STARVE THE MLP, FEED THE ATTENTION: RESOLVING VISION-KNOWLEDGE CONFLICTS IN VIDEOLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern video models have significantly advanced the field of video understanding. However, as these models generate responses, they overly depend on intrinsic textual priors rather than actual visual evidence, fundamentally restricting their performance. This over-reliance, called *vision-knowledge conflicts*, is especially evident in videos depicting counterfactual physics, where models hallucinate standard physics rules rather than processing the actual counterfactual visual input. Our mechanistic analyses reveal that as Transformer layers deepen, these language priors progressively dominate the residual stream, effectively overshadowing the visual tokens. To resolve this issue, we propose **VERTEX** (**V**isual **E**vidence **R**estoration and **T**extual **EX**clusion), a zero-shot, inference-time intervention framework. VERTEX first isolates physical motion to generate binary kinematic masks, which are directly utilized to construct distinct latent geometric subspaces. During inference, the framework actively rebalances the model by strengthening its focus on actual visual evidence while filtering out parametric textual biases. Evaluations on counterfactual physics benchmarks show that VERTEX reduces vision-knowledge conflicts, outperforming standard baselines and providing a more robust framework for grounded video understanding.

1 INTRODUCTION

Recent video models blindly trust priors. Video understanding has evolved from Video Large Language Models (VideoLLMs) (Lin et al., 2024; Maaz et al., 2024; Zhang et al., 2024) performing simple tasks to video reasoning models capable of complex reasoning (Feng et al., 2026a; Team, 2025; Feng et al., 2026b). However, both types of models can suffer from a critical flaw: the *vision-knowledge conflict* (Liu et al., 2024b). Since modern video architectures are mostly built upon text-only LLMs, they inherit strong textual priors that frequently overwhelm the visual evidence (Liu et al., 2024b; Leng et al., 2024; Wu et al., 2025). Instead of anchoring their predictions in the video content, these models may fall back on parametric knowledge when visual signals conflict. Such over-reliance on textual priors can lead to multimodal hallucinations, causing the model to generate predictions that are detached from the actual physical events occurring in the video.

Vison-knowledge conflict is difficult to detect using standard benchmarks, where many videos follow highly predictable, "commonsense" scripts (Yue et al., 2024; Liu et al., 2024c; Zellers et al., 2019). In such settings, a model may arrive at the correct answer by merely observing the first frame then relying on its textual prior. Counterfactual physical videos provide a more direct diagnostic setting since they deliberately place visual evidence in conflict with learned real-world expectations (Bansal et al., 2026; Puyin et al., 2026; Zhou et al., 2025). We therefore focus on "uncommonsense" videos (Li et al., 2026b), whose observed kinematics violate conventional physical priors, including temporally reversed events in which objects appear to move against gravity. When shown such an event, a visually grounded model should follow the observed motion, yet current VideoLLMs often revert to their textual priors and describe what would normally happen instead (Liu et al., 2024b; Ortu et al., 2026; Lietzow et al., 2026). Thus, uncommonsense physical understanding provides a direct setting for exposing vision-knowledge conflicts.

To uncover the internal mechanics driving this issue, we conduct a layer-wise analysis and demonstrate that video models suffer from a severe magnitude imbalance in the residual stream during generation. Specifically, while the Attention blocks successfully tracks the visual evidence in deeper

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.

layers, the MLP blocks overwhelm this signal by outputting parametric textual biases at higher magnitudes. This creates a dual-manifold conflict in which the textual prior overrides the grounded visual signal, explaining why conventional single-stream steering or localized ablations fail.

To address this dual-manifold conflict, we introduce **VERTEX** (**V**isual **E**vidence **R**estoration and **T**extual **EX**clusion), a training-free framework designed to directly intervene on the competing visual and textual signals. VERTEX is built upon three methodological pillars. First, Objective Kinematic Anchoring (OKA) applies dense point tracking to project physical motion onto the visual token grid, producing binary masks that separate temporal dynamics from static background content. Second, Targeted Subspace Extraction (TSE) identifies and separates the latent subspaces governing grounded kinematic evidence and parametric text priors. Finally, Orthogonal Dual-Stream Rectification (ODR) intervenes on the deeper residual stream by reinforcing the attention pathway with the isolated visual signal while orthogonally suppressing the textual prior in the MLP outputs.

Extensive experiments across leading VideoLLMs demonstrate that VERTEX significantly restores visual grounding in a training-free manner. On counterfacutal physics benchmarks, our method yields substantial performance gains, efficiently balancing the residual stream.

In summary, the key contributions of our work are as follows:

- We formalize vision-knowledge conflicts in VideoLLMs, utilizing counterfactual physical videos to quantify the reliance of these models on parametric priors over visual evidence.
- We identify a mechanistic source of the vision-knowledge conflict as a dual-manifold imbalance in deep Transformer layers. While self-attention preserves grounded visual evidence, increasingly strong textual priors from the MLP blocks overwhelm this signal in the residual stream.
- We propose VERTEX, a training-free, inference-time intervention for vision-knowledge conflicts. VERTEX reinforces visual evidence in attention while suppressing textual priors in MLPs, improving grounding across VideoLLMs and counterfactual physics benchmarks.

2 RELATED WORKS

Video Large Language Models. The transition from standard Video Large Language Models (VideoLLMs) (Yang et al., 2025a; Lin et al., 2024; Wang et al., 2025) to Video Reasoning Models marks a crucial shift toward "System-2" logical deduction (Zhang et al., 2025). By leveraging Reinforcement Learning and test-time compute (Snell et al., 2024), VLRMs generate extensive internal reasoning trajectories prior to response generation. Architectures such as Video-R1 (Feng et al., 2026a), ERNIE-4.5-VL (Team, 2025), and OneThinker (Feng et al., 2026b) integrate explicit reasoning blocks to iteratively chart temporal state changes, track objects, and verify causal relationships before summarizing a video event. However, as these models generate increasingly long reasoning sequences to process complex temporal dynamics, they become highly susceptible to losing continuous visual grounding. This growing reliance on parametric memory over visual evidence motivates our investigation into temporal vision-knowledge conflicts.

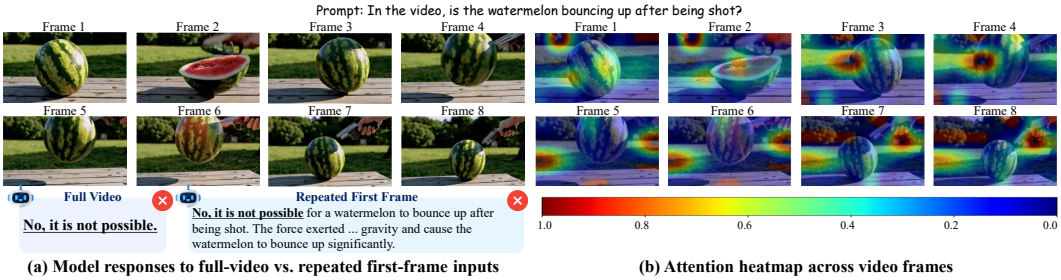
Vision-Knowledge Conflict in MLLMs. Vision-knowledge conflict occurs when pre-trained textual priors override explicit visual evidence (Liu et al., 2024b). Mechanistically, Ortu et al. (2026) attribute this to a layer-wise imbalance, where early Attention blocks extract counterfactual visual truths, but MLP blocks override it with parametric textual prior. In VideoLLMs, mitigation primarily targets general hallucinations via decoding penalties (Gao et al., 2026b; Fan et al., 2026; Gao et al., 2026a) or motion-based grounding (Wu et al., 2026). While recent benchmarks effectively diagnose kinematic failures (Liu et al., 2024d; Xue et al., 2026; Puyin et al., 2026), they offer no structural solutions. Our work moves beyond and analyzes how video models prioritize textual physics priors, dynamically resolving the conflict through a training-free, dual-manifold intervention. We provide further discussion of related work in Appendix B.

3 ANALYSES

Quantitative Evidence under Visual and Temporal Perturbations. To assess the visual grounding abilities of video models, we evaluate specialized video reasoning models (Feng et al., 2026a;

Table 1: Accuracies on the VideoHallu-Physics dataset with diverse video inputs.

Models	Standard	Reversed	Scrambled	Text-Only	First-Frame-Only
Video Reasoning Models					
Video-R1	59.18	59.69	64.80	46.94	58.16
Qwen3-VL-8B-Thinking	64.80	60.20	42.86	34.69	63.27
InternVL3.5-8B-Thinking	67.86	69.90	68.88	42.35	65.31
Standard VideoLLMs					
InternVL3.5-8B	67.86	69.90	58.67	68.88	66.33
Qwen3-VL-8B	65.31	59.69	63.27	65.31	58.16
LLaVA-NeXT	48.98	47.45	48.98	37.36	50.51



(a) Model responses to full-video vs. repeated first-frame inputs **(b) Attention heatmap across video frames**

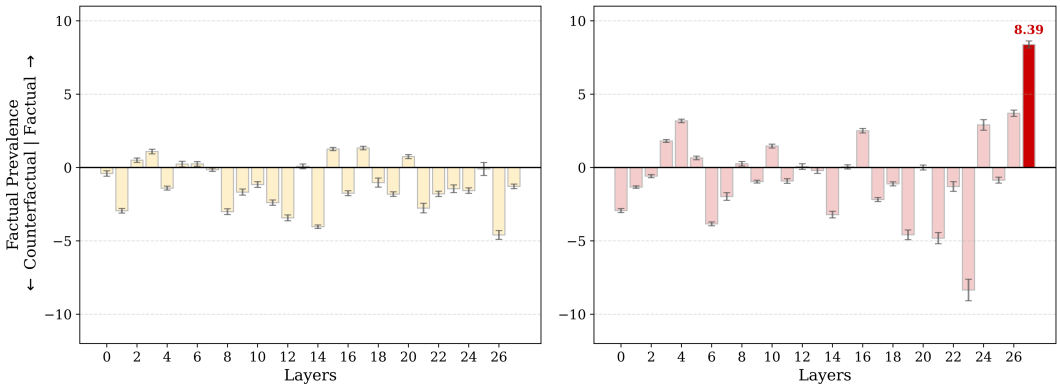
Figure 1: Behavioral and mechanistic evidence demonstrating the model's reliance on textual priors over objective visual grounding. **(a)** Even though the first frame is repeatedly provided as video input, the model hallucinates a dynamic physical event. **(b)** While the model initially attends to the salient visual objects, its spatial focus rapidly diffuses and drifts during later frames.

Yang et al., 2025a; Team, 2025) and standard VideoLLMs (Wang et al., 2025; Liu et al., 2024a) across visually constrained and temporally corrupted variants of the VideoHallu-Physics benchmark[1] (Table 1). Our analysis uncovers striking inconsistencies in how current models handle visual input. While modern reasoning architectures clearly react to visually-ablated inputs, many other models show surprisingly little accuracy loss when fed only a static first frame (`First-Frame-Only`) or no video at all (`Text-Only`). These systems rely far more on linguistic priors than actual visual context. Furthermore, even among models that do factor in video input, shuffling (`Scrambled`) or reversing the frame order (`Reversed`) often maintains performance close to the standard baseline. Ultimately, regardless of how much individual architectures depend on visual evidence, their capability for genuine physical tracking remains superficial, causing them to lean heavily on textual priors instead of empirical evidence.

Qualitative and Mechanistic Evidence of Visual Abandonment. This over-reliance on linguistic priors is qualitatively demonstrated in Figure 1(a). When presented with a sequence consisting solely of a duplicated first frame, the model still hallucinates a dynamic physical event (asserting that the bullet "penetrates the watermelon"). The resulting output is almost identical to its full video prediction, directly demonstrating how textual priors can completely substitute counterfactual visual inputs. Tracing the attention heatmaps across video frames (Figure 1(b)) reveals a severe decay in visual grounding. While the model successfully focuses on the object in early frames (Frames 3 and 4), its attention rapidly diffuses into the background during the later frames (Frames 5 to 8). After predicting the outcome using parametric textual priors, the model stops looking for visual evidence. This allows textual biases to override physical reality, resulting in visual-knowledge conflicts.

Internal Subspace Probing via Logit Lens. We apply the Logit-Lens (nostalgebraist, 2020) technique to analyze the layer-wise tension between the Attention and MLP blocks during the process. We define *Factual Prevalence* as the raw logit difference between the factual hallucination token (v_f) and the counterfactual visual evidence token (v_c). For an intermediate hidden state h_l at layer l and the model's unembedding matrix W_U, the "factual prevalence" is calculated as

[1]Filtered from VideoHallu (Li et al., 2026b) to isolate instances where `Question_category` is "Physics" and the sub-category is "Physics of Movement". Full dataset details are provided in Appendix E.



(a) Layer-wise factual prevalence of Attention block **(b) Layer-wise factual prevalence of MLP block**

Figure 2: Diagnosing the Mechanics of Vision-Knowledge Conflict in Video-R1. **(a)** Attention layers exhibit a strong counterfactual prevalence, tracking the real video. **(b)** MLP blocks maintain a high factual prevalence, relying on textual priors. At the output layer, the MLP's factual prevalence magnitude completely overrides that of Attention block's counterfactual prevalence.

$(W_U h_l)_{v_f} - (W_U h_l)_{v_c}$. A positive value indicates a bias toward the textual prior, while a negative value favors the visual evidence. As shown in Figure 2(a), the Attention blocks successfully identify the visual evidence present in the video by showing decent counterfactual prevalence. However, the MLP blocks display an overwhelming factual prevalence, relying on the parametric textual prior, as illustrated in Figure 2(b). Ultimately, in the final decoding layers, the magnitude of the MLP's factual prevalence is extremely higher than the Attention block's counterfactual prevalence, thus the textual prior effectively overrides visual reality at the end, causing the hallucination. For more analysis, we conducted a Visual-Knockout experiment, which is specifically explained in Appendix C.

4 METHOD

To mitigate the vision-knowledge conflicts in VideoLLMs, we introduce **VERTEX**, an inference-time framework that intervenes the residual stream. VERTEX consists of three core pillars: **(a)** Objective Kinematic Anchoring (OKA) to isolate physical motion from textual priors, **(b)** Targeted Subspace Extraction (TSE) to identify the distinct geometric subspaces of textual bias and visual evidence, and **(c)** Orthogonal Dual-stream Rectification (ODR) to effectively steer the model's representations. The overall pipeline of VERTEX is illustrated in Figure 3.

4.1 OBJECTIVE KINEMATIC ANCHORING (OKA)

As the decoding phase progresses, the visual signal weakens, prompting the model to fall back to its parametric textual biases. Rather than relying on the LLM's biased visual tokenization, OKA employs CoTracker (Karaev et al., 2024) to explicitly ground the model's representations in actual physical motion.

Given an input video $V \in \mathbb{R}^{T\times H\times W\times 3}$, CoTracker outputs the continuous coordinate trajectories of moving pixels. To map this into the discrete LLM latent space, we construct a 3D boolean kinematic mask $\mathcal{M}_{\text{kin}} \in \{0,1\}^{T_{\text{tok}}\times H_{\text{llm}}\times W_{\text{llm}}}$, where T_{tok} denotes the number of temporal tokens. Given the initial visual patch grid of size $H_{\text{grid}} \times W_{\text{grid}}$ produced by the vision encoder, we calculate the target latent spatial dimensions by accounting for the specific spatial downsampling factor γ of the chosen model's vision-language adapter: $H_{\text{llm}} = \lfloor \frac{H_{\text{grid}}}{\gamma} \rfloor, W_{\text{llm}} = \lfloor \frac{W_{\text{grid}}}{\gamma} \rfloor$.

Afterwards, we compute the temporal Euclidean displacement of each tracked point. Let Δx and Δy denote the frame-to-frame pixel displacements. We normalize this relative to the raw video dimensions (H, W). Points whose movement exceeds a scale-invariant threshold $\tau = 2.0$ are chosen as active kinematics:

$$\text{Displacement}_{\text{norm}} = \frac{\sqrt{\Delta x^2 + \Delta y^2}}{\sqrt{H^2 + W^2}} \times 1000 > \tau. \tag{1}$$

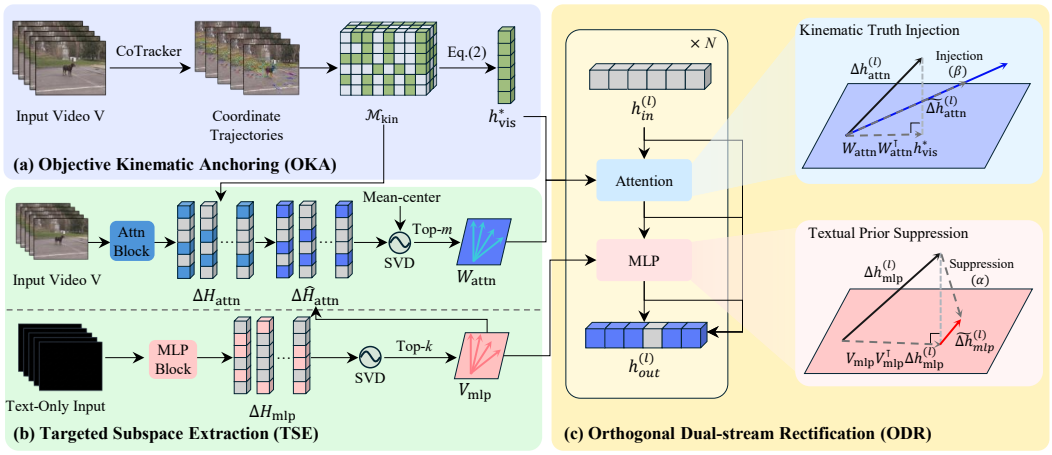


Figure 3: Overall pipeline of VERTEX. VERTEX consists of three steps: **(a)** Objective Kinematic Anchoring (OKA), **(b)** Targeted Subspace Extraction (TSE), and **(c)** Orthogonal Dual-stream Rectification (ODR).

The continuous coordinates (x, y) are then mapped onto the discrete spatial token grid, using boundary clipping to account for aspect-ratio padding: $(x_{\text{tok}}, y_{\text{tok}}) = (\lfloor (\frac{x}{W}) \cdot W_{\text{llm}} \rfloor, \lfloor (\frac{y}{H}) \cdot H_{\text{llm}} \rfloor)$. Similarly, we map the continuous frame index $t \in [0, T-1]$ to the corresponding discrete temporal token slice t_{tok} by proportional temporal discretization $t_{\text{tok}} = \min\left(\lfloor t \cdot \frac{T_{\text{tok}}}{T} \rfloor, T_{\text{tok}} - 1\right)$.

We then activate the boolean mask at the corresponding spatio-temporal coordinate by setting $\mathcal{M}_{\text{kin}}(t_{\text{tok}}, y_{\text{tok}}, x_{\text{tok}}) = 1$. Let $Z_{\text{vis}} \in \mathbb{R}^{N_{\text{vis}} \times d}$ denote the flattened sequence of raw visual embeddings from the model's vision encoder. To prevent cross-modal contamination and preserve uncorrupted kinematic evidence, we derive Z_{vis} solely from the layer-0 embeddings rather than layer-wise hidden states (see Appendix D for further information).

By flattening $\mathcal{M}_{\text{kin}}$ to align with sequence length N_{vis}, we isolate the pure kinematic truth vector h^*_{vis} via masked mean pooling over the token index j. To maintain numerical stability during tracking failures or in purely static scenes, the system reverts to global spatial pooling whenever all point displacements fall below the threshold τ. The operation is formulated as:

$$h^*_{\text{vis}} = \begin{cases} \frac{1}{\sum_{j=1}^{N_{\text{vis}}} \mathcal{M}_{\text{kin}}[j]} \sum_{j=1}^{N_{\text{vis}}} \left(\mathcal{M}_{\text{kin}}[j] \cdot Z_{\text{vis}}[j]\right), & \text{if } \sum_{j=1}^{N_{\text{vis}}} \mathcal{M}_{\text{kin}}[j] > 0 \\ \frac{1}{N_{\text{vis}}} \sum_{j=1}^{N_{\text{vis}}} Z_{\text{vis}}[j], & \text{otherwise.} \end{cases} \tag{2}$$

This produces a localized kinematic vector derived purely from visual data, isolating it from the model's pre-trained textual priors.

4.2 TARGETED SUBSPACE EXTRACTION (TSE)

Before intervening the residual stream, we isolate the specific geometric subspaces that encode the MLP's textual biases and the Attention's true kinematic signals.

Textual Prior Subspace ($\mathbf{V}_{\text{mlp}}$). We capture the exact geometry of the textual bias by executing the model on text-only inputs. To minimize memory overhead, we execute a single forward pass, restricting the sequence length M strictly to the number of tokens of the input text query. We collect the output projections of the MLP blocks, $\Delta H_{\text{mlp}} \in \mathbb{R}^{T_{\text{prompt}} \times d}$, across the deep layers, specifically layers after the critical layer L_{crit}. The specific values of L_{crit} for each model are defined in Section 5.1 .Applying Singular Value Decomposition (SVD) on the uncentered activations isolates the absolute geometric direction of the textual prior:

$$\Delta H_{\text{mlp}} = U_{\text{mlp}} \Sigma_{\text{mlp}} V_{\text{mlp}}^{\top}. \tag{3}$$

We select the top-k principal components to define the textual prior subspace $\mathbf{V}_{\text{mlp}} \in \mathbb{R}^{d \times k}$.

Attention-Kinematic Subspace ($\mathbf{W}_{\text{attn}}$). Let $O_{\text{attn}} \in \mathbb{R}^{N_{\text{vis}} \times d}$ represent the Attention outputs on input videos. We construct the unpurified kinematic matrix $\Delta H_{\text{attn}} \in \mathbb{R}^{P \times d}$ by row-concatenating

the embeddings $O_{\text{attn}}[j]$ exclusively where $\mathcal{M}_{\text{kin}}[j] = 1$. To isolate pure physical geometry, we orthogonally project this signal away from the textual prior:

$$\Delta\hat{H}_{\text{attn}} = \Delta H_{\text{attn}} - \Delta H_{\text{attn}} \mathbf{V}_{mlp} \mathbf{V}_{mlp}^{\top}. \tag{4}$$

We then mean-center the purified activations $\bar{H}$ and perform SVD:

$$\Delta\hat{H}_{\text{attn}} - \bar{H} = U_{\text{attn}} \Sigma_{\text{attn}} W_{\text{attn}}^{\top}. \tag{5}$$

The top-m principal components form the Attention-kinematic Subspace $\mathbf{W}_{\text{attn}} \in \mathbb{R}^{d \times m}$.

The choice of SVD centering reflects the distinct goals of each subspace. Uncentered SVD is applied to the MLP blocks to maintain the absolute magnitude of the textual prior, while for the Attention blocks, mean-centered SVD is used on the visual tokens to center the latent space. Further information about it is discussed in Appendix D.

4.3 Orthogonal Dual-stream Rectification (ODR)

Our analyses states that beyond L_{crit}, the latent state progressively collapses into a textual prior subspace, as the model's linguistic bias overpowers the visual evidence, resulting in vision-knowledge conflicts. Leveraging the geometric subspaces identified by OKA and TSE, we propose Orthogonal Dual-stream Rectification (ODR), an inference-time intervention that corrects this conflict.

Textual Prior Suppression (MLP). The standard MLP output, $\Delta h_{\text{mlp}}^{(l)}$, is heavily biased toward the hallucination manifold in deeper layers. To decouple this bias from the valid model pathway, we construct an orthogonal projection matrix utilizing the textual-prior subspace $\mathbf{V}_{\text{mlp}} \in \mathbb{R}^{d \times k}$. Operating directly on the residual stream, we project the MLP output away from the textual prior subspace:

$$\Delta\tilde{h}_{\text{mlp}}^{(l)} = \left(I - \alpha \mathbf{V}_{\text{mlp}} \mathbf{V}_{\text{mlp}}^{\top}\right) \Delta h_{\text{mlp}}^{(l)}, \tag{6}$$

where I is the identity matrix and $\alpha \in [0, 1]$ is the suppression coefficient. This projection weakens the dominant textual bias while leaving the orthogonal features necessary for language fluency intact, such as basic linguistic and grammatical features.

Kinematic Truth Injection (Attention). Concurrently, the degraded visual evidence in the self-attention output, $\Delta h_{\text{attn}}^{(l)}$, must be actively recovered. We retrieve the uncorrupted kinematic truth vector h_{vis}^{*} from Section 4.1. To prevent out-of-distribution artifacts, we project this signal onto the Attention-kinematic subspace $\mathbf{W}_{\text{attn}} \in \mathbb{R}^{d \times m}$. This ensures the vector remains strictly within the model's established physical geometry before being added to the Attention residual:

$$\Delta\tilde{h}_{\text{attn}}^{(l)} = \Delta h_{\text{attn}}^{(l)} + \beta \left(\mathbf{W}_{\text{attn}} \mathbf{W}_{\text{attn}}^{\top} h_{\text{vis}}^{*}\right), \tag{7}$$

where β represents the injection coefficient.

Rectified Residual Update. Let $h_{\text{in}}^{(l)}$ be the residual input at layer l. To avoid harming the model's normalization, VERTEX applies its interventions step-by-step during the forward pass, focusing only on the deeper layers ($l \geq L_{\text{crit}}$). We first update the Attention output with the kinematic injection:

$$h_{\text{mid}}^{(l)} = h_{\text{in}}^{(l)} + \Delta\tilde{h}_{\text{attn}}^{(l)}. \tag{8}$$

After the MLP processes normalization, we intercept its output to suppress the textual prior:

$$h_{\text{out}}^{(l)} = h_{\text{mid}}^{(l)} + \Delta\tilde{h}_{\text{mlp}}^{(l)}. \tag{9}$$

This two-step formulation grounds the model without disrupting standard causal LLM forwarding.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate VERTEX across four video datasets manually designed to test vision-knowledge conflicts. For baseline physical dynamics, we use the "Physics of Movement" subset of VideoHallu (Li et al., 2026b), termed VideoHallu-Physics, a manually-filtered subset of VideoPHY2 (Bansal et al., 2026) focusing strictly on physical movements, named VideoPHY-Movement.

Table 2: Quantitative Results across benchmark datasets. Red numbers with upward arrows (▲) indicate absolute accuracy gains, and blue numbers with downward arrows (▼) indicate absolte accuracy loss when applying VERTEX compared to the standard baseline models.

Models	VideoHallu-Physics	VideoPHY2-Movement	SSV2-Reversed	Real-Reversed
Standard Video Models				
Qwen3-VL-8B	61.73	69.47	21.16	14.00
LLaVA-NeXT-7B	48.98	27.33	0.24	8.00
InternVL3.5-8B	67.86	41.60	40.76	24.00
Standard Video Models w/ VERTEX				
VERTEX $_{\text{Qwen3-VL-8B}}$	64.29 ▲ 2.59	72.53 ▲ 3.06	16.77 ▼ 4.39	16.00 ▲ 2.00
VERTEX $_{\text{LLaVA-NeXT-7B}}$	51.46 ▲ 2.48	32.53 ▲ 5.20	9.87 ▲ 9.63	68.00 ▲ 60.00
VERTEX $_{\text{InternVL3.5-8B}}$	69.90 ▲ 2.04	42.40 ▲ 0.80	42.18 ▲ 1.42	18.00 ▼ 6.00
Video Reasoning Models				
Video-R1-7B	59.18	33.07	5.72	14.00
Qwen3-VL-8B-Thinking	64.80	55.20	8.16	20.00
InternVL-3.5-8B-Thinking	67.86	39.20	34.46	16.00
Video Reasoning Models w/ VERTEX				
VERTEX $_{\text{Video-R1-7B}}$	66.33 ▲ 7.15	40.40 ▲ 7.33	18.62 ▲ 12.90	18.00 ▲ 4.00
VERTEX $_{\text{Qwen3-VL-8B-Thinking}}$	65.82 ▲ 1.02	63.47 ▲ 8.27	21.65 ▲ 13.49	26.00 ▲ 6.00
VERTEX $_{\text{InternVL-3.5-8B-Thinking}}$	69.39 ▲ 1.53	41.20 ▲ 2.00	39.49 ▲ 5.03	20.00 ▲ 4.00

To assess how models handle temporal acausality, we introduce SSV2-Reversed (Goyal et al., 2017) and a custom Real-Reversed dataset, which contain temporally inverted videos that contradict standard physics. Dataset curation and input prompt details are provided in Appendix E.

Baselines and Metrics. We evaluate our method against six recent VideoLLMs, including three specialized video reasoning models (Video-R1, Qwen3-VL-8B-Thinking, InternVL-3.5-8B-Thinking) alongside three standard video models (Qwen3-VL-8B, LLaVA-NeXT-7B, InternVL-3.5-8B). This selection allows us to evaluate the intervention across both traditional visual-instruction-tuned networks and recent reasoning models. For all experiments, we utilize accuracy as our evaluation metric. Because our benchmarks are structured as constrained multiple-choice or binary verification tasks, accuracy provides a direct, unambiguous measure of whether a model successfully grounds its prediction in visual evidence or succumbs to textual priors. We utilize GPT-4o as the LLM-judge to evaluate the models' responses.

Implementation Details. All experiments are conducted on $4\times$ RTX A6000 GPUs with deterministic seeding. Videos were inputted as number of frames of $N_{frames} = 32$ for Qwen3 models and $N_{frames} = 8$ for other models. TSE isolates the textual (V_{mlp}) and kinematic (W_{attn}) subspaces via SVD, leaving the top $k = 4$ and $m = 4$ components, respectively. ODR scaling hyperparameters (α, β) are tuned via 2D grid search (See Appendix F for more information). The critical layer (L_{crit}) scales with architecture depth: 28 for Qwen3, 24 for LLaVA-NeXT, and 20 for Qwen2-based models (Video-R1, InternVL-3.5).

5.2 QUANTITATIVE RESULTS

We evaluate the proposed VERTEX framework using accuracy as the primary metric, with full quantitative results detailed in Table 2. Specifically, on the VideoHallu-Physics dataset, VERTEX consistently improves performance across all evaluated architectures, indicating that the intervention successfully suppresses textual priors and sustains visual grounding during the reasoning process. This trend extends to the VideoPHY2-Movement benchmark, where weakening the textual prior in the MLP layers yields accuracy recovery. Together, these improvements confirm that structurally isolating physical kinematics can effectively override linguistic bias across diverse video models.

On the SSV2-Reversed and Real-Reversed datasets, baseline models generally fail to recognize reversed time progression and revert to forward-chronological priors. In contrast, VERTEX demonstrates consistent accuracy improvements across most architectures. By anchoring the Attention mechanism explicitly to the reversed kinematic trajectory, the framework forces the model to process the true visible temporal sequence, directly mitigating the temporal blindness observed in standard reasoning models. Notably, VERTEX corrects a severe bias present in LLaVA-NeXT-7B on the



Figure 4: Qualitative results on VERTEX applied. When VERTEX is applied, the video models do not generate responses based on textbook physics priors.

Real-Reversed dataset, yielding an extreme accuracy gain. We provide a detailed analysis of this phenomenon in Appendix D.

Overall, we observe a trend where video reasoning models are more positively affected by our method compared to standard VideoLLMs. This amplified performance gain suggests that explicitly grounding visual evidence is exceptionally critical for models involving the reasoning process, as it prevents their logical reasoning paths from diverging based on hallucinated textual assumptions.

5.3 QUALITATIVE RESULTS

Figure 4 illustrates examples where the baseline model suffers from vision-knowledge conflicts, abandoning visual evidence to hallucinate based on its toxic textual prior. Conversely, our method resolves this conflict via dual-stream rectification. By orthogonally suppressing the divergent textual prior while injecting the discrete kinematic mask ($\mathcal{M}_{\text{kin}}$) into the Attention stream, VERTEX keeps its deep latent states strictly anchored to the moving object's spatial tokens ($x_{\text{tok}}, y_{\text{tok}}$). This visually faithful generation proves that starving the MLP and feeding the Attention effectively saves the model from prior-induced hallucinations. More qualitative examples are provided in Appendix J.

5.4 ABLATION STUDIES

To rigorously isolate the contribution of each methodological pillar in VERTEX, we conduct a component-wise ablation study on the VideoHallu-Physics benchmark. Because the framework operates as a sequential pipeline, we ablate individual modules by substituting them with naive or standard architectural alternatives. First, to evaluate the OKA phase, we replace its *Kinematic* mask with a *Dense* visual extraction baseline, where the targeted Attention subspace (W_{attn}) is computed using the entire unmasked video token grid rather than isolating temporally dynamic objects. Next,

Table 3: Ablation studies of VERTEX components on VideoHallu-Physics. We use Video-R1 as the baseline model. ✓ and × indicate enabled and disabled components, respectively.

Configuration	OKA	TSE	ODR-α	ODR-β	Acc.
Baseline	–	–	×	×	59.18
w/o OKA	Dense	Instance	✓	✓	63.27
w/o TSE	Kinematic	Global	✓	✓	59.18
w/o Suppression	Kinematic	Instance	×	✓	62.24
w/o Injection	Kinematic	Instance	✓	×	64.29
VERTEX (Full)	**Kinematic**	**Instance**	✓	✓	**66.33**

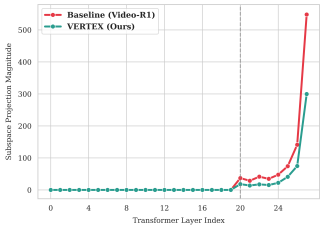


Figure 5: Toxic Prior Alignment of the baseline Video-R1 (●) and VERTEX (●).

we ablate TSE by substituting its core *Instance-Aware* mechanism with a *Global* baseline. This involves calculating a single, universal textual prior (V_{mlp}) and kinematic subspace (W_{attn}), rather than computing them dynamically per-sample. Finally, to validate the dual-stream equilibrium of the ODR phase, we disrupt the push-pull intervention by selectively zeroing out our hyperparameters ($\alpha = 0.0$ or $\beta = 0.0$).

As shown in Table 3, the complete VERTEX framework substantially outperforms the baseline, confirming that all three core components are essential for robust counterfactual reasoning. Ablating the OKA module in favor of a dense visual baseline degrades performance by conflating critical motion dynamics with static background noise, underscoring the necessity of localized kinematic anchoring. Furthermore, replacing the instance-aware TSE with a global average causes a severe regression to baseline levels, proving that textual priors are inherently input-dependent. A universal global subspace is simply too broad to effectively isolate and suppress these sample-specific biases. Finally, isolating the dual decoding mechanisms reveals that while textual prior suppression yields a stronger standalone benefit than kinematic truth injection, optimal physical reasoning strictly requires the simultaneous application of both mechanisms. Further analyses on other models and datasets are provided in Appendix G.

To empirically validate our projection operator, we measure the toxic prior alignment, which is defined as the projection magnitude onto the hallucination subspace ($\|\mathbf{V}_{\text{mlp}}\mathbf{V}_{\text{mlp}}^T h\|_2$). As Figure 5 illustrates, the baseline Video-R1 model displays a sudden rapid increase across the terminal reasoning layers (20–27), peaking above 550. This trend highlights the scale at which the language prior overwhelms visual evidence just before generation. Similar trends appear for other baseline models, as illustrated in Appendix I.

In contrast, applying VERTEX strictly suppresses this trajectory throughout the intervention span. Because our operator explicitly subtracts the textual prior projection, the comparatively shortened curve confirms that the textual prior is effectively suppressed. This validates that our dual-stream rectification successfully prevents the hidden states from collapsing into language bias, ensuring the latent space remains fully receptive to the injected physical kinematics. Toxic prior alignment experiments on other models are provided in Appendix I.

6 CONCLUSION

In this work, we addressed a fundamental architectural problem in VideoLLMs: vision-knowledge conflicts. Our analysis demonstrated that during continuous video understanding, models systematically ignore physical visual evidence and overly rely on parametric textual prior that outputs hallucinated responses. To mitigate this, we introduced VERTEX, a novel intervention built on the principle of dual-stream rectification. By orthogonally suppressing the MLP subspace while simultaneously injecting discrete kinematic truth into the Attention stream, our method effectively starves the toxic prior and feeds the visual evidence. Our empirical evaluations confirm that this approach not only weakens prior-induced hallucinations but also strengthens strict visual grounding in counterfactual scenarios where language biases contradict physical reality. Ultimately, VERTEX provides a robust, geometry-grounded architectural foundation for the next generation of hallucination-free VideoLLMs. We hope our work motivates further studies on vision-knowledge conflicts in VideoLLMs at mechanistic level. We discuss the limitations of our work in Appendix A.

AI USE STATEMENT

In this work, we used generative AI tools for aiding or polishing writing and roughly drafting sections of the paper. We have not used generative AI tools for generating experiment results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for writing and debugging Python scripts for model inference and evaluation, configuring hyperparameter grid searches, and formatting LaTeX tables and figures. We have reviewed all AI-assisted work. Specifically, all LLM-generated code was manually executed, verified, and tested for correctness by the authors on our local GPU clusters to ensure the integrity of the evaluation pipelines. Furthermore, all AI-assisted textual drafts (such as appendix content and failure mode analyses) were thoroughly reviewed, edited, and fact-checked against our quantitative findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research aims to improve VideoLLM reliability by mitigating physical hallucinations, fostering the safe deployment of grounded AI systems. While our evaluation utilizes public YouTube videos depicting human behavior, it involves no direct human interaction or sensitive Personally Identifiable Information (PII), strictly adhering to platform terms alongside established synthetic benchmarks. Moreover, although VERTEX effectively corrects kinematic and temporal hallucinations, it does not resolve underlying social biases encoded in multimodal representations, highlighting the need for complementary fairness interventions in future work. Finally, while dynamic Temporal Subspace Extraction (TSE) introduces minor inference overhead, it completely bypasses the massive carbon emissions associated with full-parameter retraining.

REPRODUCIBILITY STATEMENT

To ensure the complete reproducibility of our experimental findings, we have thoroughly documented all methodological details, dataset curation procedures, and model configurations throughout the main text and appendices. Appendix D provides comprehensive statistics and extraction methodologies for our evaluation datasets. The exact prompt templates, including the zero-shot instructions and reasoning triggers for the thinking-mode models, are explicitly detailed, too. Furthermore, all critical hyperparameter settings, specifically the textual suppression coefficient (α), kinematic injection intensity (β), and the critical intervention layer thresholds (L_{crit}) optimized for each evaluated architecture, are tabulated in the ablation studies and experimental setup sections. We plan to publicly releasing the complete code upon acceptance.

REFERENCES

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026.

Linfeng Fan, Yuan Tian, Ziwei Li, and Zhiwu Lu. Stear: Layer-aware spatiotemporal evidence intervention for hallucination mitigation in video large language models. *arXiv preprint arXiv:2604.03045*, 2026.

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.

Yuansheng Gao, Wenbin Xing, Jiahao Yuan, Kaiwen Zhou, Han Bao, Zonghui Wang, and Wenzhi Chen. Multitop: Learning to patch visual tokens to mitigate hallucinations in video large multimodal models. *arXiv preprint arXiv:2606.11792*, 2026a.

Yuansheng Gao, Jinman Zhao, Tong Zhang, Xingguo Xu, Wenbin Xing, Han Bao, Zonghui Wang, and Wenzhi Chen. Enhancing video representations with spatiotemporal-semantic residual to mitigate hallucinations in video large multimodal models. *arXiv preprint arXiv:2601.22574*, 2026b.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Xingming Li, Ao Cheng, Qiyao Sun, Xixiang He, Xuanyu Ji, Runke Huang, and Qingyong Hu. MLLMs get it right, then get it wrong: Tracing and correcting late-layer textual bias. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 1370–1378. International Joint Conferences on Artificial Intelligence Organization, 8 2026a. doi: 10.24963/ijcai.2026/153. URL https://doi.org/10.24963/ijcai.2026/153. Main Track.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *Advances in Neural Information Processing Systems*, 38:76046–76078, 2026b.

Niclas Lietzow, Danielle Bitterman, Carsten Eickhoff, William Rudman, and Michal Golovanevsky. Vision-default, prior-override: Causal mechanisms of perception-knowledge conflict in vision-language models. *arXiv preprint arXiv:2606.28273*, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b4008025c2182bfe16fcc8566ee14d64-Abstract-Conference.html.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jen-tse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. Insight over sight: Exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024d.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

nostalgebraist. interpreting GPT:: the logit lens. LessWrong blog post, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2026-09-20.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14109–14130, 2026.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33174–33184, 2026.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Baidu ERNIE Team. Ernie 4.5 technical report, 2025.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xiyang Wu, Zongxia Li, Jihui Jin, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. Mass: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9433–9443, 2026.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, et al. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*, 2025.

Zihui Sherry Xue, Romy Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *Advances in Neural Information Processing Systems*, 38:90925–90955, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724. IEEE, 2019.

Duzhen Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Xiuyi Chen, Yingying Zhang, et al. From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai, 2025c. *URL https://arxiv. org/abs/2512.01989*, 2025.

Appendix: Table of Contents

Starve the MLP, Feed the Attention

A Limitations **15**

B Extended Related Work **15**

B.1 Video reasoning and physical grounding. 15

B.2 Vision-knowledge conflicts and their mechanisms. 15

B.3 Hallucination mitigation and representation intervention. 16

C Visual Knockout **16**

C.1 Causal Probing via Visual Knockout. 16

C.2 Temporal Disconnect in the Residual Stream. 16

D Discussions **17**

D.1 Impact of Feature Source for Kinematic Anchor Extraction. 17

D.2 Uncentered vs. Mean-Centered Extraction. 17

D.3 Selection of Intervention Layers (L_{crit}). 17

D.4 Dimensionality of Geometric Subspaces (k and m). 17

D.5 The Necessity of Dual-Stream Rectification. 17

D.6 Impact of Toxic Prior Suppression on Architecturally Constrained Models. 18

E Datasets **18**

F Hyperparameter Tuning **19**

G Further Analyses **20**

H Experiments on Commonsense Videos **20**

I Layer-wise Dynamics of Textual Prior Suppression **21**

J Additional Qualitative Results **21**

A LIMITATIONS

Although VERTEX provides an efficient approach to mitigate vision-knowledge conflicts in video models, it still contains some limitations. First, VERTEX shows substantial computational and memory overhead. The framework relies on a multi-stage preprocessing pipeline that requires executing OKA for object tracking, followed by the TSE process that performs SVD to extract subspace matrices for every individual video. During inference, these matrices must be iteratively loaded from disk and projected at every critical layer at runtime. This per-video caching requirement increases memory usage compared to standard inference, making the current iteration of VERTEX highly complex for large-scale or streaming video applications.

The performance of the kinematic injection stream is also strictly dependent on the quality and availability of explicit motion vectors. Because VERTEX was designed specifically to mitigate physical hallucinations, it relies heavily on the assumption that a video contains trackable, macro-level kinematics. In scenarios involving extreme object occlusion, extreme camera movements, the OKA masks may capture noise rather than meaningful physical movement. Consequently, the extracted subspaces may fail to provide the necessary counterfactual grounding, limiting the framework's versatility beyond physics-based hallucination benchmarks.

Finally, the framework exhibits a high degree of hyperparameter brittleness. The textual suppression coefficient, the kinematic injection intensity, and the critical intervention layer threshold must be precisely calibrated for each distinct model family and parameter size. An optimal intervention pairing for LLaVA-NeXT-7B (Liu et al., 2024a) does not seamlessly transfer to InternVL3.5-8B (Wang et al., 2025). This sensitivity dictates that adopting VERTEX for novel architectures requires rigorous empirical tuning before the toxic prior alignment can be reliably broken.

B EXTENDED RELATED WORK

B.1 VIDEO REASONING AND PHYSICAL GROUNDING.

Video Large Language Models (VideoLLMs) have progressed from video instruction following and cross-modal alignment (Maaz et al., 2024; Lin et al., 2024) to reinforcement learning for temporal and multimodal reasoning (Feng et al., 2026a;b). TempCompass and AoTBench evaluate sensitivity to temporal variation and directionality beyond static appearance (Liu et al., 2024d; Xue et al., 2026), while VideoHallu and VideoPhy-2 probe hallucinations and physical consistency in synthetic or generated videos (Li et al., 2026b; Bansal et al., 2026). Beyond evaluation, ArrowRL and VideoHallu investigate reinforcement learning to improve temporal sensitivity and reduce hallucinations, respectively (Xue et al., 2026; Li et al., 2026b). Explicit motion grounding offers a complementary approach, with CoTracker recovering point correspondences across frames (Karaev et al., 2024) and MASS combining tracking, spatial representations, and reinforcement fine-tuning for physical reasoning (Wu et al., 2026). Our evaluation uses movement-focused subsets and temporally reversed videos to test whether predictions follow observed events rather than familiar physical expectations.

B.2 VISION-KNOWLEDGE CONFLICTS AND THEIR MECHANISMS.

Vision-knowledge conflicts arise when parametric knowledge overrides contradictory visual evidence, as demonstrated using commonsense-conflicting image inputs (Liu et al., 2024b). Language priors are not uniformly harmful, however, and can support understanding when visual information is incomplete (Wu et al., 2025). Mechanistic studies link feed-forward computations to key-value memory, concept promotion, and factual association recall (Geva et al., 2021; 2022; Meng et al., 2022). In vision-language models, Ortu et al. (2026) analyze competing attention and MLP contributions that favor counterfactual visual evidence and factual associations, respectively, and rebalance them through attention-head interventions. Complementary work identifies late-layer textual overrides of correct intermediate visual predictions and recovers these predictions through Conflict-Aware Layer Reference Decoding (Li et al., 2026a). These findings motivate our layer-wise investigation of attention and MLP residual contributions in physical video understanding, without assuming that either component exclusively represents one modality.

B.3 HALLUCINATION MITIGATION AND REPRESENTATION INTERVENTION.

Inference-time mitigation includes contrasting original and distorted visual inputs in Visual Contrastive Decoding (Leng et al., 2024) and penalizing attention over-trust with retrospective reallocation in OPERA (Huang et al., 2024). For videos, ViSSRes and MultiToP learn auxiliary modules for residual representation refinement and selective token replacement, respectively (Gao et al., 2026b;a), while STEAR combines layer-aware evidence restoration with temporally perturbed counterfactual decoding (Fan et al., 2026). Activation-based approaches steer selected attention-head states toward truth-related directions (Li et al., 2023) or apply fixed visual and textual directions to the vision encoder and language decoder (Liu et al., 2025). Nullu identifies a low-dimensional hallucination subspace from differences between truthful and hallucinated representations, then suppresses it through MLP weight orthogonalization (Yang et al., 2025b). Unlike fixed steering directions or offline weight edits, VERTEX derives input-specific subspaces from the current query's text-only MLP activations and motion-selected attention outputs, without paired truthful and hallucinated examples. It attenuates the resulting textual-prior directions in MLP outputs while injecting a visual anchor projected onto an attention subspace orthogonalized against those directions, leaving pretrained weights unchanged.

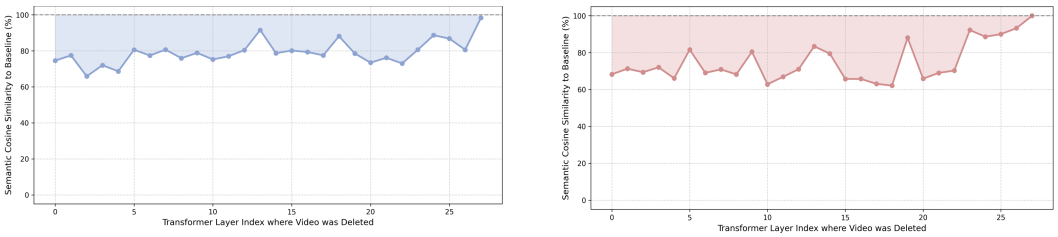
C VISUAL KNOCKOUT

C.1 CAUSAL PROBING VIA VISUAL KNOCKOUT.

To causally verify that visual tokens are ignored during deeper layers, we perform a layer-wise *Visual-Knockout* experiment. We systematically delete the visual embeddings from the residual stream at each layer and measure the semantic cosine similarity of the intermediate representations against an unablated forward pass. As shown in Figure 6(a), across the deep layers, the cosine similarity remains remarkably flat and close to 1.0, even when the video tokens are entirely erased. This directly demonstrates that once visual features are merged into early layers, the model's deeper layers actively ignore them, running their logic circuits on a purely textual rail.

C.2 TEMPORAL DISCONNECT IN THE RESIDUAL STREAM.

Further evidence of this temporal disconnect is explained using the visual-knockout analysis on reversed frames, as shown in Figure 6(b). When visual tokens are systematically ablated during understanding the reversed video, the cosine similarity of the intermediate representations remains flat and near 1.0, which is almost identical to the standard video graph. This indicates that the model is not merely misinterpreting the reversed temporal frames, but is completely disconnected from the visual sequence entirely during deep understanding. By blinding itself to the arrow of time, the model relies solely on its chronological text priors to force a conclusion.



(a) Visual knockout of standard video input **(b) Visual knockout of reversed video input**

Figure 6: Visual Knockout experiment of Video-R1 model on (a) standard videos and (b) reversed videos.

D DISCUSSIONS

D.1 IMPACT OF FEATURE SOURCE FOR KINEMATIC ANCHOR EXTRACTION.

In the OKA phase, we anchor our kinematic vector to the raw visual embeddings at layer 0 rather than layer-wise hidden states. This choice is empirically supported, as the layer-0 formulation achieved a higher accuracy compared to layer-wise embeddings. Visual tokens at layer 0 retain the raw spatial and kinematic features directly from the vision encoder. However, in deeper layers, self-attention naturally mixes these visual representations with textual tokens and parametric priors. Deriving our visual signal exclusively from layer 0 avoids this modality mixing, ensuring the intervention relies on direct video evidence rather than the text-biased representations formed later in the network.

D.2 UNCENTERED VS. MEAN-CENTERED EXTRACTION.

While TSE relies on SVD to isolate geometric subspaces, applying the exact same centering strategy to both the Attention and MLP streams is not the best choice. For the MLP textual prior ($\mathbf{V}_{\text{mlp}}$), we use *uncentered* SVD, because mean-centering the MLP activations would shift the origin of the latent space, capturing only the variance between different text prompts, not the absolute magnitude and fixed direction of the textual prior. In contrast, for the Attention-kinematic subspace ($\mathbf{W}_{\text{attn}}$), we use *mean-centered* SVD. Because visual tokens also contain static background information, uncentered SVD would extract a subspace dominated by these unnecessary features. Mean-centering factors out this static frame, forcing only the principal components to strictly capture the dynamic kinematic motion required for accurate physical understanding.

D.3 SELECTION OF INTERVENTION LAYERS (L_{crit}).

A naive application of ODR would intervene across all Transformer layers uniformly. However, our analysis states that interventions must be applied only to layers beyond a critical threshold ($l \geq L_{\text{crit}}$). The early and middle layers of multimodal LLMs are primarily responsible for basic syntactic structuring and low-level feature alignment between the vision and language modalities. Intervening too early disrupts this fundamental feature formation, often severely harming the model's output. By restricting ODR to the deeper layers where high-level understanding occurs, we achieve an optimal trade-off between physical grounding and linguistic coherence.

D.4 DIMENSIONALITY OF GEOMETRIC SUBSPACES (k AND m).

The choice of the number of principal components for the MLP prior (k) and the Attention kinematics (m) represents a trade-off between hallucination suppression and general understanding capability. If k is set too high, the orthogonal projection removes not only the counterfactual bias but also the critical semantic and grammatical features, damaging the model's performance. Conversely, if k is too low, the core textual prior survives. A similar trade-off exists for m: setting it too high injects noisy visual artifacts that confuse the generation, while setting it too low fails to provide the kinematic anchor that is strong enough to overcome the language prior.

D.5 THE NECESSITY OF DUAL-STREAM RECTIFICATION.

We formulated ODR to simultaneously suppress the MLP output and inject into the Attention output, rather than relying on a single-stream intervention. Suppressing the MLP prior without Attention injection successfully prevents the model from relying on its textual prior, but leaves it without sufficient visual evidence to generate a correct counterfactual response. Conversely, injecting the Attention kinematics without suppressing the MLP forces the visual evidence to battle with an overpowering textual prior, which the MLP usually wins. The dual-stream approach is strictly necessary: the MLP suppression clears the semantic space, allowing the Attention injection to safely guide the residual stream toward the physically grounded truth.

D.6 Impact of Toxic Prior Suppression on Architecturally Constrained Models.

As shown in Table 2, LLaVA-NeXT-7B (Liu et al., 2024a) consistently exhibits the lowest baseline performance among standard video models, especially struggling with counterfactual temporal tasks. This devastating performance stems from the model's frame-concatenation architecture. LLaVA-NeXT lacks robust spatiotemporal inductive biases, thus it overly relies on static, semantic priors. Consequently, on counterfactual physics tasks, the model exhibits severe bias, hallucinating forward-time events and performing vastly worse than random chance. However, the application of VERTEX yields a substantial absolute gain of 60.00 points on the Real-Reversed dataset. This extreme performance increase validates the efficacy of ODR. By explicitly suppressing the textual prior in the MLP layers and directly injecting kinematic vectors into the Attention, VERTEX successfully overrides the model's hallucinations, forcing it to ground its understanding in actual temporal motion rather than static textual biases.

Prompt Templates

VideoPHY2-Movement Prompt

```
Analyze the physical kinematics in this video. Does it accurately and realistically
depict the following event without any object morphing or physical hallucinations:
[caption]? Answer strictly Yes or No within <answer> </answer> tags.
```

SSV2-Reversed Prompt

```
Carefully observe the physical movement in the video.
Question: [Question]
Option A: [Standard commonsense expectation]
Option B: [Actual counterfactual visual evidence]
Which option accurately describes the physical reality of the video?
```

Real-Reversed Prompt

```
Do the events in this video look physically possible in the real world?
Please answer strictly with 'Option A' or 'Option B'.
Option A: Yes, the events look completely normal and physically possible.
Option B: No, the motion is physically impossible, as if the video is playing backward.
```

Figure 7: Input prompts used for counterfactual physics datasets. Blue text denotes input texts.

E Datasets

We conduct our evaluations across four targeted datasets. Each dataset is specifically curated to test physical commonsense, kinematic tracking, and temporal motion anomalies. We provide specific prompt templates for each dataset, and the dataset statistics are summarized in Table 4:

- **VideoHallu-Physics:** A targeted subset extracted from the VideoHallu (Li et al., 2026b) benchmark. To establish a baseline for the model's comprehension of natural physical dynamics, we strictly isolate and utilize videos exclusively from the "Physics of Movement" category. Structurally, each instance in this subset includes a short video paired with a specific question designed to check kinematic understanding, such as object trajectories, collisions, or physical state changes, alongside a ground-truth answer.
- **VideoPHY2-Movement:** To further assess the physical grounding capabilities of multimodal Large Language Models, we utilize the VideoPHY2-Movement dataset, comprising 750 manually filtered videos of physical movements from the original VideoPHY2 (Bansal et al., 2026) dataset. This dataset evaluates video understanding through a strict binary verification task using a unified prompt template, which is described in Figure 7. This constrained verification format forces the model to directly cross-reference the textual assertion against the objective visual evidence, isolating its ability to accurately verify true physical dynamics without hallucinating or succumbing to parametric confirmation bias.

- **SSV2-Reversed:** To evaluate the susceptibility of video models to temporal acausality, we construct the SSV2-Reversed benchmark by systematically filtering the Something-Something V2 (SSV2) (Goyal et al., 2017) dataset based on its "original template" annotations. Specifically, we isolate action categories representing irreversible physical processes, such as dropping an item or an object falling, where inverting the timeline explicitly yields counterfactual scenarios (e.g., a dropped object magically floating upwards). Through this targeted curation, we isolate 2,046 videos across eight directional physical interaction templates. As a preprocessing step, we temporally reverse the frame sequence of each video to generate these anomalous visual environments. To mitigate the autoregressive snowballing of parametric textual priors during evaluation, the dataset employs a constrained Multiple-Choice Question (MCQ) format using the unified prompt described in Figure 7.
- **Real-Reversed:** A custom dataset constructed from 50 real-world videos that are temporally reversed to create physically impossible motion and interactions. To directly probe the visual-knowledge conflict, we formulate the evaluation as a binary multiple-choice question-answering task. Models are asked whether the depicted events are physically possible in the real world, forcing them to choose between acknowledging the actual visual evidence or succumbing to their textual prior. To ensure systematic evaluation, all counterfactual question-answer pairs were generated using an advanced Large Language Model following a strict, unified prompt template described in Figure 7.

Table 4: Statistics of datasets. For each dataset, we report the number of questions, number of videos, and video length statistics (in seconds).

Dataset	Category	Questions	Videos	Min (s)	Max (s)	Mean (s)	Std (s)
VideoHallu	Physics	196	104	4.00	10.00	6.00	1.94
	Total	3233	549	4.00	10.00	5.84	1.70
VideoPHY2	Movement	750	750	3.20	6.12	4.73	0.91
	Total	3397	6800	3.20	6.12	4.55	0.99
SSV2	Reversed	2046	2046	0.50	9.42	4.04	1.10
	Total	220847	220847	0.67	14.17	3.82	1.05
Real Videos	Total	50	50	1.00	3.20	2.95	0.39

F HYPERPARAMETER TUNING

The effectiveness of the Orthogonal Dual-stream Rectification (ODR) phase relies on the equilibrium of its push-pull mechanism. Because intervention magnitudes are highly sensitive to a model's specific parametric latent geometry, we independently tune and report the intervention hyperparameters for all models.

The dual-stream intervention is controlled by two primary scalars:

- **Textual Suppression (α):** Controls the intensity of subtracting the targeted textual prior from the MLP output. Excessive suppression ($\alpha \geq 0.8$) risks representation collapse, such as formatting loss and language drift, while insufficient suppression fails to overcome the parametric hallucination prior.
- **Kinematic Injection (β):** Determines the magnitude of the objective kinematic truth vector injected into the Attention residual stream. If this scalar is set too high, out-of-distribution visual evidence overwhelms the target layers; if set too low, the temporal tracking evidence remains diluted by the background tokens.

To identify the optimal architectural configuration, we conducted a systematic 2D grid search over $\alpha \in \{0.2, 0.3, 0.4\}$ and $\beta \in \{0.1, 0.3, 0.5\}$. The grid demonstrates that VERTEX maintains performance improvements across a continuous neighborhood of hyperparameters, confirming that the intervention relies on robust geometric alignment rather than isolated, brittle configurations.

G FURTHER ANALYSES

We conducted further analyses across other evaluated benchmarks, as shown in Table 5, 6, 7

While the full VERTEX framework generally achieves optimal performance, results on the physical reasoning dataset VideoPHY2-Movement using Qwen3-Thinking reveal that both "w/o OKA" and the "w/o TSE" outperform the full zero-shot pipeline (Table 5). We attribute this to the constructive nature of textual priors in Chain-of-Thought (CoT) (Wei et al., 2022) architectures. Qwen3-Thinking uses its `<think>` phase to generate valid, step-by-step logic. For a physics-heavy dataset like VideoPHY2, these internal text representations are not just harmful biases, but they actively provide the physical rules required to understand the scene. Consequently, the full VERTEX intervention can cause destructive interference by overly suppressing these crucial reasoning pathways.

Our ablations demonstrate two effective mitigations for reasoning models. First, the "w/o OKA" ablation successfully grounds the Attention mechanism in visual evidence without starving the MLP of its valid logical priors. Alternatively, the Global TSE acts as a generalized low-pass filter, averaging subspaces across a calibration set to safely remove universal linguistic bias while leaving delicate instance-specific logic intact. This highlights that while normal Instance TSE remains the optimal strictly zero-shot baseline for standard VideoLLMs, intervention in CoT architectures requires either dynamic calibration of the suppression coefficient (α) or the use of global subspaces to preserve representational freedom.

Despite the lower performance of the full framework, the complete VERTEX pipeline still remains the superior systemic choice for general application. The "w/o OKA" ablation substitutes kinematic tracking with a dense mask, which also processes static background noise rather than isolating the true temporal motion necessary for robust physical grounding. Likewise, the "w/o TSE" ablation utilizes a Global subspace, which fundamentally breaks the strictly zero-shot nature of the framework by requiring prior access to an offline calibration dataset. Therefore, the full VERTEX pipeline with kinematic OKA and instance-aware TSE offers the optimal theoretical and practical balance.

Table 5: Ablation studies of VERTEX components on VideoPHY2-Movement.

Configuration	OKA	TSE	ODR-α	ODR-β	Acc.
Baseline (Qwen3-VL-Thinking)	–	–	×	×	55.20
w/o OKA	Dense	Instance	✓	✓	65.60
w/o TSE	Kinematic	Global	✓	✓	**67.73**
w/o Suppression	Kinematic	Instance	×	✓	61.47
w/o Injection	Kinematic	Instance	✓	×	62.40
VERTEX (Full)	**Kinematic**	**Instance**	✓	✓	63.47

Table 6: Ablation studies of VERTEX components on SSV2-Reversed. **Bold** indicates the best result and underline indicates the second best.

Configuration	OKA	TSE	ODR-α	ODR-β	Acc.
Baseline (Video-R1)	–	–	×	×	5.72
w/o OKA	Dense	Instance	✓	✓	16.52
w/o TSE	Kinematic	Global	✓	✓	**18.82**
w/o Suppression	Kinematic	Instance	×	✓	17.64
w/o Injection	Kinematic	Instance	✓	×	17.30
VERTEX (Full)	**Kinematic**	**Instance**	✓	✓	<u>18.62</u>

H EXPERIMENTS ON COMMONSENSE VIDEOS

To evaluate the versatility of the VERTEX framework beyond uncommonsense videos, we extend our experiments to a dataset of real-world commonsense videos. Specifically, we evaluate the framework on the Real-Original dataset, comprising the 50 unaltered source videos that correspond to the Real-

Table 7: Ablation studies of VERTEX components on Real-Reversed.

Configuration	OKA	TSE	ODR-α	ODR-β	Acc.
Baseline (InternVL-3.5-8B-Thinking)	–	–	×	×	16.00
w/o OKA	Dense	Instance	✓	✓	8.00
w/o TSE	Kinematic	Global	✓	✓	18.00
w/o Suppression	Kinematic	Instance	×	✓	**20.00**
w/o Injection	Kinematic	Instance	✓	×	**20.00**
VERTEX (Full)	**Kinematic**	**Instance**	✓	✓	**20.00**

Table 8: Performance comparison with and without VERTEX.

Model	w/o VERTEX	w/ VERTEX	Δ
InternVL3.5-8B	74.00	92.00	18.00
InternVL3.5-8B-Thinking	86.00	86.00	0.00
Video-R1	86.00	90.00	4.00
Qwen3-VL-8B-Thinking	90.00	88.00	-2.00

Reversed benchmark. We conducted these experiments across four videoLLMs, including two video reasoning models: Video-R1, InternVL-3.5, and InternVL-3.5-Thinking, Qwen3-VL-8B-Thinking.

Our results in Table 8 demonstrate that VERTEX maintains the high accuracies similar to the baselines across most models. Because standard VideoLLMs naturally excel on unaltered footage where visual evidence aligns perfectly with their textual priors, the primary objective of this evaluation is to confirm non-degradation. The fact that performance remains highly robust indicates that VERTEX does not destructively suppress commonsense logic when it is factually correct. This confirms that our ODR is not narrowly restricted to uncommonsense scenarios, but serves as a safe, generalized intervention that reliably grounds generation in visual truth without sacrificing normal open-world reasoning capabilities.

I LAYER-WISE DYNAMICS OF TEXTUAL PRIOR SUPPRESSION

We further conducted the Toxic prior alignment experiment we discussed in Section 5.4 into other models. The resulting graphs are illustrated in Figure 8. Consistent with our initial findings, all baseline VideoLLMs exhibit a sharp escalation in subspace projection magnitude within the deeper Transformer layers, indicating a over-reliance on parametric textual priors during late-stage decoding. However, the application of our framework successfully mitigates this issue. By intervening post-L_{crit} (denoted by the dashed vertical lines), VERTEX consistently suppresses the magnitude of the toxic prior subspace, ensuring that visual evidence remains grounded through the final decoding layers.

J ADDITIONAL QUALITATIVE RESULTS

We provide further qualitative examples in Figure 9 to demonstrate the efficacy of VERTEX across diverse video reasoning scenarios. Results show how the framework successfully mitigates severe vision-knowledge conflicts. In standard configurations, baseline models frequently hallucinate outcomes driven by pre-trained linguistic assumptions, completely overriding the visible kinematic evidence.

For instance, when given a video of an orange ball magically launching back into the hand, the baseline model incorrectly outputs Option A, saying that the events are physically possible, demonstrating a severe vision-knowledge conflict. Instead of relying on the actual reversed visual kinematics, the baseline defaults to its parametric textual prior, assuming standard forward-temporal physics of a ball bouncing to the ground. Conversely, VERTEX successfully anchors the reasoning process to the true temporal motion and responds that the events are physically impossible (Option B), aligning perfectly with the ground-truth and yielding a contextually faithful prediction.

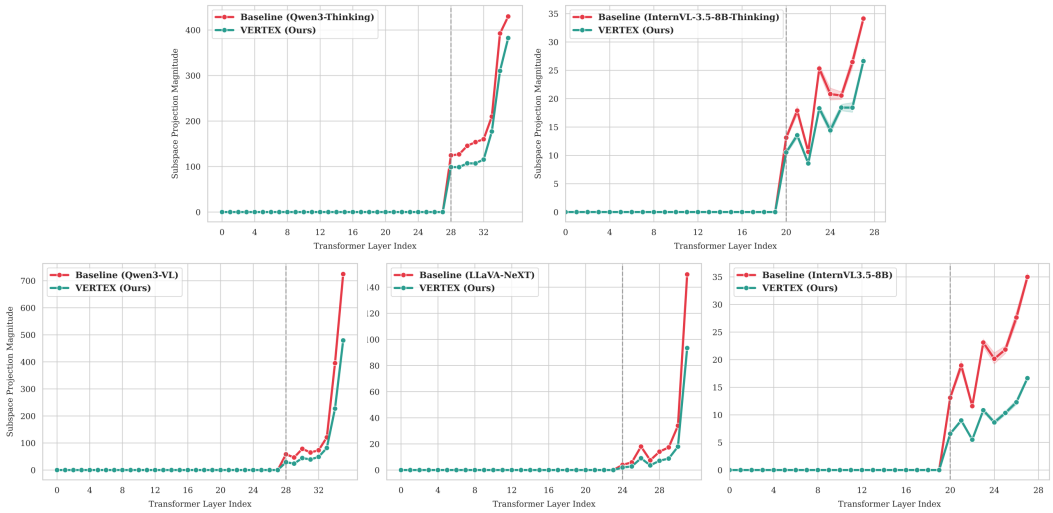


Figure 8: Toxic Prior Alignment of baseline VideoLLMs (●) and VERTEX (●)

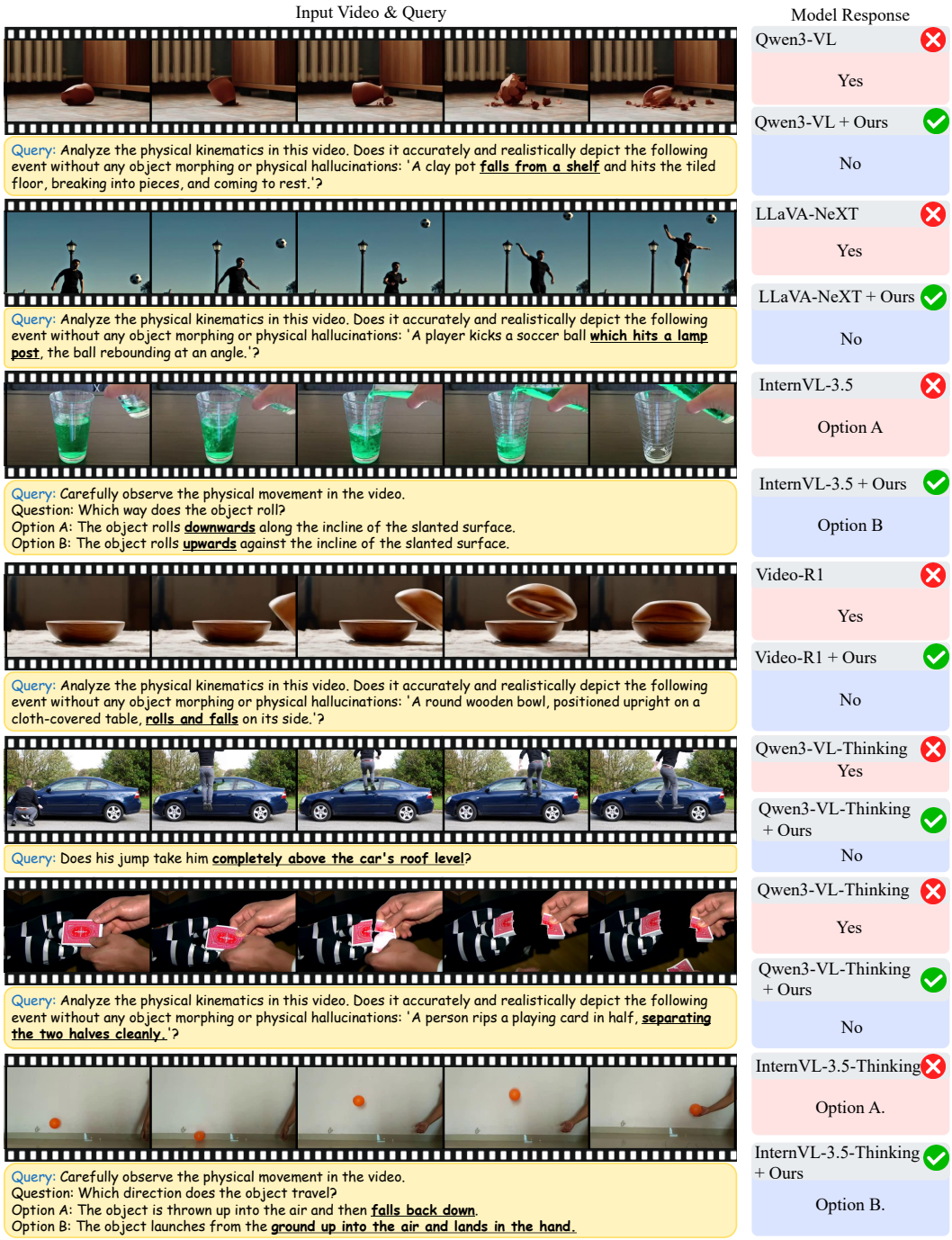


Figure 9: Qualitative comparison of responses from the baseline model and VERTEX.