CORRECT YET INSENSITIVE TO VISION: ADAPTIVE INTERVENTION TO IMPROVE VISUAL RELIANCE IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While vision-language models (VLMs) continue to improve on visual question answering (VQA) benchmarks, we identify a critical limitation: an apparently correct answer may not derive from visual evidence. We introduce the **Visual Insensitivity Score (VIS)** to quantify this behavior and surprisingly find that many correct answers remain unchanged when visual input is replaced or removed. This reveals substantial visual insensitivity in VLMs that conventional VQA accuracy fails to capture. To improve visual reliance, existing methods typically employ visual interventions with manually specified strengths, without explicitly accounting for whether model answers are grounded in visual evidence. To address this limitation, we propose **Adaptive Visual Intervention Optimization (VisIO)**, which explicitly optimizes lightweight interventions to enforce reliance on answer-determining visual evidence rather than other contextual cues. VisIO achieves this by constructing *language-preserving, answer-flipping visual counterfactual pairs*, where the linguistic context and most visual content remain unchanged while controlled visual changes flip the correct answer. Importantly, VisIO constructs such pairs using a small set of unlabeled target videos and optimizes only hundreds of head-wise parameters before inference. Across multiple VLMs and video VQA benchmarks, VisIO consistently improves accuracy while reducing visual insensitivity and strengthening reliance on visual evidence.

1 INTRODUCTION

Recent advances in vision-language models (VLMs) have significantly expanded the capabilities of multimodal systems, with continued performance improvements across a wide range of visual question answering (VQA) benchmarks Asadi et al. (2026); Feng et al. (2026b). These impressive results seemingly suggest that VLMs are becoming increasingly capable of understanding and reasoning about the visual world. However, we highlight a critical limitation of existing VLMs: **they may arrive at a correct answer without actually depending on the visual evidence.**

To investigate whether correct VLM answers depend on the provided visual evidence, we compare answers when the visual input is altered or removed. We introduce the **Visual Insensitivity Score (VIS)** by evaluating questions correctly answered with the real video under four visual conditions: real video, random video, blank video, and text-only input, as illustrated in Fig. 1 (1) and detailed in Sec. 3. Surprisingly, many correct answers remain unchanged across these substantially different visual conditions, revealing substantial visual insensitivity in existing VLMs. Such insensitivity suggests that these correct answers may have weak reliance on the provided visual evidence.

More importantly, high VQA accuracy can coexist with high VIS (Fig.1 (2)), indicating that overall accuracy can obscure visual insensitivity and potentially overestimate a VLM's visual understanding capability. To address this issue, we distinguish visual-insensitive and visual-sensitive samples for each model and evaluate their accuracy separately in this paper, allowing answer correctness to be assessed together with its reliance on visual evidence.

This finding raises a more challenging question: *how can we make VLMs rely more on visual evidence rather than produce plausible yet visually ungrounded answers?* Prior work has demonstrated the potential of attention interventions to enhance visual influence in VLMs He et al. (2025); Sarkar

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

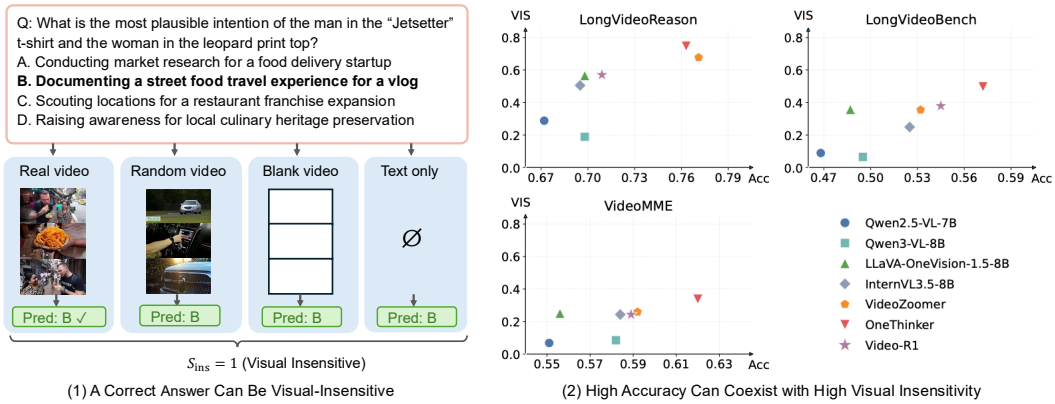


Figure 1: Analysis of visual insensitivity in VLMs. (1) An answer is visual-insensitive ($S_{\text{ins}} = 1$) if it remains unchanged across four visual conditions, indicating that its correctness may not be grounded in visual evidence. (2) The Visual Insensitivity Score (VIS) quantifies such behavior among correct answers. High VQA accuracy can coexist with high VIS, showing that overall accuracy can obscure visual insensitivity and overestimate visual grounding.

et al. (2025). However, existing methods typically rely on manually specified intervention strengths, where stronger visual influence does not explicitly ensure improved reliance on the evidence that determines the answer. We instead formulate visual intervention as an optimization problem and learn intervention strengths from questions whose answers are determined by visual evidence, adaptively optimizing visual reliance rather than heuristically increasing visual influence. Constructing such questions, however, typically requires costly human annotation, particularly for videos.

To address these challenges, we propose **Adaptive Visual Intervention Optimization (VisIO)**, which adaptively learns lightweight intervention parameters using unlabeled target videos to improve visual reliance in VLMs. VisIO constructs *language-preserving, answer-flipping visual counterfactual pairs*, where language cues remain fixed while controlled manipulations of visual evidence flip the correct answer. Such pairs cannot be resolved from language cues alone, providing direct supervision for optimizing visual reliance without human annotations.

Importantly, the constructed pairs are not equally informative for intervention optimization. We therefore introduce an *intervention-aware pair selection* strategy that prioritizes pairs for which the model is sensitive to controlled visual changes and the induced gradients are aligned with increasing visual contribution.

Finally, VisIO keeps the VLM frozen and learns head-wise scaling parameters that directly modulate visual contributions within attention heads. Importantly, VisIO requires only hundreds of trainable parameters and a small set of unlabeled target videos before inference, enabling lightweight adaptation without model-level training or external training data.

Our main contributions are summarized as follows: 1) We reveal that existing VLMs can suffer from visual insensitivity: an apparently correct answer can remain unchanged even when visual input is altered or removed. We introduce the Visual Insensitivity Score (VIS) to quantify this behavior, showing that high VQA accuracy can obscure weak reliance on visual evidence. 2) To enhance visual reliance, we propose VisIO to adaptively optimize lightweight head-wise interventions using a small set of unlabeled target videos, avoiding manually specified intervention strengths and large-scale training data. 3) To increase VLMs' reliance on answer-determining visual evidence, we construct language-preserving, answer-flipping visual counterfactual pairs and develop an intervention-aware strategy to select informative pairs for effective optimization. 4) Across multiple base and post-trained VLMs and video VQA benchmarks, VisIO consistently improves answer accuracy while reducing visual insensitivity and strengthening reliance on visual evidence.

2 Related Work

Visual Grounding and Language Priors in VLMs. Prior studies have revealed substantial language priors in VLM evaluation, showing that questions can often be answered competitively when visual input is unavailable or removed Lin et al. (2024); Asadi et al. (2026); Chen et al. (2024).

Other works study weak visual grounding mainly through visual hallucination, where generated responses conflict with the provided visual content Xie et al. (2024); An et al. (2025a); Jiang et al. (2025). These studies reveal important failures when visual information is absent or model outputs contradict it, but a critical question remains: *when vision is available and the answer is correct, does the answer actually depend on the provided visual evidence?* We study this setting by measuring whether correct answers remain unchanged as visual evidence is altered or removed. This allows us to expose visual insensitivity in VLMs that can remain hidden behind conventional VQA accuracy.

Visual Intervention in VLMs. A growing body of work improves visual grounding by intervening in the internal computation or decoding process of VLMs. Contrastive decoding methods mitigate hallucinations by contrasting output distributions from original and perturbed visual inputs Leng et al. (2024); Suo et al. (2025), while recent approaches manipulate internal visual attention or representations to enhance visual influence Chen et al. (2025); Tang et al. (2026); Basile et al. (2026); Yu et al. (2026). Although lightweight and training-free, these methods typically rely on manually specified intervention rules or strengths rather than learning them toward visual reliance. In contrast, training-based approaches explicitly align model responses with visual evidence through supervised or preference-based optimization Luo et al. (2025); Favero et al. (2024); Chowdhury et al. (2026), but require additional training data and substantially more trainable parameters. VisIO bridges these two paradigms by explicitly learning visual interventions while keeping the VLM frozen: it optimizes only hundreds of head-wise parameters from a small set of unlabeled target videos before inference, without human annotations or model-level fine-tuning.

Self-Supervised Video Learning. Recent works derive self-supervision from unlabeled visual data by automatically verifiable visual transformations He et al. (2026); Liu et al. (2026); Wu et al. (2026). VisIO differs by constructing answer-flipping pairs, where controlled visual manipulations change the correct answer while preserving the linguistic context and most visual content. Moreover, rather than treating all self-supervised questions equally, VisIO introduces an intervention-aware selection mechanism to prioritize those most informative for intervention optimization.

3 Model- and Sample-Specific Visual Insensitivity

Despite continued improvements in VQA accuracy across benchmarks, it remains unclear whether correct VLM answers actually result from reasoning over the provided visual evidence. We investigate this question by evaluating each sample under four visual conditions: real video, random video, blank video, and text only. The latter three alter or remove the visual evidence while keeping the question unchanged, allowing us to examine whether the model's answer responds to visual changes.

3.1 Quantifying Visual Insensitivity across VLMs

For video sample i, let $\hat{y}_i^m$ denote the answer under condition $m \in \{\text{real}, \text{random}, \text{blank}, \text{text}\}$. We define the sample-level visual insensitivity indicator as

$$S_{\text{ins}}^{(i)} = \mathbb{I}\big[\hat{y}_i^{\text{real}} = \hat{y}_i^{\text{random}} = \hat{y}_i^{\text{blank}} = \hat{y}_i^{\text{text}}\big]\,. \tag{1}$$

Thus, $S_{\text{ins}}^{(i)} = 1$ indicates that the answer remains unchanged despite replacing or removing the visual evidence, while $S_{\text{ins}}^{(i)} = 0$ otherwise. Since our primary interest is whether a *correct* answer depends on visual evidence, we further define the **Visual Insensitivity Score (VIS)** over samples *correctly* answered with the real video:

$$\text{VIS} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} S_{\text{ins}}^{(i)}, \qquad \mathcal{C} = \big\{i : \hat{y}_i^{\text{real}} = y_i\big\}\,. \tag{2}$$

VIS therefore measures the fraction of correct real-video answers that remain unchanged when the visual evidence is altered or removed.

As shown in Fig. 1 (2), visual insensitivity is prevalent across models and benchmarks. More importantly, higher VQA accuracy does not necessarily correspond to lower VIS. For example, post-trained models such as VideoZoomer Ding et al. (2026) and OneThinker Feng et al. (2026b) achieve higher accuracy than their base-model counterpart Qwen2.5-VL-7B Bai et al. (2025b) and Qwen3-VL-8B Bai et al. (2025a), yet also exhibit substantially higher VIS across benchmarks. This in-

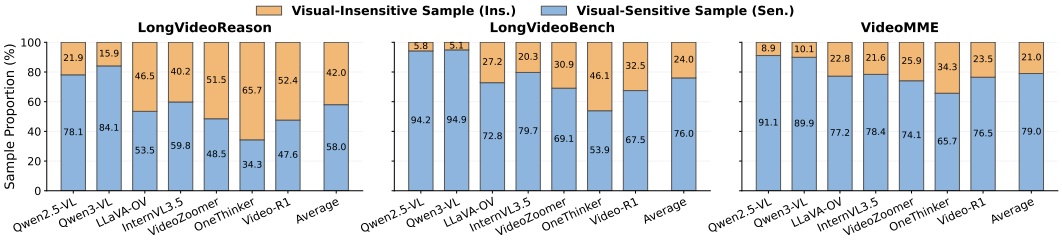


Figure 2: Distribution of visual-insensitive and visual-sensitive samples.

dicates that higher answer accuracy does not necessarily reflect stronger visual reliance, and may instead obscure substantial visual insensitivity.

These findings expose a key limitation of conventional VQA accuracy evaluation: a correct answer contributes equally to overall accuracy even when it remains unchanged as visual evidence changes, potentially overestimating the model's visual grounding capacity.

3.2 Sample-Specific Visual Sensitivity

The prevalence of visual insensitivity motivates distinguishing samples according to their responses to visual changes to better characterize visual reliance. For each model, we categorize a sample as **visual-insensitive** if $S_{\text{ins}}^{(i)} = 1$, and **visual-sensitive** otherwise. This categorization is **model-specific**: the same question can be visual-insensitive for one VLM but visual-sensitive for another.

Fig. 2 reveals pronounced variation in visual insensitivity across both benchmarks and models. Across benchmarks, LongVideoReason contains the largest proportion of visual-insensitive samples, averaging 42.0% across models, compared with 24.0% on LongVideoBench and 21.0% on VideoMME. The substantially higher visual insensitivity on LongVideoReason suggests that it may contain more questions where models can answer without closely following the provided visual evidence.

Visual insensitivity also exhibits a strong model-specific pattern. Qwen2.5-VL and Qwen3-VL consistently show relatively low proportions of visual-insensitive samples, whereas post-trained models such as VideoZoomer, OneThinker, and Video-R1 Feng et al. (2026a) exhibit considerably higher proportions. Such variation means that overall accuracy aggregates visual-insensitive and visual-sensitive samples in substantially different proportions across models, making it difficult to determine whether performance is driven by samples that respond to visual evidence or those that remain insensitive to it. Therefore, throughout the evaluations in Sec. 5, we report accuracy separately on the visual-insensitive and visual-sensitive subsets for each model.

4 Adaptive Visual Intervention Optimization

To mitigate visual insensitivity in VLMs, we move beyond manually specified visual interventions and formulate intervention as an optimization problem. We propose **Adaptive Visual Intervention Optimization (VisIO)**, which automatically learns lightweight head-wise intervention parameters using unlabeled target videos to improve reliance on visual evidence.

As illustrated in Fig. 3, VisIO consists of three stages. First, we construct answer-flipping visual counterfactual pairs from unlabeled target videos, providing direct supervision that requires the VLM to leverage answer-determining visual evidence rather than shared language cues. Second, we select informative pairs based on visual responsiveness and intervention potential for effective intervention optimization. Finally, using the selected pairs, we optimize head-wise scaling parameters to adaptively modulate visual contributions across attention heads.

4.1 Head-Wise Visual Contribution Scaling

Prior methods have shown that modulating individual attention heads in VLMs can effectively influence their use of visual information Chen et al. (2025); Yu et al. (2026). Inspired by these works, we adopt a lightweight head-wise intervention as the optimization space of VisIO.

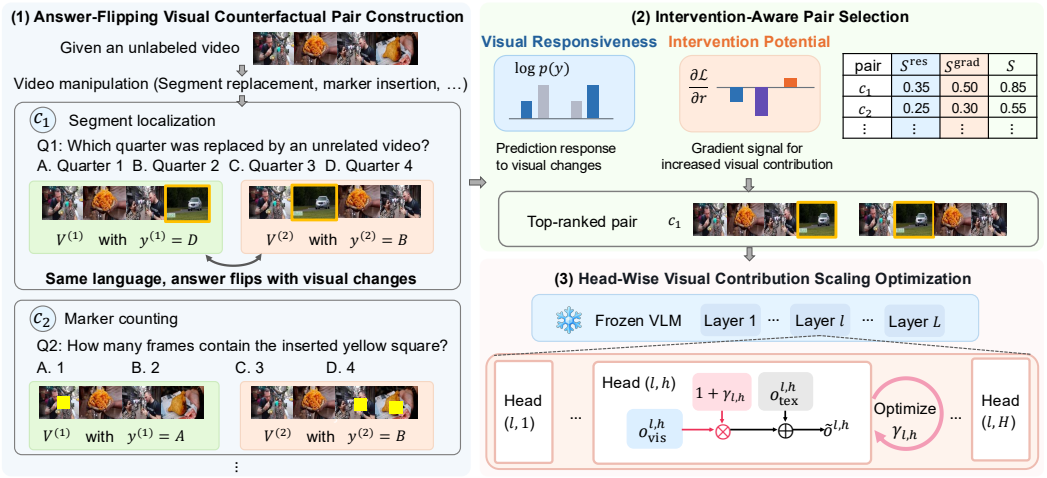


Figure 3: Overview of **Adaptive Visual Intervention Optimization (VisIO)**. Given unlabeled target videos, VisIO (1) constructs language-preserving, answer-flipping visual counterfactual pairs, whose answers change solely with controlled manipulation of visual evidence; (2) selects informative pairs based on visual responsiveness and intervention potential; and (3) uses the selected pairs to optimize head-wise visual contribution scaling parameters while keeping the VLM frozen, encouraging predictions to rely more strongly on answer-determining visual evidence.

Let $\mathcal{V}$ and $\mathcal{T}$ denote the visual- and text-token positions, respectively. For layer l, head h, and query position k, we decompose the head output into visual and textual contributions:

$$\mathbf{o}^{l,h}_{k,\text{vis}} = \sum_{j\in\mathcal{V}} A^{l,h}_{kj}\mathbf{v}^{l,h}_{j}, \quad \mathbf{o}^{l,h}_{k,\text{tex}} = \sum_{j\in\mathcal{T}} A^{l,h}_{kj}\mathbf{v}^{l,h}_{j}, \tag{3}$$

where $A^{l,h}_{kj}$ denotes the attention weight from query k to token j, and $\mathbf{v}^{l,h}_{j}$ is the corresponding value vector. The head output is then given by

$$\mathbf{o}^{l,h}_{k} = \mathbf{o}^{l,h}_{k,\text{vis}} + \mathbf{o}^{l,h}_{k,\text{tex}}. \tag{4}$$

To enhance VLMs' reliance on visual evidence, we directly modulate the visual contribution of each attention head while preserving its textual contribution:

$$\widetilde{\mathbf{o}}^{l,h}_{k} = (1+\gamma_{l,h})\mathbf{o}^{l,h}_{k,\text{vis}} + \mathbf{o}^{l,h}_{k,\text{tex}}, \qquad \gamma_{l,h} = \gamma_{\max}\sigma(r_{l,h}), \tag{5}$$

where $r_{l,h}$ is a learnable unconstrained parameter, $\gamma_{l,h}$ is the corresponding bounded intervention strength, and $\gamma_{\max}$ specifies its upper bound. For a VLM with L layers and H heads, VisIO introduces only $L \times H$ trainable scalars while keeping all original model parameters frozen.

Existing intervention methods typically rely on manually specified heads or intervention strengths Sarkar et al. (2025); He et al. (2025). In contrast, our method automatically optimizes head-wise intervention parameters to adaptively strengthen the visual contribution of each attention head.

4.2 Answer-Flipping Visual Counterfactual Pair Construction

To optimize the intervention parameters for stronger visual reliance, we construct *language-preserving, answer-flipping visual counterfactual pairs* that provide direct supervision, requiring the model to rely on answer-determining visual evidence rather than language cues or non-determining visual context. Importantly, these pairs are constructed automatically from unlabeled target videos, requiring neither human annotations nor external training data.

Given an unlabeled target video V, we apply controlled visual manipulations to obtain two counterfactual views $V^{(1)}$ and $V^{(2)}$. We construct a shared question Q with answer choices $\mathcal{A} = \{a_1, \ldots, a_N\}$, while ensuring that the correct answer flips between the two views:

$$Q^{(1)} = Q^{(2)} = Q, \qquad \mathcal{A}^{(1)} = \mathcal{A}^{(2)} = \mathcal{A}, \qquad V^{(1)} \neq V^{(2)}, \qquad y^{(1)} \neq y^{(2)}, \tag{6}$$

where $y^{(1)}, y^{(2)} \in \mathcal{A}$ denote the corresponding ground-truth answers. Since the linguistic context is identical while the correct answer changes, correctly predicting both views requires the model to rely on the answer-determining visual evidence in each view.

Specifically, we instantiate this principle with question templates and corresponding video manipulations for video VQA. As illustrated in Fig. 3 (1), these include segment localization (e.g., replacing a segment with one from an unrelated video and identifying its temporal quarter), marker counting (e.g., inserting a marker into selected frames and counting the sampled frames containing the marker), and other temporal reasoning tasks. For each template, the correct answer is directly determined by the controlled visual manipulation, allowing counterfactual pairs and their ground-truth answers to be constructed automatically without human annotations. Additional templates and examples are provided in the Appendix C.

4.3 Intervention-Aware Pair Selection

Although controlled visual manipulations provide direct supervision, the automatically constructed pairs are not equally informative for intervention optimization. We therefore introduce an intervention-aware pair selection strategy to identify informative pairs before optimization. For each target video V, let $\mathcal{C}(V)$ denote the set of candidate counterfactual pairs constructed from it. We evaluate each candidate from two complementary perspectives described below.

Visual Responsiveness. We first assess whether the model is sensitive to the controlled visual change. An informative pair should shift the model's answer preference toward the answer corresponding to each visual view, indicating that the model is responsive to the controlled visual difference. We therefore use a contrastive margin to measure the model's responsiveness to the controlled visual change. For a candidate pair c, let

$$p^{(m)}(a) = P(a \mid Q, \mathcal{A}, V^{(m)}), \quad m \in \{1, 2\}, \quad a \in \mathcal{A} \tag{7}$$

denote the probability assigned to answer a under counterfactual view m, evaluated at the initialization of the intervention parameters. We define the contrastive margin as

$$M^{\mathrm{cf}} = \frac{1}{2}\left[\log p^{(1)}(y^{(1)}) - \log p^{(1)}(y^{(2)})\right] + \frac{1}{2}\left[\log p^{(2)}(y^{(2)}) - \log p^{(2)}(y^{(1)})\right].$$

The two terms contrast the answer associated with each visual view against its counterfactual answer, jointly capturing visual responsiveness in both directions. We map the margin to a bounded visual responsiveness score:

$$S^{\mathrm{res}} = \max\left(0, \tanh\left(M^{\mathrm{cf}}\right)\right). \tag{8}$$

A larger S^{res} indicates stronger visual responsiveness.

Intervention Potential. Visual responsiveness alone does not indicate whether a pair can effectively guide intervention optimization. Since our goal is to improve visual reliance, we further assess whether optimizing a pair induces gradients that increase visual contribution in attention heads. For candidate c, we define the paired loss as

$$\mathcal{L}_{\mathrm{pair}} = -\frac{1}{2}\log p^{(1)}(y^{(1)}) - \frac{1}{2}\log p^{(2)}(y^{(2)}). \tag{9}$$

Let $\{r_{l,h}\}$ denote the intervention parameters introduced in Sec. 4.1. We compute

$$g_{l,h} = \frac{\partial \mathcal{L}_{\mathrm{pair}}}{\partial r_{l,h}}. \tag{10}$$

Since $\gamma_{l,h} = \gamma_{\max}\sigma(r_{l,h})$ increases monotonically with $r_{l,h}$, a negative $g_{l,h}$ indicates that gradient descent increases $r_{l,h}$, and strengthens the scaling applied to the visual contribution of head (l, h). We therefore measure intervention potential by aggregating vision-enhancing gradient components:

$$G = \frac{1}{LH}\sum_{l=1}^{L}\sum_{h=1}^{H}\max(0, -g_{l,h}). \tag{11}$$

A larger G indicates greater potential for the pair to drive vision-enhancing intervention updates. Importantly, this criterion does not impose increased visual contribution on all heads: positive gradient components are excluded only from the selection score, without constraining their updates during subsequent optimization.

For candidates $c_i \in \mathcal{C}(V)$, let G_i and S_i^{res} denote their intervention potential and visual responsiveness scores, respectively. To account for gradient-scale variation across target videos, we normalize the intervention potentials as

$$S_i^{\text{grad}} = \frac{G_i}{\max_{c_j \in \mathcal{C}(V)} G_j + \epsilon}, \tag{12}$$

where ϵ is a small constant for numerical stability. We combine the two criteria as

$$S_i = S_i^{\text{res}} + S_i^{\text{grad}}. \tag{13}$$

For each video, we rank its candidate pairs by S_i and retain the top-ranked pair for optimization.

4.4 Adaptive Optimization Before Inference

We optimize the intervention parameters over the selected counterfactual pairs using

$$\mathcal{L}_{\text{VisIO}} = \mathcal{L}_{\text{pair}} + \lambda_\gamma \mathcal{L}_\gamma, \tag{14}$$

where $\mathcal{L}_{\text{pair}}$ denotes the paired loss in Eq. 9 averaged over the selected pairs, and $\mathcal{L}_\gamma = \frac{1}{LH}\sum_{l,h} |\gamma_{l,h}|$ is an ℓ_1 regularizer on the intervention magnitude. During optimization, only $\{r_{l,h}\}$ are updated, while all VLM parameters remain frozen. Once optimized, the intervention parameters are fixed and applied during inference without further adaptation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate VLMs and our method on three widely used video VQA benchmarks: LongVideoBench Wu et al. (2024), Video-MME Fu et al. (2025), and LongVideoReason Chen et al. (2026), which exhibit varying degrees of visual reliance in our analysis (Sec. 3).

Evaluation Protocol. Following Sec. 3, we partition each model's samples into *visual-insensitive* and *visual-sensitive* subsets and report their accuracy alongside overall accuracy. We further report the proposed Visual Insensitivity Score (VIS) to quantify the insensitivity of correct answers to visual changes, where lower VIS indicates stronger visual reliance.

Implementation Details. For each target benchmark, we randomly sample 128 unlabeled target videos to construct counterfactual pairs. For intervention-aware selection, we retain the top-1 pair for each target video. We optimize the selected pairs for 10 epochs using AdamW with a learning rate of 0.01 and zero weight decay. The batch size is 2 with 4 gradient accumulation steps. We set the maximum intervention strength to $\gamma_{\max} = 0.3$ and the regularization coefficient to $\lambda_\gamma = 0.05$. VisIO introduces 784 trainable head-wise parameters (28 layers $\times$ 28 heads) for VideoZoomer, and $1{,}152$ parameters (36 layers $\times$ 32 heads) for Qwen3-VL-8B, OneThinker, and LLaVA-OneVision-1.5-8B. After optimization, the learned parameters are fixed and directly applied to all evaluation samples of the corresponding benchmark. On a single NVIDIA A100 80GB GPU, counterfactual pair construction and selection take approximately 4.5 hours per evaluated model, followed by approximately 1.5 hours for optimization. Unless otherwise specified, all ablation experiments are conducted with Qwen3-VL-8B as the backbone.

5.2 Comparison with State-of-the-Art Methods

Comparison across VLMs. Table 1 evaluates VLMs from different model families, including base models Qwen2.5-VL-7B Bai et al. (2025b), Qwen3-VL-8B Bai et al. (2025a), InternVL3.5-8B Wang et al. (2025), and LLaVA-OV-1.5-8B An et al. (2025b), and post-trained models VideoZoomer Ding et al. (2026), Video-R1 Feng et al. (2026a), and OneThinker Feng et al. (2026b). We apply VisIO to four models spanning different families and base/post-trained VLMs.

Table 1: Comparison of SOTA VLMs. **Ins.** and **Sen.** denote accuracy on model-specific *visual-insensitive* and *visual-sensitive* subsets, respectively, and are only comparable **within each model**. **Ins.** is shown in gray as a reference, as higher accuracy on this subset does not directly indicate stronger visual reliance and is therefore not a primary objective of VisIO. **All** and **VIS** denote overall accuracy and the Visual Insensitivity Score, respectively, and are comparable across models.

Model	LongVideoReason				LongVideoBench				VideoMME			
	Ins.	Sen.	All ↑	VIS ↓	Ins.	Sen.	All ↑	VIS ↓	Ins.	Sen.	All ↑	VIS ↓
InternVL3.5-8B	87.4	57.5	69.5	50.5	64.7	49.4	52.5	24.9	65.7	56.4	58.4	24.3
Qwen2.5-VL-7B	83.2	62.7	67.2	28.8	69.5	45.4	46.8	8.9	41.8	56.4	55.1	6.8
Video-R1	82.3	58.4	70.9	57.0	63.6	50.1	54.5	37.9	60.9	58.3	58.9	24.3
Qwen3-VL-8B	85.7	66.8	69.8	18.9	71.8	48.3	49.5	6.5	53.7	58.7	58.2	8.5
+VisIO	80.6 (-5.1)	70.5 (+3.7)	72.1 (+2.3)	**14.3** (-4.6)	68.0 (-3.8)	52.3 (+4.0)	53.1 (+3.6)	**4.5** (-2.0)	57.0 (+3.3)	61.0 (+2.3)	60.6 (+2.4)	**4.8** (-3.7)
LLaVA-OV-1.5-8B	84.6	56.9	69.8	56.5	59.9	44.5	48.7	35.7	58.6	54.7	55.6	25.0
+VisIO	84.7 (+0.1)	61.9 (+5.0)	72.5 (+2.7)	42.8 (-13.7)	65.2 (+5.3)	46.8 (+2.3)	51.8 (+3.1)	24.5 (-11.2)	61.5 (+2.9)	56.2 (+1.5)	57.4 (+1.8)	14.2 (-10.8)
VideoZoomer	89.0	64.5	77.1	67.7	62.1	49.2	53.2	35.5	61.8	58.3	59.2	25.9
+VisIO	88.1 (-0.9)	68.9 (+4.4)	**78.8** (+1.7)	52.7 (-15.0)	62.3 (+0.2)	51.0 (+1.8)	54.5 (+1.3)	23.1 (-12.4)	62.0 (+0.2)	61.2 (+2.9)	61.4 (+2.2)	14.4 (-11.5)
OneThinker	85.8	58.1	76.3	74.7	63.3	52.0	57.2	49.7	63.5	61.2	62.0	33.8
+VisIO	87.8 (+2.0)	60.4 (+2.3)	78.4 (+2.1)	58.3 (-16.4)	65.0 (+1.7)	54.2 (+2.2)	**59.2** (+2.0)	38.2 (-11.5)	64.2 (+0.7)	64.8 (+3.6)	**64.6** (+2.6)	20.4 (-13.4)

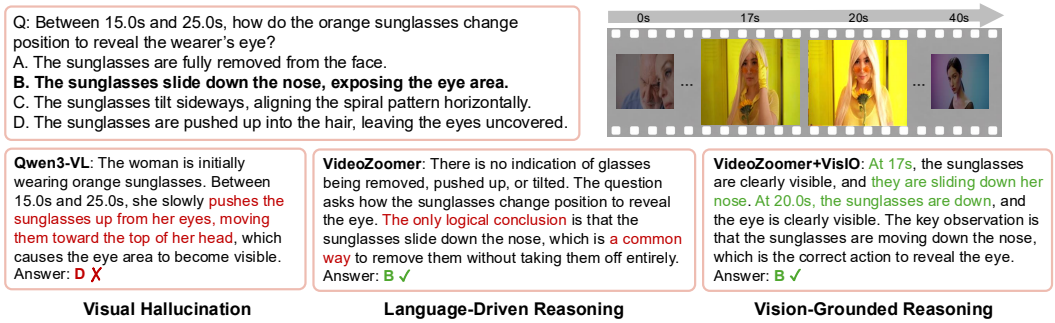


Figure 4: Qualitative comparison of visual reasoning. Qwen3-VL hallucinates the visual change and answers incorrectly. VideoZoomer produces the correct answer with a language-driven reasoning trace, whereas VisIO produces a reasoning trace that explicitly references the visual evidence.

Visual-insensitive accuracy. Most VLMs achieve higher accuracy on visual-insensitive than visual-sensitive samples, particularly on LongVideoReason. This suggests that many correct answers may have weak visual reliance, causing overall accuracy to overestimate the model's visual grounding ability. With VisIO, Ins. accuracy may increase or decrease. Importantly, this metric is inherently ambiguous: *a previously correct visual-insensitive answer may become visually sensitive but incorrect, or remain correct with stronger visual reliance*. Therefore, Ins. accuracy alone cannot directly reflect improvements in visual reliance and is reported mainly as a reference.

Visual-sensitive accuracy. VisIO consistently improves Sen. accuracy for every evaluated model across all three benchmarks. This shows that VisIO improves answer accuracy on samples whose answers are sensitive to changes in visual evidence.

Overall accuracy. Despite occasional accuracy drops on visual-insensitive samples, VisIO consistently improves overall accuracy across all evaluated models and benchmarks, showing that stronger visual reliance can be achieved alongside improved VQA performance.

Visual Insensitivity Score. VisIO consistently reduces VIS across all evaluated models and benchmarks. The reductions are particularly pronounced for post-trained models with high initial VIS, such as VideoZoomer and OneThinker, reaching up to 16.4 points. These results show that VisIO reduces visual insensitivity while simultaneously improving VQA accuracy, indicating stronger reliance of correct answers on visual input.

Qualitative Analysis. Fig. 4 provides a concrete example of why answer correctness alone is insufficient to characterize visual grounding. VisIO does not change VideoZoomer's final answer, but produces a reasoning trace that explicitly references the answer-determining visual evidence rather than relying on a generic language-based inference. This example shows that VisIO can make the generated reasoning trace more grounded in the provided visual evidence even when the original answer is already correct.

Table 2: Comparison with existing head-level visual intervention methods.

Model	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
Qwen3-VL (No intervention)	66.8	69.8	18.9	48.3	49.5	6.5
SPIN Sarkar et al. (2025)	67.3	70.1	18.2	48.5	49.6	7.4
VHR He et al. (2025)	65.2	68.2	17.6	50.2	51.2	6.6
VisIO	**70.5**	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

Table 3: Comparison of single-video and video-pair settings.

Model	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
Qwen3-VL	66.8	69.8	18.9	48.3	49.5	6.5
+VisIO (Single video)	67.3	70.2	17.6	50.1	50.8	7.7
+VisIO (Video pair)	**70.5**	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

Table 4: Effect of pair selection strategies.

Selection Strategy	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
Random Selection	67.2	70.7	17.6	48.9	50.6	9.5
Selection by S^{res}	68.8	71.5	15.3	49.6	50.2	8.3
Selection by S^{grad}	69.2	71.3	16.2	51.2	51.6	5.5
Selection by $S^{\text{res}} + S^{\text{grad}}$	**70.5**	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

Table 5: Effect of intervention targets.

Intervention Target	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
No intervention	66.8	69.8	18.9	48.3	49.5	6.5
Attention weight	64.5	68.2	22.6	46.3	47.3	10.7
Head output	67.9	69.5	20.4	48.6	50.6	5.7
Visual contribution (Ours)	**70.5**	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

Comparison with Visual Intervention Methods. Table 2 compares VisIO with SPIN Sarkar et al. (2025) and VHR He et al. (2025), two head-level visual intervention methods. VisIO achieves consistent improvements in accuracy while substantially reducing VIS, whereas SPIN and VHR yield smaller or inconsistent gains. Unlike these methods that rely on predefined intervention strengths, VisIO optimizes parameters using answer-flipping counterfactual supervision, enabling adaptive modulation of head-wise visual contributions toward stronger visual reliance.

5.3 Ablation Study

Effect of Counterfactual Pairing. Table 3 compares two forms of supervision for intervention optimization. In the *single-video* setting, only one randomly selected video-question instance from each counterfactual pair is used to compute the loss, whereas the *video-pair* setting uses both instances. Using both instances consistently yields larger performance gains. As the linguistic context is identical while the correct answers differ, the paired supervision explicitly associates the answer difference with the changed visual evidence, providing a stronger signal to improve visual reliance.

Effect of Pair Selection. Table 4 shows that both visual responsiveness (S^{res}) and intervention potential (S^{grad}) individually improve performance over random selection, where one pair is randomly selected from all constructed candidates for each target video. Combining the two criteria consistently achieves the highest accuracy and lowest VIS. Specifically, S^{res} prioritizes pairs that elicit stronger model responses to controlled visual changes, whereas S^{grad} prioritizes pairs whose induced gradients favor increased visual contribution. Their combination therefore selects pairs that are both visually informative to the model and effective for guiding intervention optimization.

Effect of Intervention Target. We further compare different intervention targets under the same optimization setting. Following Sec. 4.1, *attention-weight intervention* scales visual attention as $\widetilde{A}^{l,h}_{kj} = (1+\gamma_{l,h})A^{l,h}_{kj}$ for $j \in \mathcal{V}$, while *head-output intervention* scales the entire head as $\widetilde{\mathbf{o}}^{l,h}_{k} = (1+\gamma_{l,h})\mathbf{o}^{l,h}_{k}$. VisIO instead directly scales the visual contribution as defined in Eq. 5. As shown in Table 5, both alternatives degrade accuracy, whereas VisIO improves accuracy while achieving the lowest VIS. Attention-weight intervention affects visual contribution only indirectly through attention weights, while head-output intervention jointly scales visual and textual contributions. In contrast, VisIO directly targets the visual contribution itself, more effectively strengthening the influence of visual evidence.

6 Conclusion

We reveal that existing VLMs can exhibit substantial visual insensitivity: many correct answers remain unchanged even when the provided visual evidence is altered or removed. We introduce VIS to quantify this behavior and VisIO to mitigate this issue by adaptively optimizing head-wise visual contributions with answer-flipping counterfactual supervision. Across multiple VLMs and video VQA benchmarks, VisIO effectively improves accuracy while strengthening visual reliance.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, literature search, research brainstorming, feedback on experimental design, and the development and debugging of research code. All AI-assisted content, code, and references were reviewed and verified by the authors. The authors take full responsibility for the research, results, and final content of this paper.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, 2025a.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025b.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. *Advances in Neural Information Processing Systems*, 38:112351–112383, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for VLMs? An attention mechanism perspective on focus areas. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 9910–9932, 2025.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026.

Rahul Chowdhury, Timothy A Rupprecht, Xuan Shen, Pu Zhao, and Yanzhi Wang. Scale: Attention head scaling as a minimal adapter for spatial reasoning in vision language models. 2026.

Yang Ding, Xin Lai, Yizhen Zhang, Wei Li, Ruihang Chu, and Yujiu Yang. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 20087–20111, 2026.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026a.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026b.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501, 2025.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Videossr: Video self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26530–26540, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 29914–29934, 2024.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9570–9581, 2026.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. In *Proceedings of the International Conference on Machine Learning*, 2025.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12492–12511, 2025.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29904–29914, 2025.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9439–9447, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves mllms. In *International Conference on Learning Representations*, volume 2026, pp. 93221–93245, 2026.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36021–36029, 2026.

A Additional Ablation Study

Effect of the Number of Target Videos. Table 6 studies the effect of the number of unlabeled target videos used for intervention optimization. Performance tends to improve as the number of target videos increases, with the strongest and most balanced gains emerging around 128 videos. Using 128 videos achieves strong and balanced performance across both benchmarks, reaching 70.5/72.1 Sen./All with a VIS of 14.3 on LongVideoReason and 52.3/53.1 with a VIS of 4.5 on LongVideoBench. Further increasing the number of videos does not yield consistent improvements: although individual metrics can improve, such as Sen. with 256 videos and All with 512 videos, gains are not maintained across accuracy and VIS. These results suggest that VisIO can learn effective intervention parameters from a relatively small set of unlabeled target videos, with performance largely saturating around 128 videos.

Variants of Pair Selection Scores. Table 7 compares alternative formulations of the two pair selection scores. For S^{res}, the contrastive formulation consistently outperforms using the correct-answer probabilities alone, yielding higher accuracy and substantially lower VIS on both benchmarks. This indicates that responsiveness is better characterized by the relative preference between the answers associated with the two counterfactual views, rather than by their individual prediction confidence. For S^{grad}, using the absolute gradient magnitude improves Sen. on LongVideoReason but degrades overall accuracy and VIS, with larger drops on LongVideoBench. In contrast, retaining only the vision-enhancing gradient components provides more consistent gains across metrics and benchmarks. These results support considering both counterfactual contrast in visual responsiveness and gradient direction in intervention potential.

Table 6: Ablation on the number of unlabeled target videos used for VisIO optimization.

# Target Videos	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
No intervention	66.8	69.8	18.9	48.3	49.5	6.5
8	66.4	70.2	20.6	49.0	49.2	7.8
16	66.2	70.9	16.1	48.6	50.9	6.0
32	67.9	70.2	16.2	50.4	51.1	5.6
64	69.7	71.7	15.1	51.5	52.2	5.0
128	**70.5**	72.1	14.3	52.3	**53.1**	**4.5**
256	70.2	71.9	**13.5**	**52.8**	52.2	6.3
512	69.9	**72.8**	14.7	50.6	52.8	5.4

B Limitations

1) Our evaluation currently focuses on long-video VQA benchmarks, and the effectiveness of VisIO on image-based VQA and other multimodal reasoning settings remains to be investigated. 2) The

Table 7: Ablation of variants of the pair selection scores.

Score	Key Formulation	LongVideoReason			LongVideoBench		
		Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
S^{res}	$\frac{1}{2}\left[p^{(1)}(y^{(1)}) + p^{(2)}(y^{(2)})\right]$	68.3	71.2	15.8	51.7	51.3	8.4
	$\frac{1}{2}\left[\log p^{(1)}(y^{(1)}) - \log p^{(1)}(y^{(2)})\right] + \frac{1}{2}\left[\log p^{(2)}(y^{(2)}) - \log p^{(2)}(y^{(1)})\right]$	**70.5**	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**
S^{grad}	$\frac{1}{LH}\sum_{l,h}\lvert g_{l,h}\rvert$	**71.1**	70.6	14.8	51.2	51.9	7.3
	$\frac{1}{LH}\sum_{l,h}\max(0, -g_{l,h})$	70.5	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

current counterfactual construction relies on a predefined set of automatically verifiable video manipulations and question templates. Although they cover multiple temporal reasoning tasks, extending the construction to more diverse visual concepts and reasoning scenarios could further improve the generality of VisIO. 3) VisIO introduces additional computational overhead before inference. On a single NVIDIA A100 80GB GPU, counterfactual pair construction and selection, together with intervention optimization, take approximately six hours per model-benchmark setting. Improving the diversity and efficiency of counterfactual construction and adaptation provides promising directions for future work.

C Answer-Flipping Visual Counterfactual Pairs

C.1 Question Types

We design eight question templates to provide diverse vision-dependent supervision for VisIO. They are grouped into two categories: *event-based questions*, which manipulate natural video events, and *marker-based questions*, which introduce controlled visual markers. For all templates, counterfactual pairs share the same question and answer choices, while controlled visual changes alter the correct answer.

Event-Based Questions. Event-Based questions are constructed from semantic events in videos, with representative examples shown in Fig. 5. We first divide videos into temporal segments and use the target VLM itself to generate a lightweight caption for each segment, consisting of a short summary of its visual content. Candidate events used for counterfactual construction are drawn from unrelated videos, reducing visual and semantic similarity between events and making the answer choices more distinguishable. Importantly, all captions are generated by the VLM being optimized, requiring no external captioning model.

Based on these segment-level event descriptions, we construct four types of questions:

- **Segment Localization.** A temporal segment is replaced with an unrelated video segment, and the model is asked to identify the replaced quarter, as shown in Fig. 3(1). Counterfactual pairs replace different quarters while keeping the question and answer choices fixed.
- **Event Identification.** The model is asked to identify the event occurring in a specified temporal quarter (Fig. 5). Counterfactual views contain different events in the queried quarter, resulting in different correct answers.
- **Event Repetition.** One event is repeated across multiple temporal quarters, and the model is asked to identify the repeated event (Fig. 5). Counterfactual pairs repeat different candidate events.
- **Event Ordering.** The model is asked which candidate event occurs first in the video (Fig. 5). Counterfactual pairs alter the temporal arrangement of the candidate events while preserving the question and answer choices.

Together, these event-based questions cover complementary aspects of video understanding, including event recognition, temporal localization, event matching, and temporal ordering over natural video content.

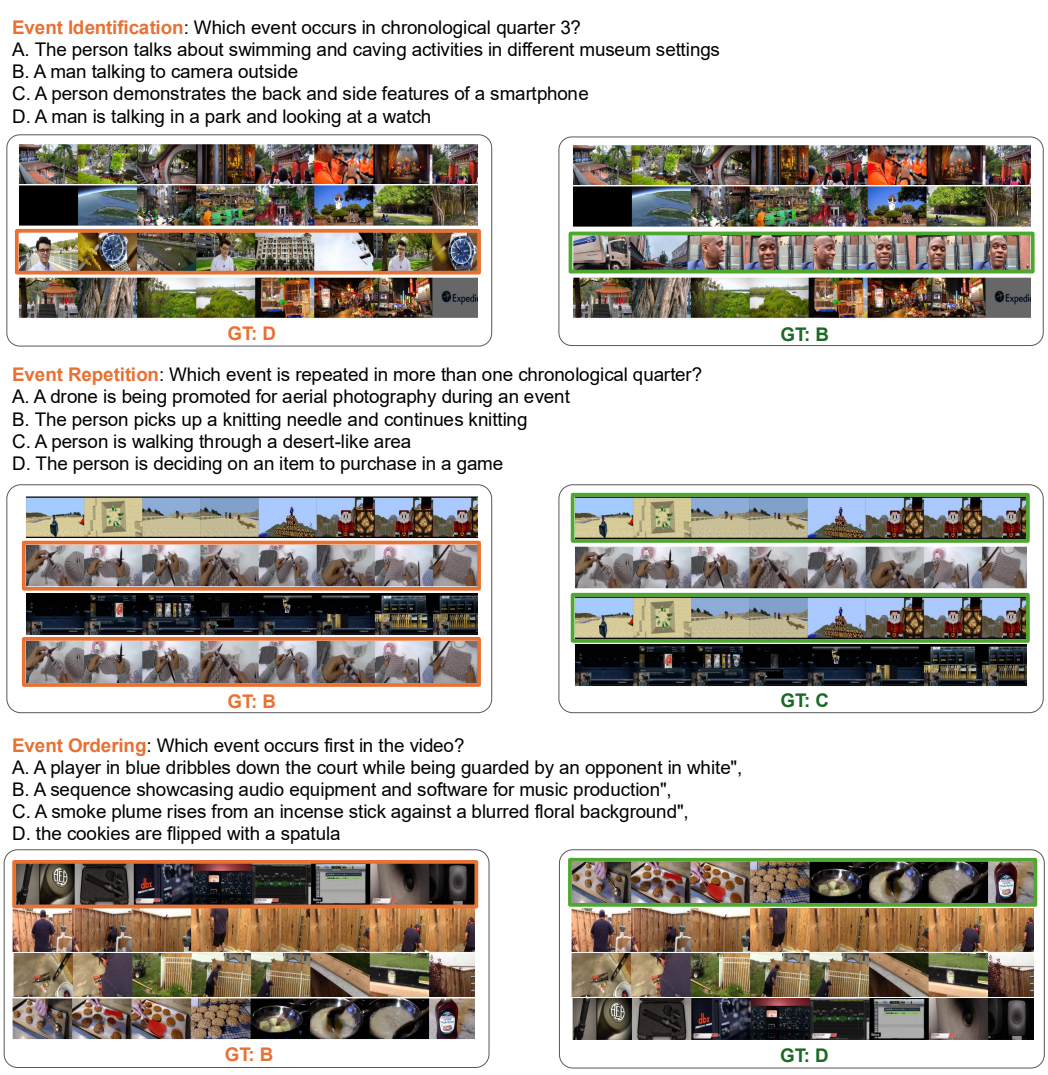


Figure 5: Examples of event-based counterfactual question pairs. Answer choices are generated through lightweight segment-level captioning by the target VLM itself. Answer-determining segments are inserted from unrelated videos, flipping the ground-truth answer while preserving the linguistic context.

Marker-Based Questions. Marker-based questions introduce controlled visual markers into the original video, providing precisely controllable visual evidence and deterministic ground-truth answers without human annotation, as illustrated in Fig. 6.

- **Marker Counting.** A marker is inserted into different numbers of sampled frames, and the model is asked how many frames contain it, as shown in Fig. 3(1).
- **Marker Identification.** A colored geometric marker is inserted into selected frames, and the model is asked to identify the inserted marker (Fig. 6). Counterfactual pairs use different marker appearances.
- **Marker Localization.** A marker is inserted into a specific sampled frame, and the model is asked to identify its temporal location (Fig. 6). Counterfactual pairs place the marker at different temporal positions.
- **Quarter Counting.** A marker is inserted into selected temporal quarters, and the model is asked how many quarters contain it (Fig. 6).

Together, these marker-based questions provide controlled supervision for visual recognition, fine-grained temporal localization, and evidence aggregation at different temporal granularities.

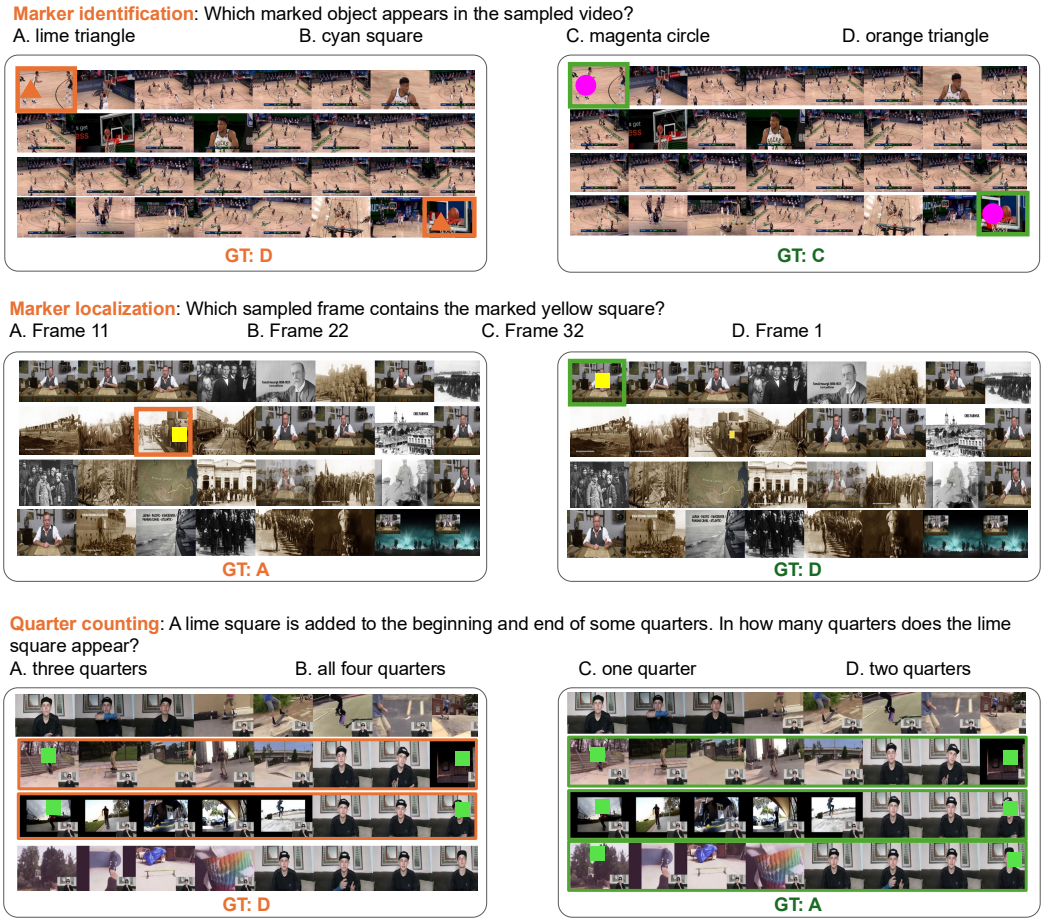


Figure 6: Examples of marker-based counterfactual pairs, where controlled marker changes flip the ground-truth answer while preserving the underlying video and linguistic context.

Table 8: Distribution (%) of question types selected by intervention-aware pair selection.

Model	Dataset	Seg. Loc.	Event ID.	Event Rep.	Event Ord.	Marker Count.	Marker ID.	Marker Loc.	Quarter Count.
Qwen3-VL-8B	LongVideoReason	6.3	1.6	27.6	7.1	34.6	4.7	15.0	3.1
	LongVideoBench	7.0	3.1	23.4	10.9	36.7	0.8	18.0	0.0
	VideoMME	6.3	5.5	25.8	7.0	41.4	0.8	11.7	1.6
VideoZoomer	LongVideoReason	9.4	7.0	31.3	6.3	18.0	2.3	23.4	2.3
	LongVideoBench	3.1	7.0	28.1	9.4	19.5	1.6	28.9	2.3
	VideoMME	3.9	8.6	25.8	7.0	11.7	1.6	39.1	2.3

C.2 Analysis of Selected Pairs

Table 8 analyzes the question types retained by intervention-aware pair selection. The selected distributions differ substantially across models. For Qwen3-VL-8B, marker counting is consistently the most frequently selected type, accounting for 34.6-41.4% of the selected pairs across the three benchmarks. In contrast, VideoZoomer selects substantially more marker-localization pairs, whose proportion increases from 23.4% on LongVideoReason to 39.1% on VideoMME. Event repetition is also consistently selected for both models, accounting for 23.4-31.3% of the selected pairs. These differences indicate that pair selection does not simply favor a fixed set of question templates, but adapts to the visual responsiveness and intervention potential exhibited by each model and target dataset.

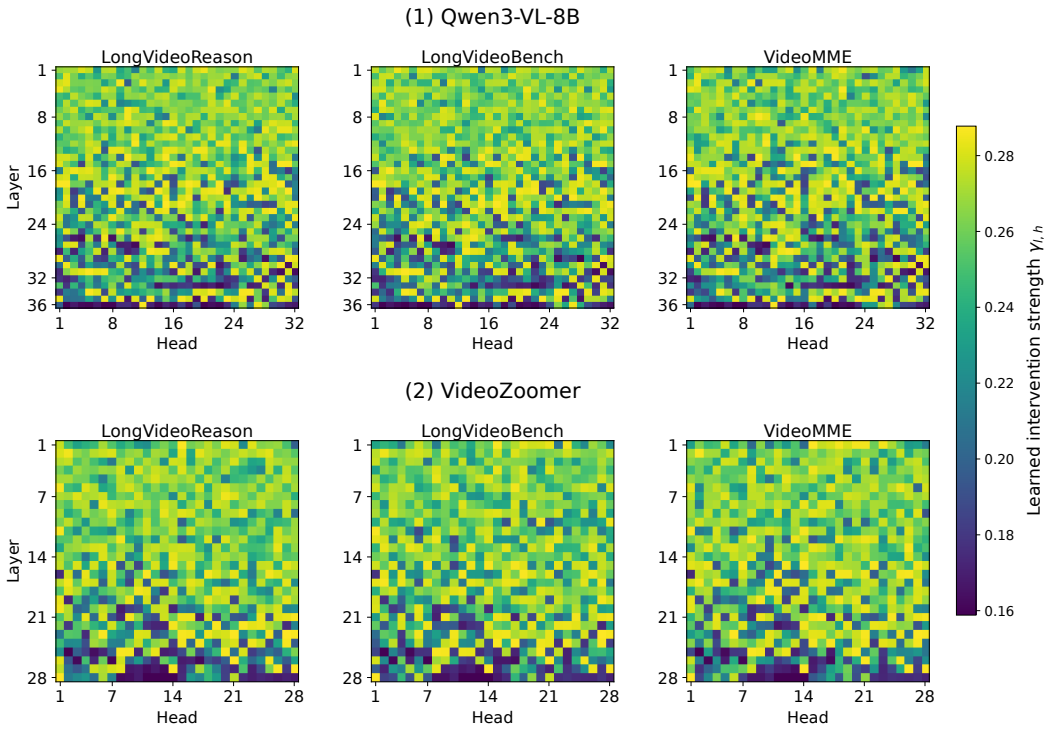


Figure 7: Visualization of the learned head-wise intervention strengths $\gamma_{l,h}$ across three video benchmarks for Qwen3-VL-8B and VideoZoomer.

C.3 Ablation Study of Pair Types

Table 9 compares the two categories of counterfactual questions. Using either category alone improves accuracy over no intervention, but neither consistently dominates across accuracy and VIS. Combining both categories provides the most balanced improvement, achieving the highest overall accuracy and lowest VIS on both benchmarks. These results suggest that the two categories provide complementary supervision, with their combination improving visual reliance more effectively.

Table 9: Ablation of different counterfactual question categories.

Question Category	LongVideoReason			LongVideoBench		
	Sen. ↑	All ↑	VIS ↓	Sen. ↑	All ↑	VIS ↓
No intervention	66.8	69.8	18.9	48.3	49.5	6.5
Event-based only	69.7	71.4	15.3	50.3	52.4	7.4
Marker-based only	**71.0**	70.8	16.5	51.1	52.3	6.8
Event-based + Marker-based	70.5	**72.1**	**14.3**	**52.3**	**53.1**	**4.5**

D Visualization Results

Analysis of Learned Intervention Parameters. Fig. 7 visualizes the learned head-wise intervention strengths for Qwen3-VL-8B Bai et al. (2025a) and VideoZoomer Ding et al. (2026). The learned strengths vary substantially across attention heads, with more pronounced head-wise differentiation in later layers. Moreover, similar patterns emerge across the three benchmarks, indicating that VisIO consistently assigns different intervention strengths to individual heads rather than uniformly scaling visual contributions.

Qualitative Analysis of Visual Responsiveness. Fig. 8 provides an intuitive example of S^{res}. The two counterfactual pairs are constructed from the same video with the same question template and differ primarily in the inserted marker color. Nevertheless, they receive substantially different

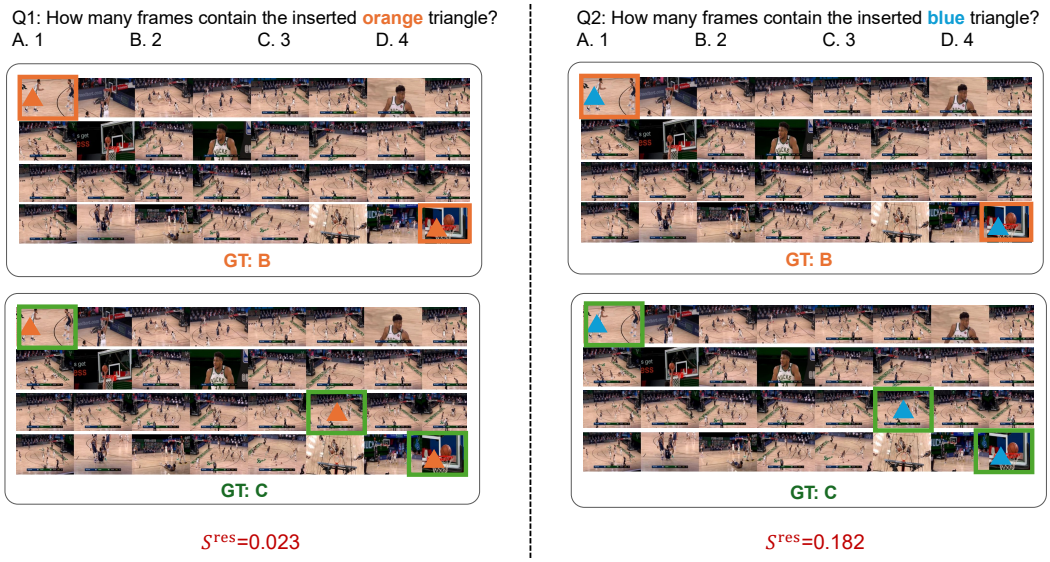


Figure 8: Illustration of visual responsiveness (S^{res}). The two counterfactual pairs differ primarily in marker color, yet exhibit substantially different responsiveness scores.

responsiveness scores (0.023 vs. 0.182). In the right example, the blue marker is more visually distinguishable from the video background, which may lead to a stronger change in the model's answer preference across the two counterfactual views. This example illustrates that seemingly similar counterfactual pairs can induce different levels of model responsiveness, and S^{res} provides a model-dependent measure for identifying such differences.

E Additional Implementation Details

VIS Evaluation Details. To compute the Visual Insensitivity Score (VIS), we evaluate each question under four input conditions while keeping the question and answer choices unchanged: (1) the original video, (2) a randomly selected unrelated video, (3) blank visual input, and (4) text-only input.

For each condition, we extract the model's final multiple-choice prediction. We first parse the option enclosed by `<answer>` and `</answer>`. If these tags are absent or malformed, we apply answer-oriented fallback patterns, such as "the final answer is C," "the correct answer is C," a boxed option, or a standalone option letter on the final line. We do not extract arbitrary option letters mentioned within the reasoning trace. The set of valid options is determined by the answer choices of each question (e.g., A–D or A–E).

Following Eq. 2, VIS is computed only over samples that are answered correctly under the original-video condition. A sample is assigned a visual-insensitivity indicator of 1 if and only if all four conditions produce valid predictions and the four predicted options are identical. Otherwise, the indicator is 0. In particular, ***refusals, irrelevant responses, and outputs from which no valid final option can be reliably extracted are treated as invalid predictions and cannot be counted as visual-insensitive***. VIS is the mean of this indicator over samples correctly answered under the original-video condition.

Video Sampling. For both intervention optimization and evaluation, videos are sampled at 0.5 FPS with at most 64 frames. The sampled frames are provided to the VLM together with the multiple-choice question and answer choices.

Optimization Prompt. During intervention optimization, we directly optimize the likelihood of the ground-truth option *without generating a reasoning trace*. We use the system prompt ``You are a helpful assistant.'' and format the user instruction as:

```
Answer the multiple-choice question based on the provided
video frames.  Choose one option from A, B, C, or D. Write
your reasoning inside <think> and </think>, then write the
final option letter inside <answer> and </answer>.
```

The video frames, question, and answer choices are included in the input. We then teacher-force an empty reasoning block, `<think></think>`, followed by the opening `<answer>` tag, and compute the logits of the candidate option letters at the next token position. For each counterfactual pair, the paired loss in Eq. 9 is computed from the corresponding ground-truth option probabilities. This avoids generating or supervising a reasoning trace during optimization and allows gradients to be propagated directly to the head-wise intervention parameters.

Inference Prompt. During evaluation, we use the same system prompt and multiple-choice instruction, but allow the model to autoregressively generate both its reasoning and final answer:

```
Answer the multiple-choice question based on the provided
video frames.  Choose one option from A, B, C, or D. Write
your reasoning inside <think> and </think>, then write the
final option letter inside <answer> and </answer>.
```

For questions with five answer choices, the candidate set is extended from A–D to A–E accordingly. We use greedy decoding without sampling, with a maximum generation length of 2,048 tokens. The final prediction is extracted using the parsing procedure described above, prioritizing the option enclosed by `<answer>` and `</answer>` when available.

Intervention Implementation. During inference, the learned intervention remains active at every autoregressive decoding step. Therefore, the intervention affects the generation of both the reasoning tokens within `<think>` and the final answer within `<answer>`, rather than being restricted to the first generated token or the final answer token. The same learned head-wise parameters are shared across all decoding steps and remain fixed throughout inference.