated as  $S$ , input into the RoPE-attention architecture. The output attention logits  $A$  experienced long-range dependencies decay, we then modulated three DRS: (1) SD-DRS ( $A^{sd}$ ) amplifies semantically relevant distant tokens; (2) DC-DRS ( $A^{dc}$ ) applies local smoothing technique, avoiding the local structure being perturbing, and (3) reRD-DRS ( $A^{re}$ ) is specialized for re-strengthens long-range attention with rich semantics. Integrating the pre-softmax attention map,  $A^{T-DRS}$  is used to generate the final output.

attention no longer treats all token pairs equally. Although it assumes that the dot product sufficiently captures relevance, this assumption becomes invalid for distant yet semantically important pairs, leading to unfair suppression of long-range dependencies. To mitigate this, we explicitly model semantic correlation beyond positional proximity. The key intuition is that semantically related token pairs should retain high attention weights. To this end, SD-DRS first computes a semantic affinity map:

$$\text{sem\_sim}_{i,j} = \cos(Q_i, K_j) = \frac{Q_i \cdot K_j}{\|Q_i\| \cdot \|K_j\|}, \quad (6)$$

which captures the cosine similarity between query and key embeddings. To make the values compatible with attention logits, we normalize the similarity to a positive range:

$$\text{sem\_pos}_{i,j} = \frac{1}{2} (\text{sem\_sim}_{i,j} + 1), \quad (7)$$

yielding a matrix values in  $[0, 1]$ , higher values indicate stronger semantic correspondence. In the final output of the SD-DRS, we add the bias to the original logits as:

$$A_{i,j}^{sd} = A_{i,j} + \text{sem\_pos}_{i,j}, \quad (8)$$

effectively amplifying attention between semantically meaningful token pairs, especially those that are distant in position. This introduces a content-aware bias that complements

the position-centric nature of traditional transformers, thus restoring the model's ability to capture long-range semantic dependencies. To modulate attention in a stable and bounded manner, we transform the semantic position similarity map  $\text{sem\_pos}_{i,j}$  into a continuous scaling factor

$$\text{scale}_{i,j} = \frac{\text{sem\_pos}_{i,j} - \min(\text{sem\_pos})}{\max(\text{sem\_pos}) - \min(\text{sem\_pos})}, \quad (9)$$

instead of directly injecting a semantic bias into the attention logits, which may destabilize training or inference. This normalization rescales each semantic position similarity score into the  $[0, 1]$  range, enabling adaptive and bounded control over the two subsequent attention decay strategies.

### Distance-aware Control DRS (DC-DRS)

While semantic cues are vital for guiding attention, local structure remains an essential inductive bias in sequential modeling. In Fig. 2, although SD-DRS recovers distant token pairs, it slightly perturbs the local structure. To address this, we propose DC-DRS, a structure-preserving attention modulation smoothing method that explicitly incorporates token separation effects into the attention calculation.

DC-DRS formulates a smooth, distance-dependent attenuation profile with explicit analytic guarantees. Concretely, for a query at position  $i$  and a key vector at position  $j$ , we

define the relative distance as  $d_{i,j} = |i - j|$ , and enforce the following design criteria on the decay function  $w(d)$  inspired by (Bishop 2006):

- **Monotonicity:**  $w(d)$  must strictly decrease with  $d$ , respecting the intuition of locality bias.
- **Smoothness:** To maintain compatibility with gradient-based optimization, the decay profile is required to be continuous and differentiable.
- **Lower-bound preservation:** A non-zero minimum attention value  $w_{\min}^{dc}$  should be enforced at the maximum distance  $d_{\max}$ , to have persistent connectivity throughout the entire sequence.

To satisfy these constraints, we construct a closed-form attenuation profile parameterized by a decay scale  $\sigma_0$ , calibrated such that:

$$w(d_{i,j}) = \exp\left(-\frac{1}{2}\left(\frac{d_{i,j}}{\sigma_0}\right)^2\right),$$

$$\text{where } \sigma_0 = \frac{\max(d_{i,j})}{\sqrt{-2 \ln w_{\min}^{dc}}}, \quad (10)$$

where  $d_{i,j}$  denotes the positional distance between tokens at positions  $i$  and  $j$ . The hyperparameter  $w_{\min}^{dc}$  specifies the minimum allowable attenuation weight at the maximal distance. This formulation ensures rapid decay for nearby positions while retaining a controlled residual weight for distant ones, embodying an interpretable, differentiable, and computationally efficient geometric prior. To adaptively modulate this decay according to semantic context, we incorporate scaling factors from Eq. (9), defining an effective distance:

$$\hat{d}_{i,j} = \frac{d_{i,j}}{\text{scale}_{i,j}}. \quad (11)$$

With a high semantic alignment, the scaling factor can increase, effectively reducing  $\hat{d}_{i,j}$ , softening the decay and allowing semantically important distant token pairs to maintain stronger attention. The distance-aware modulation weight is formulated as:

$$r_{i,j}^{dc} = \exp\left(-\frac{1}{2}\left(\frac{\hat{d}_{i,j}}{\sigma_0}\right)^2\right), \quad (12)$$

which is multiplied by the  $A$ , returning

$$A_{i,j}^{dc} = \lambda_{dc} \cdot A_{i,j} \cdot r_{i,j}^{dc} \quad (13)$$

as the final result.

Generally, DC-DRS aims to achieve a delicate balance, which enforces an inductive prior that encodes a smooth and bounded notion of locality, modulated by contextual relevance, and the local contextual information retains the long-range connectivity recovered by SD-DRS.

### re-Reinforce Distant DRS (reRD-DRS)

Despite the dual guidance of semantic and positional priors, some semantically important but distant token pairs still receive weak attention due to compounded decay effects, as

shown in the two final points of the last curve in Fig. 2. Then we have the final component, reRD-DRS, to further reinforce. It has a target modulation term that is specialized for re-weighting extreme cases, where token pairs with extremely high semantic affinity and significant distance.

We first define a re-weighting coefficient  $r_{i,j}^{re}$  as a reinforcement gate by a rational quadratic function which serves as a heavy-tailed kernel. The rational quadratic function (Rasmussen and Williams 2006) decays slower, allowing stronger reinforcement for long-range dependencies. :

$$r_{i,j}^{re} = \left(1 + \frac{(d_{i,j})^2}{2 \cdot (\sigma_{re} \cdot \text{scale}_{i,j})^2}\right)^{-\alpha}, \quad (14)$$

Specifically, we compute:

$$\alpha = \frac{-\log(w_{\min}^{re})}{\log\left(1 + \frac{d_{\max}^2}{2 \cdot (\sigma_{re} \cdot \text{scale}_{i,j})^2}\right)}. \quad (15)$$

This formulation guarantees that the reinforced attention maintains a lower bound at long range, while remaining sensitive to semantic similarity and contextual scales. It also eliminates the need for manually tuning the decay sharpness, making the model more robust and interpretable. This constraint is formalized as:

$$r_{i,j}^{re} \Big|_{d_{i,j}=d_{\max}, \text{scale}_{i,j}=1} = w_{\min}^{re}, \quad (16)$$

which yields the closed-form expression for  $\sigma_{re}$ :

$$\sigma_{re} = \frac{d_{\max}}{\sqrt{2\alpha(w_{\min}^{re(-1/\alpha)} - 1)}}. \quad (17)$$

This ensures attention to the most distant tokens is smoothly decayed to exactly  $w_{\min}^{re}$ , while nearby tokens retain stronger reinforcement. The final attention logit of reRD-DRS is

$$A_{i,j}^{re} = \lambda_{re} \cdot A_{i,j} \cdot r_{i,j}^{re}. \quad (18)$$

$\lambda_{re}$  is a tunable coefficient controlling the reinforcement strength. This residual formulation ensures that no token pair is prematurely discarded due to positional distance alone, as long as its semantic relevance warrants attention. It introduces a controlled, content-aware reinforcement path for long-range interactions, without disrupting the local attention patterns established earlier. The attenuation profile in reRD-DRS decays smoothly with distance while maintaining a heavy-tail behavior, allowing attention to persist beyond the short-range regime.

The final attention logits of T-DRS are modulated through a residual combination of all three DRS:

$$A_{i,j}^{T-DRS} = A_{i,j} + A_{i,j}^{sd} + A_{i,j}^{dc} + A_{i,j}^{re}. \quad (19)$$

In essence, T-DRS integrates semantic cues and distance-aware priors in a three-stage pipeline to robustly preserve long-range attention in vision-language models.

## Experiment

### Experimental Settings

**Datasets.** We adopt three standard benchmark datasets: ScienceQA-IMG (Lu et al. 2022), GQA (Hudson and Manning 2019), and TextVQA (Ganz et al. 2023) to evaluate

model performance across diverse vision-language reasoning tasks. To further assess the effectiveness of our proposed T-DRS strategy in alleviating long-range dependency degradation, we additionally employ the Positional Object hallucination Prevalence Evaluation (POPE) dataset (Li et al. 2023c), which specifically targets positional hallucination phenomena. ScienceQA-IMG is a curriculum-based benchmark with 21,208 multimodal questions requiring integration of visual and scientific knowledge. GQA offers 22M structured reasoning questions grounded in visual genome images; we use its balanced version for fair evaluation. TextVQA emphasizes OCR-based reasoning over scene text with 28K Q&A pairs from OpenImages. POPE tests factual grounding and spatial awareness by detecting hallucinations in object reasoning. These datasets comprehensively evaluate structural reasoning, and cross-modal understanding. We report standard accuracy for all datasets, with the F1-score added for the POPE dataset.

**Comparison Methods.** Fourteen LVLM approaches i.e., Instruct-BLIP-7B & 13B (Dai et al. 2023), BLIP-2-13B (Li et al. 2023a), Shikra (Chen et al. 2023a), GPT3.5 (Zheng et al. 2023), Ying-VLM (Li et al. 2023b), MiniGPT-4 (Zhu et al. 2023a), Qwen-VL-Chat (Bai et al. 2023), Qwen-VL (Bai et al. 2023), MobileVLM-v2-7B (Chu et al. 2024), Otter (Li et al. 2025a), LLaVA1.5-7B (Liu et al. 2023), InterVL2-8B (Chen et al. 2024), and Qwen2.5-VL-7B (Bai et al. 2025) are chosen for comparison.

**Backbone Models.** To demonstrate the applicability of our method, we plug-in T-DRS into three representative LVLMs: LLaVA1.5-7B, Inter2VL2-8B, and Qwen2.5-VL-7B. These models vary in architecture and multimodal fusion strategies, allowing us to test the robustness and compatibility of T-DRS across various transformer designs.

**Implementation Details.** All models are tested in their original settings without further fine-tuning. T-DRS modules are injected at inference time, require no additional training, and are fully parameter-free except for a small number of fixed hyperparameters ( $w_{min}^{dc}$ ,  $w_{min}^{re}$ ,  $\lambda_{dc}$ ,  $\lambda_{re}$ ), which are shared across all experiments for consistency. All the experiments are done on one NVIDIA A100 GPU.

## Quantitative Comparisons

To evaluate the generality of T-DRS, we integrate it into three diverse LVLMs, differing in scale and architecture. Without any fine-tuning, T-DRS consistently improves performance across ScienceQA, GQA, and TextVQA in Tab. 1, with the improvement of 1.3%, 1.1%, and 0.8% integrating with LLaVA, 0.7%, 0.2%, and 0.6% with InterVL, and 1.3%, 0.4%, and 0.5% with Qwen. demonstrating its plug-and-play flexibility and highlighting long-range attention degradation as a common limitation in current VLMs. Moreover, we evaluate our model on the POPE benchmark, which is specifically designed to expose hallucination errors that arise when models fail to ground their predictions in the actual visual content. The additional analysis provides stronger empirical evidence for the robustness and reliability of T-DRS across diverse reasoning scenarios. To further validate the effectiveness of our proposed method, we conduct comprehensive ablation studies to disentangle the contributions of

Method	QA Datasets			POPE	
	Sci.	GQA	TextVQA	Acc.	F1-score
Instruct-BLIP-7B	60.0	49.2	60.5	70.1	72.3
Instruct-BLIP-13B	63.1	49.5	63.1	71.0	73.0
BLIP-2-13B	61.3	36.4	42.5	65.3	67.2
Shikra	45.8	-	-	63.2	65.8
GPT3.5	72.5	-	-	74.6	76.5
Ying-VLM	55.7	-	-	69.4	71.0
MiniGPT-4	42.3	32.2	-	59.7	62.1
Qwen-VL-Chat	68.2	-	-	75.1	76.9
Qwen-VL	67.1	-	-	76.2	77.1
MobileVLM-v2-7B	61.0	62.6	-	78.0	78.5
Otter	66.3	-	-	71.9	73.7
LLaVA1.5-7B	<u>67.9</u>	<u>62.0</u>	58.2	<u>83.3</u>	<u>85.7</u>
InterVL2-8B	96.6	62.6	79.1	<b>88.0</b>	87.0
Qwen2.5-VL-7B	79.4	57.9	84.5	87.7	86.4
LLaVA1.5. + T-DRS	<b>69.2</b>	<b>63.1</b>	<b>59.0</b>	<b>83.7</b>	<b>86.1</b>
InterVL2. + T-DRS	<b>97.3</b>	<b>62.8</b>	<b>79.7</b>	<b>88.0</b>	<b>87.4</b>
Qwen2.5. + T-DRS	<b>80.7</b>	<b>58.3</b>	<b>85.0</b>	<b>88.5</b>	<b>87.3</b>

Table 1: Accuracy (Acc.%) performance comparison on the ScienceQA-IMG (Sci.), GQA, TextVQA, and POPE datasets. All models operate in a training-free setting. T-DRS is plugged into three distinct LVLMs for evaluation.

*note:* Best and second-best results are **bolded** and underlined.

each component within T-DRS.

## Ablation Studies

To validate the contribution of each of the DRS, we conduct ablation studies in Tab. 2, utilizing the ScienceQA-IMG and POPE dataset, plugging in with the three distinct LVLMs. SD-DRS introduces semantic awareness into attention modulation, enhancing the model’s ability to attend to semantically aligned token pairs regardless of positional separation. However, this may occasionally perturb local structural dependencies, leading to performance fluctuations. To mitigate this, DC-DRS explicitly controls local structural coherence and achieves the second-best performance. Nonetheless, distant token pairs with semantically salient relevance still require enhanced attention recovery, which is addressed by reRD-DRS, specialized for residual long-range attention reinforcement. The integration of all three modules achieves the highest performance, underscoring their complementary roles and validating the effectiveness of the proposed decay-resilient framework.

## Qualitative Analysis

Within the T-DRS framework, there are four hyperparameters  $\{w_{min}^{dc}, \lambda_{dc}, w_{min}^{re}, \lambda_{re}\}$  for sensitivity evaluation shown in Fig. 3. The values of  $\lambda$  are determined following the selection of  $w_{min}$ , and the two hyperparameters in reRD-DRS are configured subsequent to the determination of those in DC-DRS.  $|A|_{min}$  denoted as the minimum value of the attention map within  $A$  in Eq.(5).

$w_{min}^{dc}$  is the minimum attention weight of  $A$  for distant tokens in the decay component,  $\lambda_{dc}$  controls the strength of the distance-aware modulation weight  $r_{dc}$  in Eq.(13),  $w_{min}^{re}$

Configuration		Datasets		
		Sci. Acc.	POPE Acc.	POPE F1-score
Baseline	LLaVA1.5-7B	67.9	83.3	85.7
	InterVL2-8B	96.6	88.0	87.0
	Qwen2.5-VL-7B	79.4	87.7	86.4
w. SD-DRS	LLaVA1.5-7B	68.1	83.4	85.8
	InterVL2-8B	96.9	87.6	86.8
	Qwen2.5-VL-7B	79.8	87.8	86.6
w/o. reRD-DRS	LLaVA1.5-7B	68.8	83.6	86.0
	InterVL2-8B	97.1	87.9	87.2
	Qwen2.5-VL-7B	80.4	88.0	86.9
Full model	LLaVA1.5-7B	69.2	83.7	86.1
	InterVL2-8B	97.3	88.0	87.4
	Qwen2.5-VL-7B	80.7	88.5	87.3

Table 2: Ablation study of T-DRS components on ScienceQA-IMG (Sci.) and POPE, with each abatement item plugged in with LLaVA1.5-7B, InterVL2-8B, and Qwen2.5-VL-7B.

specifies minimum reinforcement weight for distant but semantic relevance tokens, and  $\lambda_{re}$ , a coefficient to further decide the reinforcement strength of semantic relevance. When  $w_{min}^{dc}$  is set to three times  $|A|_{min}$  and  $\lambda$  is fixed at 1, T-DRS (w/o. reRD-DRS) configuration achieves the best performance in Fig. 3(A). Upon further introducing reRD-DRS, setting  $w_{min}^{re}$  to twice  $|A|_{min}$  with  $\lambda_{re}$  set to 0.8 for ScienceQA-IMG and 1 for POPE yields the highest overall performance for the full T-DRS model in Fig. 3(B). A larger  $w_{min}$  increases tolerance to long-range interactions, but may introduce redundant attention to semantically irrelevant tokens. In contrast, an excessively small  $w_{min}$  may over-filter useful nearby tokens, disrupting local semantic coherence.

## Visualization Results

To better understand the effects of distinct stages within T-DRS in the reasoning process, we have the visualizations in Fig. 4. In Fig. 4(A), RoPE-like approaches exhibit long-range decay, with tokens focusing primarily on image peripheries instead of semantically important regions. For instance, when presented with a mammal with furry hair, the model tends to focus on the fur, potentially leading to the incorrect selection of "Painted Stork", which has furry feathers. This strong locality bias risks missing key visual cues at the center of the image relevant to the task. Ultimately, the model may fail to capture the visual semantics needed for accurate cross-modal alignment. With the integration of SD-DRS and DC-DRS in Fig. 4(B), tokens shift toward the semantically important central regions, exhibiting a smooth activation pattern, indicating consistent and focused attention. Under the full model setting in Fig. 4(C), after applying reRD-DRS, attention becomes more focused on semantically meaningful regions, strengthening previously weak long-range token interactions. As a result, T-DRS concentrates on the object's face, correctly recognizing it as a mammal and producing the correct answer. "Red Kangaroo".

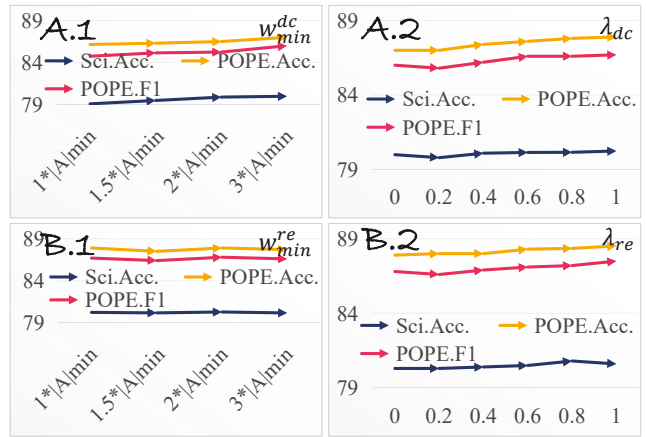


Figure 3: Four hyperparameters are evaluated using ScienceQA-IMG and POPE datasets. The values of  $\lambda$  are determined following the selection of  $w_{min}$ , and the two hyperparameters in reRD-DRS are configured subsequent to the determination of those in DC-DRS. We assign the value of three times  $|A|_{min}$  and 1 to the two hyperparameters in DC-DRS, and twice of  $|A|_{min}$  with set 0.8 for  $\lambda_{re}$  in ScienceQA-IMG, and 1 for POPE in reRD-DRS.

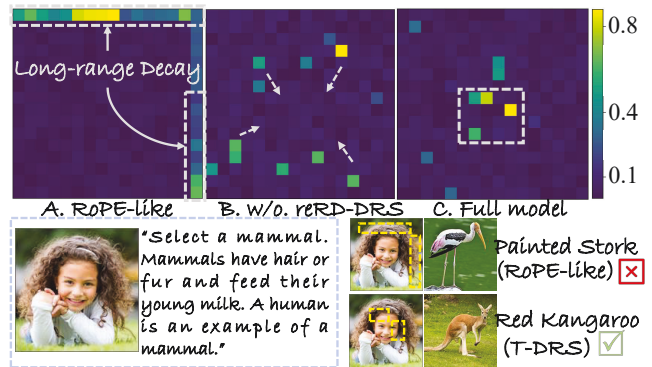


Figure 4: Visualization results of cross-attention of  $S$  across different stages. (A) RoPE-only (A). (B) With SD-DRS and DC-DRS ( $A + A^{sd} + A^{dc}$ ), attention begins to converge towards the central focus. (C) Under the full model setting ( $A^{T-DRS}$ ), the model can further reinforce informative long-range connections.

## Conclusion

In this paper, we propose Three-stage Decay-Resilient Strategies (T-DRS), a three-step inference-time framework to alleviate the long-range attentions decay in Large Vision-Language Models (LVLMs). T-DRS integrates the Semantic Driven DRS (SD-DRS), recovering long-range dependencies, Distance-aware Control DRS (DC-DRS), smoothing local structure that perturbed by SD-DRS, and re-Reinforce Distant DRS (reRD-DRS) specialized for reweighting tokens that are distant but with rich semantics. Experimental results demonstrate that T-DRS can be seamlessly applied to LVLM backbones and consistently improves performance.

## Acknowledgments

The work is supported by Guangdong and Hong Kong Universities' "1+1+1" Joint Research Collaboration Scheme (2025A0505000003), the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College (2022B1212010006), Guangdong Key Lab of AI and Multimodal Data Processing (2020KSYS007), and the internal grant of BNBU (R6025A, UICR0300019).

## References

- Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4971–4980.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.
- An, X.; Xie, Y.; Yang, K.; Zhang, W.; Zhao, X.; Cheng, Z.; Wang, Y.; Xu, S.; Chen, C.; Wu, C.; et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Barbero, F.; Vitvitskiy, A.; Perivolaropoulos, C.; Pascanu, R.; and Veličković, P. 2024. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*.
- Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.
- Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.
- Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.
- Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.
- Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19358–19369.
- Ganz, R.; Nuriel, O.; Aberdam, A.; Kittenplon, Y.; Mazor, S.; and Litman, R. 2023. Towards models that can see and read. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21718–21728.
- Gao, P.; Lee, Y.; Liu, X.; Zhang, H.; Chen, Z.; Hu, Y.; Jing, G.; and Lai, Y. 2025. Contextual Reasoning for Robust Composed Image Retrieval with Vision-Language Models. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 349–357.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.
- Lee, Y.; Gao, P.; Chen, Z.; Fan, W.; Jing, G.; and Hu, Y. 2025a. Boosting Audio-Visual Segmentation via Triple-Modalities Alignment. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.
- Lee, Y.; Gao, P.; Xu, Y.; and Fan, W. 2025b. How Do Optical Flow and Textual Prompts Collaborate to Assist in Audio-Visual Semantic Segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 23342–23352.
- Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Cahyono, J. A.; Yang, J.; Li, C.; and Liu, Z. 2025a. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Li, H.; Qin, Y.; Ou, B.; Xu, L.; and Xu, R. 2025b. HoPE: Hybrid of Position Embedding for Length Generalization in Vision-Language Models. *arXiv preprint arXiv:2505.20444*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.
- Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; et al. 2023b. M<sup>3</sup>IT: A Large-

- Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv preprint arXiv:2306.04387*.
- Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Rasmussen, C. E.; and Williams, C. K. 2006. *Gaussian Processes for Machine Learning*. MIT Press.
- Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.
- Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.
- Tang, F.; An, X.; Yang, H.; Xie, Y.; Yang, K.; Hu, M.; Cheng, Z.; Zhou, X.; Ran, Z.; Razzak, I.; et al. 2025a. UniViT: Unifying Image and Video Understanding in One Vision Encoder. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Tang, F.; Huang, Z.; Liu, C.; Sun, Q.; Yang, H.; and Lim, S.-N. 2025b. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*.
- Tang, F.; Liu, C.; Xu, Z.; Hu, M.; Huang, Z.; Xue, H.; Chen, Z.; Peng, Z.; Yang, Z.; Zhou, S.; et al. 2025c. Seeing Far and Clearly: Mitigating Hallucinations in MLLMs with Attention Causal Decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26147–26159.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wei, X.; Liu, X.; Zang, Y.; Dong, X.; Zhang, P.; Cao, Y.; Tong, J.; Duan, H.; Guo, Q.; Wang, J.; et al. 2025. VideoRoPE: What Makes for Good Video Rotary Position Embedding? *arXiv preprint arXiv:2502.05173*.
- Xing, Y.; Li, Y.; Laptev, I.; and Lu, S. 2024. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37: 92012–92035.
- Xue, H.; Tang, F.; Hu, M.; Liu, Y.; Huang, Q.; Li, Y.; Liu, C.; Xu, Z.; Zhang, C.; Feng, C.-M.; et al. 2025. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*.
- Yang, H.; Tang, F.; Zhao, L.; An, X.; Hu, M.; Li, H.; Zhuang, X.; Wang, B.; Lu, Y.; Zhang, X.; et al. 2025. StreamAgent: Towards Anticipatory Agents for Streaming Video Understanding. *arXiv preprint arXiv:2508.01875*.
- Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Zhao, Q.; Zhang, X.; Li, Y.; Xing, Y.; Yuan, X.; Tang, F.; Fan, S.; Chen, X.; Zhang, X.; and Wang, D. 2025. MCA-LLaVA: Manhattan Causal Attention for Reducing Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2507.09184*.
- Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Zhu, S.; Ye, J.; Jiang, W.; Xue, S.; Zhang, Q.; Wu, Y.; and Li, J. 2023b. CoCA: Fusing Position Embedding with Collinear Constrained Attention in Transformers for Long Context Window Extending. *arXiv preprint arXiv:2309.08646*.
- Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024a. Enhancing Zero-Shot Vision Models by Label-Free Prompt Distribution Learning and Bias Correcting. In *NeurIPS*.
- Zhu, X.; Zhu, B.; Tan, Y.; Wang, S.; Hao, Y.; and Zhang, H. 2024b. Selective Vision-Language Subspace Projection for Few-shot CLIP. In *ACM Multimedia*, 3848–3857. ACM.