VisReflect: Latent Visual Reflection for Fine-Grained Perception in Long Visual Context

Xiaoqian Shen Mohamed Elhoseiny

King Abdullah University of Science and Technology, Saudi Arabia
{xiaoqian.shen,mohamed.elhoseiny}@kaust.edu.sa
https://xiaoqian-shen.github.io/VisReflect

Abstract. Large Vision Language Models (LVLMs) have achieved remarkable success on vision–language tasks, yet fine-grained perception over high-resolution images and long-context videos remains challenging. As the number of visual tokens increases, the visual attention sink phenomenon becomes increasingly severe, causing irrelevant tokens to absorb a disproportionate amount of attention mass. Recent approaches attempt to mitigate this issue by explicitly predicting bounding boxes or temporal spans and re-encoding the cropped visual regions. Such methods depend on unreliable numeric localization in the discrete token space and incur significant computational overhead due to additional forward passes. In this work, we propose **VisReflect**, a simple yet effective framework that improves fine-grained perception in long visual contexts through latent visual reflection. Instead of decoding intermediate predictions into discrete tokens, the model generates continuous visual reflection that represents question-relevant visual features in the latent space. These reflections selectively emphasize salient regions or frames, guiding attention towards relevant visual tokens within a single forward pass. We conduct comprehensive evaluations on challenging high-resolution image benchmarks, including BLINK, V*, and HRBench-4K/8K, as well as video understanding benchmarks such as MVBench, VideoMME, and MLVU. Our method consistently improves over strong baselines, achieving gains of 4.1% on image benchmarks and 1.8% on video benchmarks. Compared with zooming-based methods, our model achieves comparable performance while reducing inference time by roughly 44% on video understanding.

Keywords: Large Vision Language Model· Fine-Grained Perception · Long Video Understanding

1 Introduction

Large Vision Language Models (LVLMs) have recently achieved remarkable progress on advanced vision–language tasks, demonstrating strong performance in visual question answering and video reasoning. By integrating powerful language models with visual encoders, the models are able to reason over multimodal inputs and exhibit impressive capabilities in complex perception and

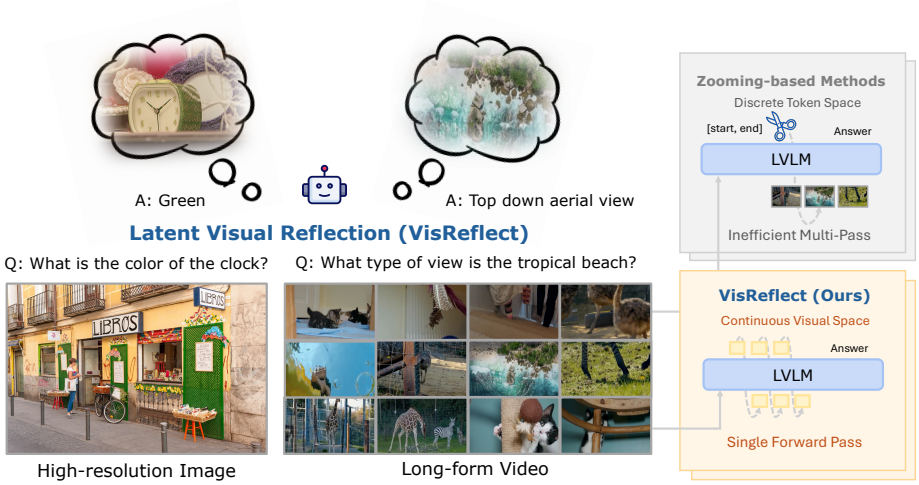

Fig. 1: Existing zooming-based methods localize relevant regions or frames by predicting coordinates in discrete token space and performing repeated forward passes. In contrast, VisReflect generates latent visual reflection tokens in continuous visual space, enabling the model to internally recall question-relevant visual features within a single forward pass.

reasoning scenarios. However, despite these successes, a persistent challenge remains: fine-grained perception in long visual contexts.

When processing high-resolution images, LVLMs must handle thousands of visual tokens to achieve fine-grained perceptual understanding. This challenge is further amplified in long-context video reasoning, where visual token sequences grow substantially. Ideally, the model should selectively attend to visual information that is most relevant to the textual query. However, as the number of visual tokens increases, the *Visual Attention Sink* [14] phenomenon becomes more severe, causing irrelevant tokens to absorb a disproportionate amount of attention mass. Consequently, the model struggles with small-object perception in high-resolution images or temporally localized reasoning in long-form videos.

To address this challenge, recent studies have introduced the "*Thinking-with-Images/Videos*" paradigm [11, 26, 30, 44, 45], which aims to preserve global context while capturing fine-grained details in long visual sequences. Specifically, these methods prompt the model to predict the bounding box of a region of interest or the temporal span of salient frames. Based on these predictions, a second fine-grained forward pass is performed, where visual tools are invoked to crop or zoom into the selected image regions or video segments, and the refined inputs are re-encoded for detailed understanding. Although this paradigm yields performance improvements, it exposes two critical limitations. **(i) Reliability of numeric predictions.** The model generates deterministic tokens to represent bounding box coordinates or temporal intervals, which are expressed purely in the textual space and do not faithfully model the underlying visual feature space.

As a result, the second forward pass is highly sensitive to localization errors. **(ii) Increased computational cost due to additional prefilling.** Iterative cropping or zooming requires the visual features to be re-forwarded to the language model, substantially increasing inference latency and computational overhead.

These limitations lead us to rethink, how should LVLMs retrieve relevant visual evidence from massive token sequences. Instead of explicitly zooming into pixel-space regions through discrete token predictions, *can a model directly recall relevant visual representations within its latent space?*

Motivated by this perspective, we propose **VisReflect**, a simple yet effective framework that enhances fine-grained perception in high-resolution images and long-context videos without explicit coordinate-based cropping or temporal zooming. Instead of predicting discrete localization tokens, it generates continuous visual reflection embeddings that recall the visual features of salient regions or frames in the latent space. These reflection embeddings guide attention toward informative visual tokens, enabling the model to focus on relevant regions or frames within a single forward pass. As a result, our approach provides a computationally efficient alternative to multi-pass zooming strategies, offering a unified solution for long-context visual understanding.

We evaluate our method on challenging high-resolution image benchmarks, including BLINK [12], V*, and HRBench [36], as well as on long-video benchmarks such as VideoMME [10] and MLVU [46], which demand precise small-object perception and temporally localized segment reasoning. Experimental results demonstrate an average improvement of 4.1% over the base model on image benchmarks and a 1.8% gain over strong baselines on video benchmarks. Moreover, our VisReflect achieves comparable performance than multi-pass "thinking-with-video" approaches while reducing inference time by approximately 44%.

Contribution. Our contribution is summarized as follows:

- We propose **VisReflect**, a novel framework that reflects question-relevant visual information from long visual contexts directly in the continuous embedding space, avoiding unreliable numeric localization and costly multi-pass inference required by prior tool-using methods.
- We analyze the attention maps from the visual reflection tokens to the original visual inputs. The reflected visual representations guide the model to focus its attention on informative regions or frames within large visual contexts, mitigating the attention sink caused by massive visual tokens.
- Comprehensive experiments demonstrate that VisReflect improves fine-grained perception in both high-resolution images and long-form videos, achieving gains of **+4.1%** and **+1.8%**. It achieves comparable performance to zooming-based methods that require multiple forward passes, while reducing video inference time by **44%**.

2 Related Work

2.1 Thinking with Images

Large vision language models (LVLMs) [23, 47] have recently achieved strong performance across a wide range of visual understanding and reasoning tasks. Building upon these foundations, several studies [4,5,27,28] enhance spatial perception by enabling models to predict explicit coordinate outputs, allowing them to localize small or fine-grained objects within images. More recently, research on fine-grained multimodal reasoning has moved toward a "Thinking-with-Images" paradigm, in which models dynamically retrieve localized visual evidence during inference rather than depending exclusively on a single global image representation [15, 35, 38, 42, 44, 45]. In this context, DeepEyes [45] and Mini-o3 [15] employ reinforcement learning to incentivize the usage of visual tools such as zooming-in, cropping or searching, while Thyme [44] and PixelReasoner [35] enable models to generate code or perform pixel-space operations to manipulate visual inputs dynamically. Although effective, this paradigm incurs substantial inference overhead due to repeated visual encoding passes or iterative tool-calling loops, leading to increased latency. Moreover, these methods are constrained to predict numeric bounding box coordinates in the textual space, which do not truly reflect the underlying visual features. In our work, we encourage the model to perform latent visual reflection by generating visual representations in the LLM's output space that align with past salient region features. This helps the model re-attend to relevant information within the massive visual token sequence of high-resolution images, without explicit cropping during inference that would otherwise increase computational cost.

2.2 Thinking with Videos

Large vision-language models process videos by extracting and encoding frames, and then concatenating them into final video representations. To address massive number of visual tokens for long videos, several works train on sparsely sampled frames [1, 8, 19], while others try to handle long videos by token pooling [22,25,34], token compression [31] or segment retrieval [2,32]. More recently, a new paradigm "*Thinking-with-Video*" has emerged [11, 26, 30, 41]. The model first generates intermediate temporal cues to identify query-relevant segments, and then performs a second forward pass by "zooming in" on these salient frames. This test-time scaling strategy effectively preserves global context while refining local details through re-inference. Despite its effectiveness, the reliance on multiple forward passes substantially increases inference latency and computational cost. In contrast, we propose a latent visual reflection mechanism that achieves comparable performance improvements within a single forward pass.

2.3 Latent Reasoning

Latent reasoning [7,13,17,33] performs intermediate reasoning entirely in the latent space, without explicitly decoding intermediate steps into natural language.

Instead of generating and appending reasoning tokens, the model performs reasoning by recursively using the latent state from the previous iteration as the input embedding to the next, and returns to text space only at the final step to generate the answer. This paradigm, instead of converting continuous embeddings into a deterministic token space, enabling reasoning beyond language and exploiting rich, high-dimensional representations. In this work, we leverage the continuous latent embeddings to align with salient visual features, enhancing the model's perceptual capabilities in long visual contexts such as high-resolution images or long video sequences.

3 Method

3.1 Preliminaries

A Large Vision-Language Model (LVLM) consists of three components: (i) a vision encoder $\mathcal{V}$, (ii) a projection module $\mathcal{P}$ that maps visual representations into the language embedding space, and (iii) a large language model $\mathcal{M}$.

Let $\mathbf{I}$ denote an input image and $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$ denote a video with T frames. Given a visual input (image or video) together with a textual query $\mathbf{Q} = \{q_i\}_{i=1}^{L_q}$ of length L_q, the vision encoder extracts visual tokens, which are projected into the language embedding space as:

$$\mathbf{Z}_v = \mathcal{P}(\mathcal{V}(\mathbf{I})) \quad \text{or} \quad \mathbf{Z}_v = \mathcal{P}(\mathcal{V}(\mathbf{V})). \tag{1}$$

Here, $\mathbf{Z}_v \in \mathbb{R}^{N \times d}$ denotes N visual token embeddings of dimension d. For video inputs, $N = T \times N_s$, where N_s is the number of spatial tokens per frame.

The textual query is embedded via a token embedding function $\mathcal{E}(\cdot)$:

$$\mathbf{Z}_q = \{\mathcal{E}(q_i)\}_{i=1}^{L_q}, \qquad \mathbf{Z}_q \in \mathbb{R}^{L_q \times d}. \tag{2}$$

The concatenated sequence $[\mathbf{Z}_v; \mathbf{Z}_q]$ is fed into the language model $\mathcal{M}$, which autoregressively generates an answer sequence $\mathbf{Y} = \{y_i\}_{i=1}^{L_y}$:

$$p(\mathbf{Y} \mid \mathbf{I}, \mathbf{Q}) = \prod_{i=1}^{L_y} p(y_i \mid y_{<i}, \mathbf{Z}_v, \mathbf{Z}_q). \tag{3}$$

Let $\mathbf{h}_i \in \mathbb{R}^d$ denote the last-layer hidden state of $\mathcal{M}$ at decoding step i. The token distribution is computed through a language modeling head $\mathcal{W}_{\text{LM}}$:

$$p(y_{i+1} \mid y_{\leq i}) = \text{Softmax}(\mathcal{W}_{\text{LM}} \mathbf{h}_i). \tag{4}$$

Limitation. When the visual input are high-resolution images or long videos, the number of visual tokens N becomes large. The self-attention mechanism allocates attention across all tokens, reducing effective focus on query-relevant regions and impairing fine-grained perception.

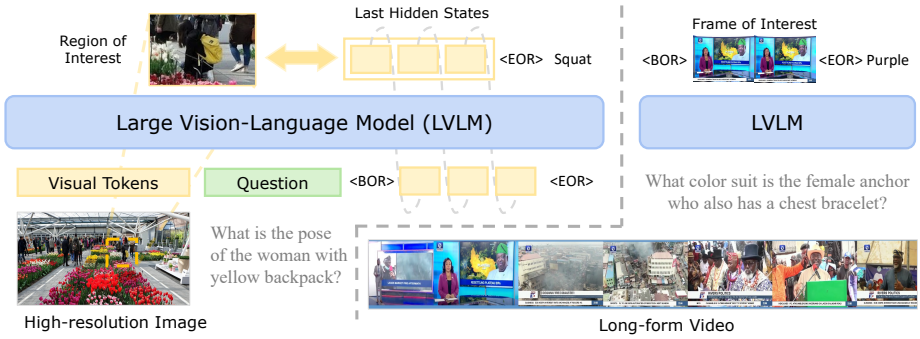

Fig. 2: Overview of VisReflect. Given visual tokens extracted from a high-resolution image or a long-form video together with a textual query, a LVLM generates a sequence of visual reflection tokens between the special tokens `<BOR>` and `<EOR>`. These tokens are trained to approximate the latent visual representations of the region of interest (for images) or frames of interest (for videos) that are relevant to the question. Instead of explicitly cropping images or re-encoding selected frames, the model recalls question-relevant visual features directly in the latent embedding space using its last hidden states. The reflected visual features then guide the model's attention toward informative parts of the visual context, enabling accurate fine-grained perception in high-resolution images and long-form videos while requiring only a single forward pass.

Existing grounding approaches predict explicit coordinates, such as bounding boxes $[x_1, y_1, x_2, y_2]$ for images or temporal intervals $[t_s, t_e]$ for videos. However, these predictions are constrained to the discrete token space and therefore do not faithfully capture the underlying visual feature space. Moreover, such methods typically require cropping and re-encoding the selected region in an additional forward pass, increasing computational overhead and breaking strict end-to-end differentiability.

3.2 VisReflect

To overcome these limitations, we introduce a *latent visual reflection* mechanism that enables the model to reflect question-relevant visual information from long visual contexts before producing the final answer. Concretely, the model generates a sequence of latent visual embeddings that reside in the same embedding space as the visual input, allowing them to directly represent salient regions or frames within the visual context. Unlike prior approaches that rely on predicting numeric coordinates for cropping or zooming, our method reflects the relevant visual content in the continuous embedding space within a single forward pass.

More specifically, we add three special tokens to the vocabulary: `<BOR>` (beginning of reflection), `<VR>` (visual reflection token), and `<EOR>` (end of reflection). These tokens are designed to trigger the visual reflection mode within LVLMs, guiding the model to internally generate embeddings that reflect the

relevant visual features, effectively recalling past visual content within the latent space.

Visual Reflection for Image Understanding. Let $\mathbf{Z}_v = \{\mathbf{z}_n\}_{n=1}^{N}$ denote the spatial visual tokens extracted from an input image, where $\mathbf{z}_n \in \mathbb{R}^d$.

Given a bounding box annotation, we identify the indices of the K visual tokens that fall inside the region of interest (ROI), denoted as $i_1, i_2, \ldots, i_K$, where K is the number of tokens within the ROI.

During training, the model generates a reflection sequence:

$$\texttt{<BOR>}\ \texttt{<VR>}_1\ \texttt{<VR>}_2\ \ldots\ \texttt{<VR>}_K\ \texttt{<EOR>}.$$

Let $\mathbf{h}_k^{\mathrm{VR}} \in \mathbb{R}^d$ denote the last-layer hidden state corresponding to the k-th <VR> token. We enforce a one-to-one alignment between reflection embeddings and ROI visual tokens by maximizing their cosine similarity:

$$\mathcal{L}_{\mathrm{VR}}^{\mathrm{img}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\mathbf{h}_k^{\mathrm{VR}} \cdot \mathbf{z}_{i_k}}{\|\mathbf{h}_k^{\mathrm{VR}}\|_2\,\|\mathbf{z}_{i_k}\|_2}\right) \tag{5}$$

Intuitively, the generated visual reflection embeddings act as latent anchors. Since these embeddings approximate visual tokens corresponding to salient regions, they naturally propagate attention mass from the reflection tokens toward the associated visual tokens. As a result, the model implicitly re-focuses its attention distribution on question-relevant subsets of the visual context, alleviating the attention sink caused by large numbers of visual tokens.

Visual Reflection for Video Understanding. For a video input, the encoder produces frame-wise spatial tokens:

$$\mathbf{Z}_v = \{\mathbf{Z}^{(t)}\}_{t=1}^{T}, \qquad \mathbf{Z}^{(t)} = \{\mathbf{z}_n^{(t)}\}_{n=1}^{N_s}, \tag{6}$$

where N_s denotes the number of spatial tokens per frame. Given temporal boundaries $[t_s, t_e]$, we uniformly sample J frames with indices $t_1, t_2, \ldots, t_J$, where each $t_j \in [t_s, t_e]$.

For each sampled frame, we recall all spatial tokens. Therefore, the model generates $J \times N_s$ tokens for visual reflection.

$$\texttt{<BOR>}\ \texttt{<VR>}_1\ \texttt{<VR>}_2\ \ldots\ \texttt{<VR>}_{J\times N_s}\ \texttt{<EOR>}.$$

The salient frame feature alignment is defined as:

$$\mathcal{L}_{\mathrm{VR}}^{\mathrm{vid}} = \frac{1}{J \times N_s}\sum_{j=1}^{J}\sum_{n=1}^{N_s}\left(1 - \frac{\mathbf{h}_{(j-1)\times N_s+n}^{\mathrm{VR}} \cdot \mathbf{z}_n^{(t_j)}}{\|\mathbf{h}_{(j-1)\times N_s+n}^{\mathrm{VR}}\|_2\,\|\mathbf{z}_n^{(t_j)}\|_2}\right). \tag{7}$$

Language Mode v.s. Visual Reflection Mode. The model operates in two modes. In text generation mode, the last hidden states are projected to vocabulary logits through the language model's head, the embedding of the next input token is then set to the embedding of the generated token:

$$p(y_{i+1} \mid y_{\leq i}) = \text{Softmax}(\mathcal{W}_{\text{LM}}\mathbf{h}_i), \quad \mathbf{E}_{i+1} = \mathcal{E}(y_{i+1}). \tag{8}$$

In visual reflection mode, we use the last hidden state from the previous token to replace the input embedding:

$$\mathbf{E}_{i+1} = \mathbf{h}_i^{\text{VR}}, \quad \text{if } y_i = \texttt{<VR>} \tag{9}$$

The final answer is then generated conditioned on the continuous visual reflection, which emphasizes the regions or frames most relevant to the query, implicitly recalling fine-grained information from the long visual context.

Training Objective. The overall objective combines language modeling and visual alignment, $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\mathcal{L}_{\text{VR}}$, where $\mathcal{L}_{\text{LM}}$ is the standard cross-entropy loss for answer generation, $\mathcal{L}_{\text{VR}} \in \{\mathcal{L}_{\text{VR}}^{\text{img}}, \mathcal{L}_{\text{VR}}^{\text{vid}}\}$ depends on modality, and λ balances linguistic supervision and latent visual alignment.

The entire framework is trained end-to-end in a single forward pass, enabling computational efficiency and enhancing attention to relevant regions in high-resolution images and long-form videos. Localization annotations are used only as a training signal, and the model remains fully coordinate-free during inference.

Inference. During inference, we append `<BOR>` to the end of the question prompt. Then the model naturally generates `<VR>` tokens following the training format. The hidden state of `<VR>` token will replace its token embedding for the next decoding step. Once the number of generated `<VR>` tokens reaches the limit, we force the next token to be `<EOR>` to terminate the reflection process.

4 Experiments

4.1 Experimental Setup

Benchmarks and Metrics. We conduct comprehensive evaluations on both high-resolution image benchmarks and long video understanding benchmarks.

Image Understanding. We evaluate our method on BLINK [12], HRBench-4K [36], HRBench-8K [36], and V* [39]. BLINK [12] contains perception-demanding tasks such as multi-view reasoning and relative depth estimation. V* [39] and HRBench [36] focus on fine-grained attribute recognition and spatial reasoning under high-resolution settings, where target objects typically occupy only a small fraction of the image. Such setups pose significant challenges due to the large number of visual tokens and the need for precise regional perception.

Video Understanding. For long-video understanding, we evaluate our method on MLVU [46], MVBench [20], and VideoMME [10]. These benchmarks require temporal reasoning over extended contexts, testing the model's ability to identify relevant actions or events within long video sequences. We report multi-choice accuracy as the evaluation metric.

Baselines. We adopt Qwen-2.5-VL-7B [3] as the backbone and primary baseline, and further compare our approach against several recent state-of-the-art models for image and video understanding.

Image Understanding. For image-level evaluation, we compare our method with recent visual reasoning approaches, including PixelReasoner [35], PAPO [37], Thyme-VL [44], and DeepEyes [45]. Notably, Thyme-VL [44] and DeepEyes [45] follow a "thinking-with-images" paradigm, which involves cropping the image and performing re-inference on selected regions, thereby requiring multiple forward passes. For fair comparison, we additionally report their performance without tool usage (i.e., single-pass inference), evaluating the models' standalone capabilities without additional inference.

Video Understanding. For video understanding, we compare against strong video-language models, including LLaVA-OneVision [18], LongVU [31], and VideoChat-R1 [21]. Furthermore, we evaluate against "thinking-with-video" approaches that first perform temporal grounding with respect to the query, then zoom into salient frames for a second-stage inference. This category includes LOVE-R1 [11], VideoChat-R1.5 [41], and TimeSearch-R [26]. These methods similarly rely on multi-stage processing, in contrast to our single-pass visual reflection framework.

Implementation Details. We adopt Qwen-2.5-VL-7B [3] as the backbone model. Training contains an image stage followed by a video stage.

Stage 1: Image Training. We train on image data with the maximum input resolution set to $4096 \times 28 \times 28$ pixels and the minimum resolution set to $128 \times 28 \times 28$ pixels, corresponding to a visual token range of $128 \leq N \leq 4096$. We use the VisualCoT [29] as image training data and filter the training samples to retain only those whose bounding box area is smaller than 10% of the entire image.

Stage 2: Video Training. We initialize from the image-stage checkpoint and continue training on video data. The maximum number of visual tokens is set to 16,384, while the maximum pixel resolution per frame is constrained to $128 \times 28 \times 28$. We sample $J = 3$ visual reflection frames from the annotated temporal span. The training datasets include QVHighlights [16], ActivityNet [43], NExT-GQA [40], and the multiple-choice split of PLM-Video [9]. We further restrict the temporal clue span to be shorter than 10 seconds to emphasize localized reasoning.

Models	BLINK	HRBench-4K	HRBench-8K	V^*		
				Attribute	Spatial	Overall
Multiple Prefilling						
Thyme-VL [44]	-	77.0	72.0	83.5	80.3	82.2
DeepEyes [45]	-	75.1	72.6	91.3	88.2	90.1
Single Prefilling						
Qwen2.5-VL [3]	56.5	70.9	66.9	80.8	76.3	79.1
PixelReasoner [35]	56.1	71.2	64.2	81.7	76.3	79.6
PAPO [37]	55.6	70.5	69.1	80.8	77.6	79.5
Thyme-VL† [44]	55.2	72.6	67.8	81.7	78.9	80.6
DeepEyes† [45]	56.5	71.2	68.3	82.6	76.3	80.1
LVR [17]	55.3	71.4	67.2	81.7	79.0	80.6
VisReflect (Ours)	**57.1**	**73.8**	**70.1**	**86.7**	**82.3**	**84.9**

Table 1: Performance on image understanding benchmarks. † denotes a single prefilling without cropping. Their default tool-using results are marked in gray.

In both stages, the visual encoder and multimodal projector are kept frozen, and only the parameters of the language model are updated. The learning rate is set to 1×10^{-5} with a warmup ratio of 0.03. We set $\lambda = 0.1$ to weight the alignment loss. All models are trained on 8 NVIDIA H100 GPUs (80GB).

4.2 Visual Reflection for Fine-Grained Image Perception

Table 1 demonstrates the performance for fine-grained image perception.

Comparison with the backbone model. Compared with the backbone model Qwen2.5-VL, our method achieves consistent improvements across all benchmarks, including BLINK (+0.6), HRBench-4K (+2.9), HRBench-8K (+3.2), and V^* (+5.8). These gains indicate that visual reflection effectively enhances the model's ability to focus on question-relevant regions in high-resolution images.

Comparison with state-of-the-art baselines. Our method outperforms the strongest runner-up (Thyme-VL†) on HRBench-4K (+1.2), HRBench-8K (+2.3), and on V^* (+4.3), demonstrating the advantage of reflecting visual information in the latent space for fine-grained perception.

Comparison with region-cropping methods. Our method approaches the performance of multi-pass cropping-based models such as Thyme-VL and DeepEyes. This highlights that latent visual reflection can serve as an effective alternative to explicit zoom-in operations for long visual context understanding while reducing computational overhead.

Models	MLVU	MVBench	VideoMME (w/o & w/ sub.)
LLaVA-OneVision [18]	64.7	56.7	58.2 / 61.5
LongVU [31]	65.4	66.9	60.6 / 63.7
LongVILA [6]	67.0	67.1	60.1 / 65.1
NVILA [24]	70.1	68.1	64.2 / 70.0
VideoChat-R1 [21]	69.5	67.9	64.3 / 69.1
Qwen2.5-VL [3]	69.8	68.4	65.2 / 70.7
VisReflect (Ours)	**70.6**	**70.4**	**67.4** / **72.7**

Table 2: Performance on video understanding benchmarks.

Models	MLVU		MVBench		VideoMME (w/o sub.)	
	Acc↑	Time↓	Acc↑	Time↓	Acc↑	Time↓
Multiple Prefilling						
LOVE-R1 [11]	67.4	32.5s	66.6	1.4s	66.2	29.8s
VideoChat-R1.5-M [41]	70.9	37.1s	70.6	1.7s	67.1	30.0s
TimeSearch-R [26]	71.5	32.8s	69.8	1.6s	66.6	28.5s
Single Prefilling						
VisReflect (Ours)	70.6	18.5s	70.4	0.9s	67.3	16.5s

Table 3: Comparison with zooming-based video models. Our method achieves comparable performance to zooming-based approaches that require multiple forward passes, while reducing inference time by 44%.

4.3 Visual Reflection for Long Video Understanding

Long videos pose a more severe challenge: as the number of frames increases, attention over all tokens becomes increasingly distracting, making it difficult for models to focus on the most salient moments.

Comparison with state-of-the-art baselines. As reported in Table 2, our approach consistently outperforms strong baselines on long video understanding benchmarks, achieving gains of +0.8 on MLVU [46], +2.0 on MVBench [20], and +2.1/+2.0 on VideoMME [10] w/o or with subtitles. These improvements demonstrate that visually reflecting on target frames enables the model to preserve global context while capturing critical detail within a single forward pass.

Comparison with frame-zooming methods. Notably, experimental results in Table 3 show that our latent visual reflecting achieves performance comparable to "Thinking-with-Video" approaches that require multiple forward passes by zooming into salient frames, while substantially reducing inference latency.

4.4 Ablation Studies

Effect of Visual Reflection. Table 4 presents the contribution of visual reflection. Fine-tuning Qwen2.5-VL alone yields only limited improvements, while incorporating visual reflection significantly boosts performance. We further compare mean squared error (MSE) and cosine similarity (COS) as the alignment loss, and observe that COS provides more stable training and consistently better performance.

Models	BLINK	HRBench-4K	HRBench-8K	V^*
Qwen2.5-VL [3]	56.5	70.9	66.9	79.1
Qwen2.5-VL Finetune	55.2	71.3	65.9	80.1
+ Visual Reflection (MSE)	56.4	72.8	69.3	83.2
+ Visual Reflection (COS)	57.1	73.8	70.1	84.9

Table 4: Effect of visual reflection and alignment loss choice. We compare different loss functions for aligning the generated visual reflection tokens with target visual features.

Number of <VR> Tokens during Training. We study how the number of generated <VR> tokens affects model performance. During training, we control the maximum number of visual reflection tokens that the model is required to generate. As shown in Table 5, increasing the number of reflection tokens from 16 to 128 consistently improves performance across all benchmarks. However, further increasing the number of tokens beyond this point leads to marginal gains or slight degradation. Excessively long reflection sequences might introduce redundancy and make optimization more difficult.

# <VR>	16 tokens	32 tokens	64 tokens	128 tokens	196 tokens	256 tokens
V^* [39]	81.6	81.2	82.2	**84.9**	83.2	83.7
HRBench-4K [36]	71.4	72.5	73.3	**73.8**	73.5	73.7
HRBench-8K [36]	67.8	68.4	69.9	70.1	**70.3**	70.1

Table 5: Number of <VR> tokens during training. We control the maximum number of visual reflection tokens that the model is required to generate.

Number of <VR> Tokens during Inference. We modify the inference procedure by fixing the number of visual reflection tokens (number of <VR>) and explicitly appending the <EOR> token to terminate the reflection phase. The <EOR> token serves as a trigger that forces the model to transition to the final-answer generation stage.

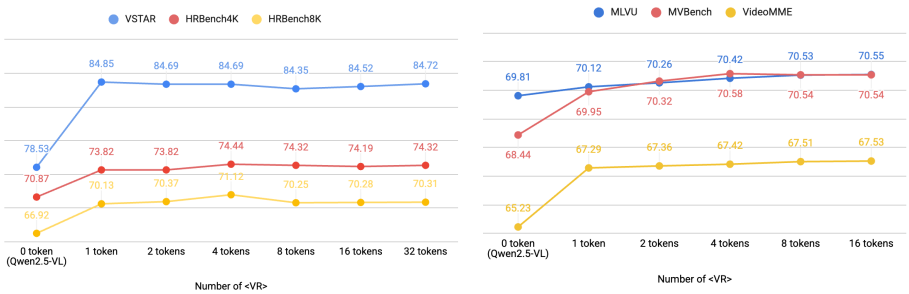

Fig. 3: Number of <VR> during inference. We vary the number of generated visual reflection tokens during inference to evaluate how reflection length influences model performance.

As shown in Fig. 3, generating one <VR> token already leads to considerable improvements over the backbone Qwen2.5-VL. For image benchmarks, increasing the number of tokens brings limited additional benefit. This suggests that one visual reflection token is sufficient to guide attention toward the relevant small regions in high-resolution images. Therefore, we set the number of <VR> tokens as one for image evaluation.

For video benchmarks, we observe a gradual improvement as the number of reflection tokens increases. However, the gains become marginal when further increasing the token number. Considering the trade-off between performance and computational cost, we use four <VR> tokens for video evaluation.

Attention refocusing, not visual reconstruction. Although training employs multiple reflection tokens to encourage comprehensive alignment between reflection embeddings and visual features, we find that only a small number of tokens are necessary at inference time, which effectively guides the model's attention toward question-relevant visual content. This observation suggests that the visual reflection mechanism primarily acts as an attention refocusing signal, steering the model toward salient regions in images or informative frames in videos, rather than requiring dense reconstruction of the entire set of visual tokens.

4.5 Analysis

Inference Efficiency. For image VQA, inference is dominated by LLM decoding, making single-pass methods comparable in runtime. We report per-image latency in Table 6. VisReflect remains efficient with a single reflection token while achieving stronger performance.

Table 3 compares our method with zooming-based video models that require multiple prefilling. "Time" denotes the average inference time per video. Our method achieves comparable accuracy while significantly reducing inference time, cutting latency by 44% across benchmarks. This demonstrates that

our proposed visual reflection provides a computationally efficient alternative to multi-pass zooming strategies.

Model	PixelReasoner	PAPO	DeepEyes (S)	DeepEyes (M)	Thyme (S)	Thyme (M)	Ours
Seconds	3.04	2.51	2.53	4.78	2.74	4.81	2.60

Table 6: Inference speed on image understanding. S denotes single-prefilling (forward pass), while M denotes multiple-prefilling (forward pass).

Attention quantification. We define the ROI attention ratio as $\frac{\frac{1}{|\mathrm{ROI}|}\sum_{i\in\mathrm{ROI}} a_i}{\frac{1}{N}\sum_{i=1}^{N} a_i}$, where a_i denotes the attention weight of visual token i, and N is the total number of visual tokens. A ratio greater than 1 indicates that attention is more concentrated in important regions than across the full image. As show in Table 7, our method achieves a higher ratio on V^*, indicating more effective refocusing toward salient visual regions.

Model	Qwen2.5-VL	PixelReasoner	PAPO	DeepEyes (S)	Thyme (S)	Ours
Ratio	0.73	0.75	0.81	0.79	0.82	1.15

Table 7: Attention quantification. S denotes single-prefilling (forward pass). A higher ratio indicates that attention is more concentrated in important regions.

Attention Visualization. To better understand how visual reflection guides the model's perception, we visualize the attention maps from the generated <VR> tokens to the input images. As shown in Fig. 4, the attention of reflection tokens consistently concentrates on the question-relevant regions. Notably, these objects often occupy only a small portion of the high-resolution image. Despite the large number of visual tokens, the reflected representations guide the model to selectively focus on informative regions rather than being distracted by irrelevant background content. This observation suggests that visual reflection effectively mitigates the attention sink phenomenon caused by massive visual tokens, enabling the model to better localize and perceive about fine-grained visual details.

5 Limitations and Discussion

During training, the frequency of <VR> tokens is higher than that of the <EOR> token. As a result, the model occasionally tends to continue generating reflection tokens rather than terminating the reflection phase. Therefore, we explicitly insert the <EOR> token after a predefined number of <VR> tokens. The number of

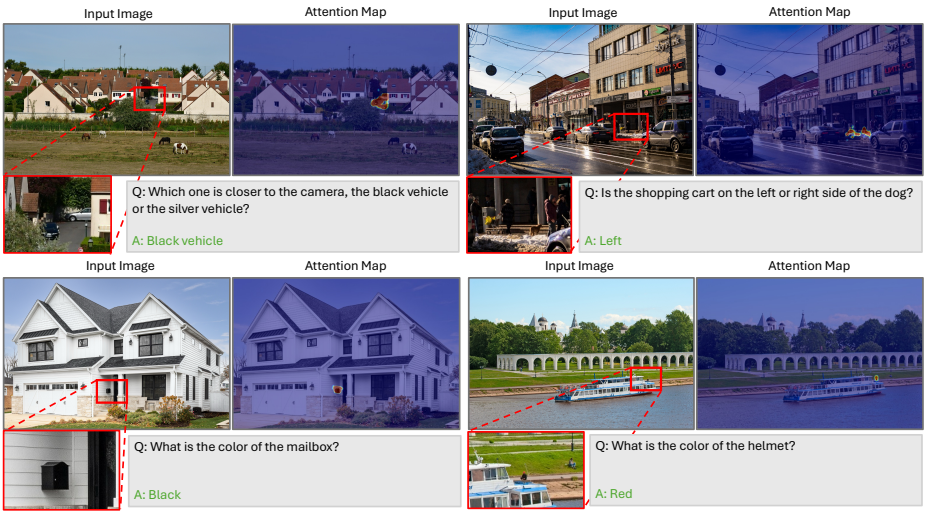

Fig. 4: Attention visualization of visual reflection tokens (<VR>). It illustrates the spatial focus of the generated visual reflection tokens on the input image.

reflection tokens should adapt dynamically to the complexity and richness of the question-relevant visual features. Developing an adaptive reflection mechanism is a potential direction for future work.

The video training aligns reflection tokens with visual features at the frame level. However, large-scale video question answering datasets with precise temporal annotations remain limited, and spatiotemporal grounding annotations for QA are even rarer. Richer video datasets with fine-grained spatiotemporal annotations could further enhance the effectiveness of latent visual reflection for long-video perception.

6 Conclusion

In conclusion, we presented **VisReflect**, a simple and principled framework to enhance fine-grained perception in large visual contexts for high-resolution images and long videos. Unlike prior multi-pass paradigms that rely on deterministic token predictions and repeated cropping or zooming, VisReflect generates continuous visual reflection embeddings that selectively emphasize salient regions or frames in latent space. This approach enables efficient single-pass reasoning, guiding cross-modal attention toward relevant visual features without the need for explicit coordinate prediction or iterative re-encoding. Extensive experiments on challenging image benchmarks and long-video benchmarks demonstrate that VisReflect consistently improves performance, achieving an average gain of 4.1% on images and 1.8% on videos compared to strong single-pass baselines. Furthermore, our method reduces inference time by roughly 44% relative to multi-pass approaches while maintaining comparable accuracy.

References

1. Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhu, D., Ding, J., Elhoseiny, M.: Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. arXiv preprint arXiv:2404.03413 (2024) 4
2. Ataallah, K., Shen, X., Abdelrahman, E., Sleiman, E., Zhuge, M., Ding, J., Zhu, D., Schmidhuber, J., Elhoseiny, M.: Goldfish: Vision-language understanding of arbitrarily long videos. arXiv preprint arXiv:2407.12679 (2024) 4
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923 9, 10, 11, 12
4. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning (2023), https://arxiv.org/abs/2310.09478 4
5. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 4
6. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al.: Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188 (2024) 11
7. Cheng, J., Van Durme, B.: Compressed chain of thought: Efficient reasoning through dense representations. arXiv preprint arXiv:2412.13171 (2024) 4
8. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024) 4
9. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., et al.: Perceptionlm: Open-access data and models for detailed visual understanding. arXiv preprint arXiv:2504.13180 (2025) 9
10. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025) 3, 9, 11
11. Fu, S., Yang, Q., Li, Y.M., Wei, X., Xie, X., Zheng, W.S.: Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. arXiv preprint arXiv:2509.24786 (2025) 2, 4, 9, 11
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024) 3, 8
13. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024) 4
14. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. arXiv preprint arXiv:2503.03321 (2025) 2
15. Lai, X., Li, J., Li, W., Liu, T., Li, T., Zhao, H.: Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. arXiv preprint arXiv:2509.07969 (2025) 4

16. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. NeurIPS (2021) 9
17. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. arXiv preprint arXiv:2509.24251 (2025) 4, 10
18. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024) 9, 11
19. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 4
20. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024) 9, 11
21. Li, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025) 9, 11
22. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: ECCV (2024) 4
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 4
24. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4122–4134 (2025) 11
25. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 4
26. Pan, J., Zhang, Q., Zhang, R., Lu, M., Wan, X., Zhang, Y., Liu, C., She, Q.: Timesearch-r: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. arXiv preprint arXiv:2511.05489 (2025) 2, 4, 9, 11
27. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023) 4
28. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024) 4
29. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024) 9
30. Shen, X., Chen, M.H., Wang, Y.C.F., Elhoseiny, M., Hachiuma, R.: Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. arXiv preprint arXiv:2512.14273 (2025) 2, 4
31. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024) 4, 9, 11

32. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. arXiv preprint arXiv:2510.14032 (2025) 4
33. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: Codi: Compressing chain-of-thought into continuous space via self-distillation. arXiv preprint arXiv:2502.21074 (2025) 4
34. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: CVPR (2024) 4
35. Wang, H., Su, A., Ren, W., Lin, F., Chen, W.: Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. arXiv preprint arXiv:2505.15966 (2025) 4, 9, 10
36. Wang, W., Ding, L., Zeng, M., Zhou, X., Shen, L., Luo, Y., Yu, W., Tao, D.: Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7907–7915 (2025) 3, 8, 12
37. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., et al.: Perception-aware policy optimization for multimodal reasoning. arXiv preprint arXiv:2507.06448 (2025) 9, 10
38. Wei, Y., Zhao, L., Lin, K., Yu, E., Peng, Y., Dong, R., Sun, J., Wei, H., Ge, Z., Zhang, X., et al.: Perception in reflection. arXiv preprint arXiv:2504.07165 (2025) 4
39. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13084–13094 (2024) 8, 12
40. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: CVPR (2024) 9
41. Yan, Z., Li, X., He, Y., Yue, Z., Zeng, X., Wang, Y., Qiao, Y., Wang, L., Wang, Y.: Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. arXiv preprint arXiv:2509.21100 (2025) 4, 9, 11
42. Yu, X., Guan, D., Gu, Y.: Zoom-refine: Boosting high-resolution multimodal understanding via localized zoom and self-refinement. arXiv preprint arXiv:2506.01663 (2025) 4
43. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9127–9134 (2019) 9
44. Zhang, Y.F., Lu, X., Yin, S., Fu, C., Chen, W., Hu, X., Wen, B., Jiang, K., Liu, C., Zhang, T., et al.: Thyme: Think beyond images. arXiv preprint arXiv:2508.11630 (2025) 2, 4, 9, 10
45. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing "thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025) 2, 4, 9, 10
46. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025) 3, 9, 11
47. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 4