TRAINING-FREE POSITIVE SPATIOTEMPORAL ENHANCEMENT FOR MITIGATING TEMPORAL HALLUCINATIONS IN VIDEO-LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Large Language Models (Video-LLMs) recognize objects and scenes reliably, yet they still misread event order, motion direction and object-state transitions, which produces temporal hallucinations. We refer to the change of the same object's state across frames as *object-centric motion*. Existing training-free approaches boost *per-frame* object-centric evidence to reduce hallucination. Yet, they do not explicitly boost *cross-frame* object-centric motions, the temporal relationships of the same object. Our empirical analysis suggests that this information is partially decodable from object-level visual representations yet, the final layer does not decode them in the final prediction. From this diagnosis, we propose Feature Enhance, a training-free scheme that strengthens object-centric visual evidence by emphasizing foreground regions and making inter-frame changes more explicit in the visual representations. It reinforces foreground regions that persist across frames and injects foreground-gated inter-frame feature residuals, leaving the model parameters untouched. Across three backbones, LLaVA-OneVision, Qwen2.5-VL and LLaVA-Video, Feature Enhance improves average accuracy over the corresponding unmodified model in all six backbone and benchmark combinations on VideoHallucer and EventHallusion, with gains of up to 4.67 and 4.00 percentage points.

1 INTRODUCTION

Visual-language research has progressively evolved from image captioning and visual question answering (You et al., 2016; Antol et al., 2015) to large-scale video description and video question answering (Xu et al., 2016; Yu et al., 2019), explicit spatiotemporal video modeling (Bertasius et al., 2021), and chat-centric video understanding (Li et al., 2024b). Building on these foundations, Multimodal Large Language Models (MLLMs) have extended from static images to video (Bai et al., 2025; Li et al., 2024a), supporting video dialogue (Maaz et al., 2024), dense captioning (Yang et al., 2023), temporal grounding (Ren et al., 2024; Huang et al., 2023), and long-form understanding (Song et al., 2024), with strong results on comprehensive benchmarks (Li et al., 2024c; Fu et al., 2025) and applications in embodied agents (Driess et al., 2023) and autonomous driving (Sima et al., 2024). Yet these models remain unreliable when answers depend on how visual content evolves over time rather than on what appears in individual frames. These failures are not explained by model scale alone. On MotionBench, GPT-4o using text only reaches 33% accuracy against a 25% random baseline (Hong et al., 2026), while on SpookyBench, where information is encoded purely through temporal correlations between noise-like frames, humans score above 98% and state-of-the-art Video-VLMs score 0% (Upadhyay et al., 2026).

Existing training-free methods intervene at inference either by suppressing predictions that survive temporal corruption (Zhang et al., 2024; Wu et al., 2025) or by reweighting spatially salient evidence (Li et al., 2025a; Du et al., 2025). Both manipulate *where* the model attends or *which* prediction to suppress, implicitly assuming that redirecting attention recovers the correct temporal decision. We argue this view is incomplete: attending to the right object in a frame does not imply representing its object-centric motion, *i.e.*, the temporal relationship between states of the same object across frames. Observing an object at location B does not reveal whether it moved from A to B, from C to

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.

B, or remained stationary. Object-centric motion, in other words, is left implicit in the representation the decoder actually reads.

To better understand why temporal hallucinations occur, we investigate three hypotheses concerning foreground attention and the use of motion information during answer generation. First, models often pay insufficient attention to foreground regions when making incorrect temporal predictions. Correct predictions show greater foreground attention and maintain it more consistently across frames than incorrect predictions. Second, models encode object-centric motion in their visual representations but this information is not reliably reflected at the answer-decision state. Linear probing provides initial evidence that motion-direction information is encoded in visual tokens, supporting the encoding aspect of this hypothesis. Motivated by these findings, we test whether making foreground and motion-related evidence more explicit in the visual input improves temporal predictions.

Motivated by this analysis, we propose Feature Enhance, a training-free scheme that strengthens positive evidence inside the visual embedding space. It combines instantaneous foreground weighting with a persistence signal that emphasizes spatial regions repeatedly identified as foreground across sampled steps. It also injects foreground-gated inter-frame feature residuals, computed at the same spatial positions across adjacent sampled temporal steps after any backbone-specific temporal merging. These residuals provide a coarse proxy for foreground temporal variation without explicitly establishing same-object correspondence. The combined representation is passed to the language model without updating model parameters or introducing additional decoding branches.

Our contributions are summarized as follows:

- A diagnostic analysis of how foreground attention and motion-direction information relate to temporal prediction errors in Video-LLMs.
- Feature Enhance, a training-free visual-embedding intervention that combines BiRefNet - derived foreground weighting, cross-frame foreground persistence, and foreground-gated inter-frame feature residuals.
- An evaluation on three Video-LLM backbones showing improved benchmark-average accuracy over the unmodified models in all six backbone–benchmark combinations on VideoHallucer and EventHallusion, with gains of up to 4.67 and 4.00 percentage points, respectively. Additional evaluation on MotionBench shows a 1.90-point macro-average gain on Qwen2.5-VL-7B.

2 Related Work

2.1 Video Hallucination Benchmarks

Several benchmarks evaluate hallucination and temporal reasoning in Video-LLMs. VidHalluc (Li et al., 2025a) covers action, temporal-sequence, and scene-transition hallucinations, and VideoHallucer (Wang et al., 2024) spans intrinsic and extrinsic categories, including object-relation, temporal, and semantic-detail hallucinations. EventHallusion (Zhang et al., 2024) targets event-level hallucination under misleading or compositional contexts, while TempCompass (Liu et al., 2024b) controls for static appearance by pairing videos that differ in a single temporal aspect. MotionBench (Hong et al., 2026) shifts the focus to fine-grained motion perception across six motion-oriented question types and finds that current models perform poorly on them. The authors attribute this weakness in part to frame-level feature compression and address it with Through-Encoder Fusion, which models temporal dependencies inside the visual encoder but requires training. Together, these benchmarks indicate that Video-LLM failures extend beyond object recognition to how visual content changes over time, which is the focus of our work.

2.2 Hallucination Mitigation Methods

Image-level mitigation. Early work on image LVLMs attributes hallucination to over-reliance on language priors (Gunjal et al., 2024; Liu et al., 2024a). Training-based methods improve grounding with robust instruction data (Liu et al., 2024a) or fine-grained human feedback (Yu et al., 2024), while training-free methods intervene at decoding time via visual contrastive decoding (Leng et al., 2023), over-trust penalties (Huang et al., 2024), attention calibration (Woo et al., 2025), or detector-based guidance (Zhao et al., 2025). Designed for static images, they do not model cross-frame



Figure 1: **Comparison of training-free video hallucination mitigation paradigms.** DINO-HEAL (a) reweights spatially salient visual features, TCD (b) contrasts predictions from original and temporally corrupted videos, and SEASON (c) combines temporal and spatial negatives through self-diagnostic contrastive decoding. In contrast, our method (d) directly strengthens the positive visual representation by combining foreground evidence, cross-frame persistence, and foreground-gated temporal differences. The resulting enhanced tokens V' provide stronger object-centric spatiotemporal evidence to the LLM without model fine-tuning. Heatmaps and logits are illustrative.

changes and remain insufficient for temporal hallucination in videos (Li et al., 2025a; Wang et al., 2024).

Training-free video methods. Several works extend inference-time mitigation to Video-LLMs (Qi et al., 2025). Most rely on contrastive decoding and differ in how the negative evidence is built: temporal downsampling in TCD (Zhang et al., 2024), forward–reversed motion elimination in MotionCD (Kong et al., 2025), model-guided object-level counterfactuals in MACD (Xiao & Zhou, 2026), temporally homogenized and spatially perturbed negatives with self-diagnostic weights in SEASON (Wu et al., 2025), and modality-aware masking for audio-visual inputs in AVCD (Jung et al., 2025). Other methods act on visual evidence or response selection: DINO-HEAL (Li et al., 2025a) reweights features by DINOv2 saliency (Oquab et al., 2024), MotionSight (Du et al., 2025) highlights motion through visual prompts, and SmartSight (Sun et al., 2025) stops decoding on hallucination-prone responses. These methods either adjust output distributions with altered inputs or emphasize evidence within individual frames. None examines whether object-centric motion is already present in the visual representations but unused at answer time, or makes this change explicit in the positive visual tokens themselves. Section 3 analyzes this gap, and Section 4 addresses it by injecting foreground-gated inter-frame residuals into the projected visual tokens, without object tracking or additional decoding branches.

Training-based video methods. Another line of work adapts model parameters. Vista-LLaMA (Ma et al., 2025) removes relative positional encoding between visual and text tokens so that visual influence does not decay with output length; MASH-VLM (Bae et al., 2025) disentangles spatial and temporal representations to reduce action–scene hallucination; and MMA (Dang et al., 2025) enforces multi-level video–text alignment through a semantic discriminative loss. Preference-based methods, including PaMi-VDPO (Ding et al., 2025), VistaDPO (Huang et al., 2025), TPO (Li et al., 2025b), and RRPO (Sarkar & Etemad, 2025), strengthen video–text alignment with curated preference data, while ArrowRL (Xue et al., 2025) and VideoHallu (Li et al., 2025c) apply reinforcement learning to improve temporal and physical reasoning. TAAE (Cai et al., 2025) instead learns temporal-aware activation interventions. These methods show that optimization can improve temporal grounding, but they require training data, reward signals, or model-specific adaptation. Our work examines a complementary setting: strengthening the visual evidence supplied to a frozen Video-LLM at inference time, without updating its parameters.

3 Problem Formulation and Analysis

3.1 Problem Formulation

We seek to reduce temporal prediction errors by modifying the visual evidence supplied to a frozen Video-LLM at inference time. Let $X = (x_1, \ldots, x_K)$ be sampled video frames, Q a question, and Y its ground-truth answer. The visual pipeline produces $V = f_\theta(X) \in \mathbb{R}^{T \times P \times D}$, where T denotes temporal steps after any temporal merging, P spatial tokens per step, and D feature dimensions. The decoder predicts an answer from $p_\phi(\cdot \mid V, Q)$.

Object-centric motion. Object-centric motion concerns changes in an object's position or state across temporal observations. We describe this evidence at the foreground level using pooled visual states. Let $F_{t,p} \in [0, 1]$ denote the foreground occupancy of token (t, p), obtained by averaging a pixel-level foreground mask over each token's pixel footprint and aligned with V after spatial pooling and temporal merging (Section 4). Define

$$o_t = \sum_{p=1}^{P} \frac{F_{t,p}}{\sum_{p'=1}^{P} F_{t,p'} + \epsilon} V_{t,p}, \qquad m_t = o_t - o_{t-1}, \quad t = 2, \ldots, T. \tag{1}$$

The ordered sequence $\mathcal{M}(X) = (m_2, \ldots, m_T)$ describes foreground-state changes. This is a coarse characterization: pooling may combine multiple objects and does not establish same-object correspondence. The method implements temporal enhancement through a separate token-wise residual in Section 4.

Objective. Let g_η be a visual-token transformation with coefficients $\eta = (\alpha, \alpha_s, \beta)$ that uses only V and F. With θ and ϕ frozen, we choose

$$\eta^\star = \arg\min_{\eta} \; \mathbb{E}_{(X,Q,Y)\sim\mathcal{D}_{\text{dev}}} \big[\text{err}(\hat{Y}_\eta, Y)\big], \tag{2}$$

where $\hat{Y}_\eta$ is the answer decoded with a fixed procedure from $p_\phi(\cdot \mid g_\eta(V, F), Q)$, err is the benchmark's binary error indicator, and $\mathcal{D}_{\text{dev}}$ is development data disjoint from the evaluation data; $\eta^\star$ is then held fixed.

We then investigate which factors make an answer temporally wrong. Following the path from frames to answer, we examine three candidate factors: (F1) the model does not attend to the moving foreground (Section 3.2); (F2) the model attends to the object but misses its state change between frames (Section 3.3); and (F3) motion information is present in the representations but not reflected in the answer (Section 3.4). Each analysis ends with a conclusion on its factor, and together they motivate the method in Section 4.

3.2 Attention to the Moving Foreground

We first examine whether temporal errors coincide with weak attention to moving foreground. On 600 VidHalluc TSH questions answered by LLaVA-OV-7B, we define, for each step $t \geq 2$, the moving foreground $\mathcal{A}_t$ as the 20% of tokens with $F_{t,p} > 0.1$ whose pixel footprints show the largest RGB change from step $t - 1$, and the background as $\mathcal{B}_t = \{p : F_{t,p} \leq 0.1\}$ (Appendix A.2). Let $a_t^{\mathcal{A}}$ and $a_t^{\mathcal{B}}$ denote mean question-to-visual attention per token over $\mathcal{A}_t$ and $\mathcal{B}_t$. The bad-step ratio is

$$\rho = \frac{1}{|\mathcal{T}_{\text{valid}}|} \sum_{t \in \mathcal{T}_{\text{valid}}} \mathbb{1}\left[a_t^{\mathcal{A}} < a_t^{\mathcal{B}}\right], \tag{3}$$

where $\mathcal{T}_{\text{valid}}$ contains steps with both regions defined; a higher ρ means that the moving foreground receives less attention than the background at more steps.

Correct predictions have lower ρ than incorrect ones (55.41% versus 59.77%; one-sided Mann-Whitney $p = 0.011$, $d = -0.19$). Samples with $\rho > 0.75$ ($n = 117$) have a 62.4% error rate, compared with 47.6% for the remaining 483 (Fisher $p = 0.005$; Table 1). The association is small and the threshold is post hoc (Appendix A.2).

Conclusion (F1). Weak attention to the moving foreground contributes to temporal errors, but only weakly, and does not explain them alone. This motivates emphasizing foreground tokens. Attending

Table 1: Error rate for samples with bad-step ratio $\rho > \tau$ (LLaVA-OV-7B, VidHalluc TSH, $N = 600$, overall error 50.5%). Gain is in percentage points over the overall error rate; p: uncorrected Fisher exact test comparing $\rho > \tau$ with $\rho \leq \tau$. Thresholds were examined post hoc; the full sweep is in Table 7.

τ	$n_{\rho>\tau}$	Wrong rate	Gain (pp)	p
0.86	44	70.5%	$+20.0$	0.007
0.75	117	62.4%	$+11.9$	0.005
0.57	253	55.7%	$+5.2$	0.032

to the object, however, does not guarantee that its change between frames is captured, which we examine next.

3.3 Missing Inter-Frame Object State

Recognizing an object in individual frames does not ensure that its state change informs the answer. We inspect last-layer answer-to-visual attention in Qwen2.5-VL-7B on TempCompass. In the illustrated correct cases, attention covers the relevant object across frames; incorrect cases instead concentrate on early frames or shift to unrelated regions, suggesting incomplete coverage of temporal evidence. Brief transitions may be missed when relevant frames receive little attention, while changes absent from the sampled input cannot be recovered (Section 6).

Quantitatively, LLaVA-OV-7B achieves 88.5-93.0% accuracy on non-temporal basic questions of VideoHallucer, compared with 54.0% on temporal ones. Qwen2.5-VL-7B exhibits temporal-blindness rates of 59.5% (P1) and 52.7% (P2) under paired reversal and option-order checks on TempCompass.

Conclusion (F2). Recognizing an object does not ensure that its temporal change informs the answer. The visualizations suggest incomplete coverage of object states, but do not establish whether motion information is absent from the representations. We therefore examine whether motion-direction information remains recoverable from the model's internal representations, even when its temporal predictions are incorrect.

3.4 Motion Information and Answer Generation

We use linear probes to examine motion direction, a measurable aspect of object-centric motion, at visual-token positions and at the state used to generate the answer.

Visual representations. For sample i, step t, and layer ℓ, we pool foreground hidden states as

$$R_{i,t}^{(\ell)} = \frac{\sum_{p \in \mathcal{O}_{i,t}} a_{i,t,p}^{(\ell)} H_{i,t,p}^{(\ell)}}{\sum_{p \in \mathcal{O}_{i,t}} a_{i,t,p}^{(\ell)} + \epsilon}, \tag{4}$$

where $\mathcal{O}_{i,t}$ is the BiRefNet foreground-token set, $H_{i,t,p}^{(\ell)}$ a hidden state, and $a_{i,t,p}^{(\ell)}$ its attention-derived pooling weight. Linear probes predict binary direction labels with per-step accuracy ranging from 0.780 at layer 0 to 0.893 at layer 28, indicating that direction information is recoverable from these representations.

Decision states. We also probe $Z_i = h_{i,n_i}^{(L)}$, the final-layer hidden state immediately preceding the scored answer token. The scored token is excluded from its context, and no ground-truth answer is supplied. A probe trained on split A correctly predicts direction on 63 of 101 held-out split-B videos (62.38%), compared with 55 (54.46%) for the model. The probe is correct while the model is wrong in 33 cases; the reverse occurs in 25 cases. These results establish decodability rather than causal use; the different evaluation units also prevent interpreting the visual and decision-state accuracy gap as information loss.

Conclusion (F3). Motion-direction information is present in the representations, including some states that produce incorrect answers, but is not consistently reflected in the answer. This motivates

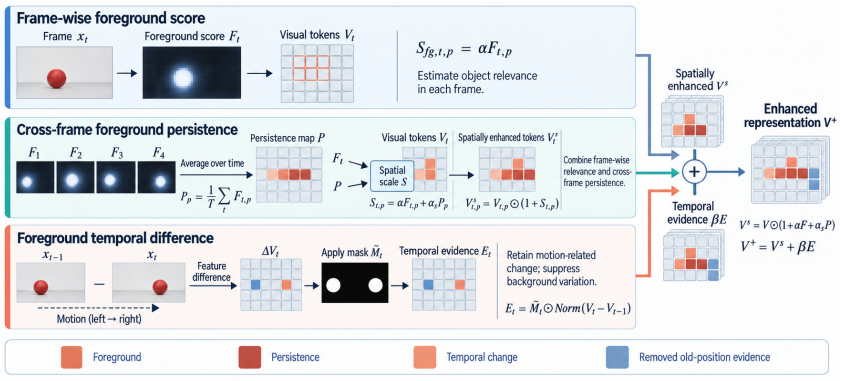

Figure 2: Overview of Feature Enhance. Sampled frames are processed by BiRefNet and the frozen visual pipeline to obtain foreground occupancy F and projected visual tokens V. The foreground map and its temporal average P scale the tokens to give V^{sp}. Adding the foreground-gated temporal residual gives $V' = V^{\mathrm{sp}} + \beta\hat{e}$, without parameter updates.
exposing inter-frame changes directly in the visual tokens that the language model reads, rather than modifying decoding; whether this helps is tested by the intervention experiments.

4 METHODOLOGY

Feature Enhance modifies the projected visual tokens V before they enter the frozen language model. It strengthens foreground tokens and adds an explicit signal of how they change between steps (Figure 2). All terms are computed from the sampled frames, without access to the question or updates to model parameters.

Foreground map. The main subjects of a video, such as people and objects, usually lie in the foreground, so we first locate the foreground in each frame. We apply BiRefNet (Zheng et al., 2024), an off-the-shelf segmentation model, to obtain a binary foreground mask for each sampled frame. We then average the mask over the pixels covered by each visual token, aligned with the backbone's spatial pooling and temporal merging. This gives a foreground map $F \in [0,1]^{T\times P}$, where $F_{t,p}$ is the fraction of token (t,p) covered by foreground: close to 1 inside an object and close to 0 in the background. The map does not depend on the question.

Spatial term. The spatial term strengthens foreground tokens, with the aim of increasing their influence relative to the background. The frame-wise map $F_{t,p}$ strengthens objects where they appear at the current step. Because a mask can miss an object at some steps, we also use the average foreground occupancy of each position over the clip, the persistence map

$$P_p = \frac{1}{T}\sum_{t=1}^{T} F_{t,p}. \tag{5}$$

P_p is high for positions that are foreground in many steps and is applied to every step, including steps where $F_{t,p}$ is low. It is a spatial prior only: it does not track objects or encode the order of their states. The two signals are combined into one gain per token:

$$S_{t,p} = \alpha F_{t,p} + \alpha_s P_p, \qquad V^{\mathrm{sp}}_{t,p} = V_{t,p} \odot \big(1 + S_{t,p}\big), \qquad \alpha, \alpha_s \geq 0. \tag{6}$$

Each token is multiplied by a single positive number, so its direction is unchanged and only its magnitude grows with its foreground evidence. Tokens with no foreground evidence ($F_{t,p} = P_p = 0$) are left unchanged, and α and α_s control how strongly the two signals are applied. Because the normalization layers of the language model may partly offset a pure rescaling, we measure the effect of this term empirically in Section 5.3.

Temporal term. This is the term that carries temporal information and the only one that is not a rescaling. It is built in three steps:

$$\underbrace{d_{t,p} = V_{t,p} - V_{t-1,p}}_{\text{(a) temporal difference}}, \qquad \underbrace{\tilde{d}_{t,p} = \frac{d_{t,p}}{\|d_{t,p}\|_2 + \epsilon}\,\|V_{t,p}\|_2}_{\text{(b) norm matching}}, \qquad \underbrace{\hat{e}_{t,p} = F_{t,p}\,\tilde{d}_{t,p}}_{\text{(c) foreground gating}}, \tag{7}$$

with $d_{1,p} = 0$ so the first step receives no injection. Temporal differencing exposes adjacent-step feature changes, norm matching scales them relative to the current token, and foreground gating weights their contribution by foreground occupancy.

Fusion.

$$V'_{t,p} = \underbrace{V_{t,p} \odot \big(1 + \alpha F_{t,p} + \alpha_s P_p\big)}_{\text{multiplicative: re-weights}} + \underbrace{\beta\, \hat{e}_{t,p}}_{\text{additive: injects}} \tag{8}$$

equation 8 combines foreground-based spatial scaling with a residual computed from adjacent sampled steps. The former emphasizes tokens at positions occupied by salient objects; the latter adds a signal of change at those positions. Both operate on projected visual tokens before they enter the language model and require no parameter updates. We assess their individual contributions in Section 5.3.

5 EXPERIMENTS

5.1 SETUP

Benchmarks and metrics. Our main evaluation uses VideoHallucer (Wang et al., 2024) and EventHallusion (Zhang et al., 2024). VideoHallucer counts a question pair as correct only when both the basic and hallucination questions are answered correctly. We additionally report MotionBench (Hong et al., 2026) on Qwen2.5-VL-7B to assess fine-grained motion perception. We report per-subtask accuracy and their macro-average as the benchmark-level score. Component ablations use a development subset described in Table 4.

Backbones. We evaluate Qwen2.5-VL-7B-Instruct (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2024a), and LLaVA-Video-7B (Lin et al., 2024), with all model weights frozen. Foreground maps are extracted using BiRefNet (Zheng et al., 2024). All methods share the same settings for each backbone: 8 uniformly sampled frames and greedy decoding. Feature Enhance is applied to the projected visual tokens after each backbone's spatial pooling and temporal merging, immediately before they enter the language model; special tokens (e.g., newline separators) are left unchanged.

Baselines. We evaluate TCD (Zhang et al., 2024), DINO-HEAL (Li et al., 2025a), and SEASON Wu et al. (2025) in a unified harness with identical sampled frames, prompts, decoding settings, and scoring procedures; hyperparameters are selected on the same disjoint development subsets. TCD and DINO-HEAL are adapted from publicly available codebases. As SEASON had no public implementation at the time of our experiments, we reimplement it from the published description and report its results as a best-effort reproduction. Implementation details are provided in Appendix C.

5.2 MAIN RESULTS

Hallucination Benchmark Evaluation.

Table 2 reports results on VideoHallucer and EventHallusion across three Video-LLM backbones. Without updating model parameters, our method improves the unmodified backbones on both benchmarks. On VideoHallucer, average accuracy rises from 40.49% to 45.16% on LLaVA-OV-7B, from 51.00% to 54.70% on Qwen2.5-VL-7B, and from 51.22% to 54.46% on LLaVA-Video-7B: gains of 4.67, 3.70, and 3.24 percentage points, respectively.

The largest category gains occur on SDH: $43.50\% \rightarrow 52.00\%$ for LLaVA-OV-7B and $63.00\% \rightarrow 71.50\%$ for Qwen2.5-VL-7B ($+8.50$ points each). LLaVA-OV-7B also gains 5.50 points on ORH. Gains on TPH are smaller, ranging from $+1.71$ to $+3.87$ points. On MotionBench (Table 3), Qwen2.5-VL gains 5.34 points in motion recognition and 2.3 points in camera motion.

On EventHallusion, average accuracy rises from 59.06% to 63.06% on LLaVA-OV-7B, from 65.77% to 68.71% on Qwen2.5-VL-7B, and from 65.16% to 67.50% on LLaVA-Video-7B, corresponding to gains of 4.00, 2.94, and 2.34 points. Our method improves ENTIRE on all three backbones and MISLEADING on LLaVA-OV and Qwen2.5-VL, while matching the base on LLaVA-Video-7B. SEASON performs better on MIX across all three backbones and achieves a higher EventHallusion average on Qwen2.5-VL-7B.

Table 2: Accuracy (%) on VideoHallucer and EventHallusion. AVG is the arithmetic mean of the five VideoHallucer tasks or the three EventHallusion subsets, respectively, calculated from the displayed task scores. Within each backbone, the highest score in each column is shown in bold and the second-highest distinct score is underlined; ties receive the same formatting.

Model	VideoHallucer						EventHallusion			
	ORH	TPH	SDH	EFH	ENFH	AVG	ENTIRE	MISLEADING	MIX	AVG
LLaVA-OV-7B Li et al. (2024a)	50.00	45.45	43.50	15.00	48.50	40.49	50.00	85.22	41.97	59.06
+TCD Zhang et al. (2024)	48.50	46.02	44.50	15.00	53.00	41.40	49.12	87.25	43.01	59.79
+DINO-HEAL Li et al. (2025a)	51.00	43.75	44.00	12.50	52.50	40.75	48.25	88.24	40.83	59.11
+SEASON Wu et al. (2025)	51.00	47.10	43.50	**17.50**	47.50	41.32	50.00	87.25	**48.55**	61.93
Our Method	**55.50**	**49.32**	**52.00**	15.00	**54.00**	**45.16**	**52.88**	**89.33**	46.97	**63.06**
Qwen2.5-VL-7B Bai et al. (2025)	64.50	46.02	63.00	24.50	57.00	51.00	64.04	89.22	44.04	65.77
+TCD Zhang et al. (2024)	64.00	46.59	67.00	25.00	55.00	51.52	60.53	85.29	40.41	62.08
+DINO-HEAL Li et al. (2025a)	65.50	**48.86**	69.00	24.50	57.00	52.97	64.04	89.22	44.04	65.77
+SEASON Wu et al. (2025)	66.50	47.75	64.00	22.50	57.00	51.55	65.40	91.18	**53.98**	**70.19**
Our Method	**67.00**	48.02	**71.50**	**28.50**	**58.50**	**54.70**	**66.79**	**92.20**	47.15	68.71
LLaVA-Video-7B Lin et al. (2024)	66.00	59.09	64.00	14.50	52.50	51.22	58.77	**92.16**	44.56	65.16
+TCD Zhang et al. (2024)	65.50	56.82	62.50	15.00	52.50	50.46	57.89	**92.16**	47.19	65.75
+DINO-HEAL Li et al. (2025a)	64.50	56.82	66.50	12.50	51.50	50.36	54.39	**92.16**	40.93	62.49
+SEASON Wu et al. (2025)	65.50	53.41	55.50	**18.50**	48.00	48.18	54.39	89.22	**56.99**	66.87
Our Method	**68.00**	**60.80**	**68.50**	17.50	**57.50**	**54.46**	**62.77**	**92.16**	47.58	**67.50**

Table 3: Accuracy (%) of Qwen2.5-VL-7B by motion-related question category on MotionBench. AVG is the unweighted mean of the six categories.

Method	**Action order**	**Camera motion**	**Location-related motion**	**Motion recognition**	**Motion-related objects**	**Repetition count**	**AVG**
Qwen2.5-VL-7B Bai et al. (2025)	43.74	52.73	59.34	62.72	78.84	28.25	54.27
+TCD Zhang et al. (2024)	43.35	52.73	58.79	61.37	**78.99**	28.50	53.96
+DINO-HEAL Li et al. (2025a)	43.74	52.73	59.34	62.72	78.84	28.25	54.27
+SEASON Wu et al. (2025)	46.61	45.19	55.86	54.26	70.72	29.00	50.27
Our Method	**45.55**	**55.03**	**59.52**	**68.06**	78.84	**30.00**	**56.17**

5.3 Component analysis

We examine the contributions of individual terms and their combinations in equation 8 on the development subset (Table 4).. Foreground weighting improves TSH by 3 percentage points, while persistence improves TPH by 4 points. The full configuration is best or tied-best across all four tasks, improving TSH, TPH, and EventHallusion over the baseline by 5, 4, and 3 points, respectively, and raising average accuracy from 61.00% to 64.25%. These gains favor the combined configuration, although differing coefficients and small subsets limit conclusions about individual component effects.

Table 4: Component ablation on the Qwen2.5-VL development subset (100 TSH, 100 MCQ, 100 TPH, and 100 EventHallusion samples). Results compare the contribution of each method term. Full results with coefficient values, and the grid search used to select $(\alpha, \alpha_s, \beta) = (0.4, 0.4, 0.8)$, are given in Appendix D

Active terms	TSH	MCQ	TPH	EventHallusion
Base	0.39	0.87	0.44	0.74
Foreground αF	0.42	0.85	0.44	0.72
Persistence $\alpha_s P$	0.40	0.85	0.48	0.74
Spatial $\alpha F + \alpha_s P$	0.39	0.85	0.46	0.72
Temporal $\beta \hat{e}$	0.43	0.85	0.48	0.75
Full (Eq. 8)	**0.44**	**0.88**	**0.48**	**0.77**

6 DISCUSSION

What underlies temporal hallucination? Foreground attention alone does not explain temporal errors: it is stronger for correct answers but does not reliably separate them from incorrect ones. Motion direction is linearly decodable from visual representations yet not consistently reflected in

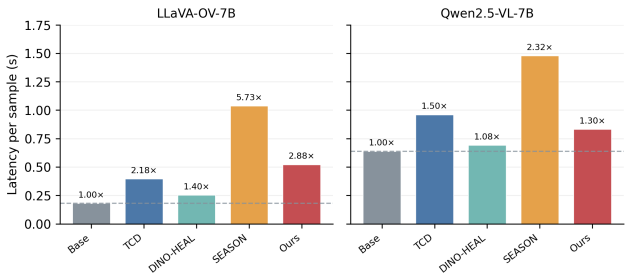

Figure 3: Inference latency on EventHallusion across two Video-LLM backbones. Each bar shows the macro-average time per sample over dataset; labels above the bars indicate the latency relative to the corresponding unmodified backbone. Dashed lines mark the backbone latency. Lower is better.

answers (Section 3.4). This points to a gap between the availability of *object-centric motion* and its use in answer generation, motivating our combination of foreground emphasis and inter-frame residuals.

Spatial and temporal evidence play different roles. The ablation results further show that no individual component consistently dominates across all tasks. Foreground and persistence terms primarily strengthen spatially grounded object evidence, whereas the temporal residual captures inter-frame changes. Their combination achieves the best macro performance, suggesting that reliable video reasoning requires both identifying *what* information is important and modeling *how* that information changes over time. At the same time, the small gap between different full configurations indicates that the method is not overly sensitive to a single coefficient setting.

Limitations. Feature Enhance targets object-centric temporal change and does not address all forms of temporal reasoning; in particular, differences between adjacent sampled steps indicate change but not duration, so duration-related questions are not directly supported. Because these differences compare fixed spatial positions, they do not establish object correspondence across frames, and under strong camera motion or egocentric video they may reflect viewpoint change rather than object motion. The temporal signal is further bounded by frame sampling: transitions that fall between sampled frames cannot be recovered. Finally, the method depends on BiRefNet foreground masks, which are not question-conditioned; segmentation errors, or salient objects unrelated to the query in multi-object scenes, can direct the enhancement toward the wrong regions.

7 CONCLUSION

We investigated why MLLMs often recognize relevant objects yet fail to understand their changes over time, and how these failures can be mitigated during inference. Our observations reveal temporally imbalanced and inconsistent object-focused attention in hallucinated cases. Complementary probing results suggest that motion-direction information remains accessible in visual representations but is not reliably reflected in the answer-decision state. Together, these findings motivate strengthening both object-relevant evidence and inter-frame changes.

We propose a training-free enhancement that combines foreground emphasis, a persistence prior, and inter-frame residuals to make such evidence explicit in visual tokens. This approach targets the gap between frame-level perception and temporal integration without updating model parameters.

The method nevertheless depends on foreground estimation and the temporal coverage of sampled frames. Fixed-position differences do not explicitly track moving objects, and enhancement cannot recover transitions missing from the sampled input. Furthermore, attention and probing analyses provide suggestive rather than conclusive causal evidence. Future work will explore motion-aligned correspondence, adaptive frame sampling, and causal interventions to better characterize and improve the use of temporal information.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. MASH-VLM: Mitigating action-scene hallucination in video-LLMs through disentangled spatial-temporal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL https://proceedings.mlr.press/v139/bertasius21a.html.

Jianfeng Cai, Wengang Zhou, Zongmeng Zhang, Jiale Hong, Nianji Zhan, and Houqiang Li. Mitigating hallucination in videollms via temporal-aware activation engineering, 2025. URL https://arxiv.org/abs/2505.12826.

Jisheng Dang, Shengjun Deng, Haochen Chang, Teng Wang, Bimei Wang, Shude Wang, Nannan Zhu, Guo Niu, Jingwen Zhao, and Jizhao Liu. Hallucination reduction in video-language models via hierarchical multimodal consistency. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9167–9175, 2025. doi: 10.24963/ijcai.2025/1019.

Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning, 2025. URL https://arxiv.org/abs/2504.05810.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. Motionsight: Boosting fine-grained motion understanding in multimodal llms, 2025. URL https://arxiv.org/abs/2506.01674.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143, 2024. doi: 10.1609/aaai.v38i16.29771.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models, 2026. URL https://arxiv.org/abs/2501.02955.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, 2023.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models, 2025. URL https://arxiv.org/abs/2504.13122.

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 (CVPR)*, pp. 13418–13427, 2024. doi: 10.1109/CVPR52733.2024.01274.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2121.

Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. MHBench: Demystifying motion hallucination in VideoLLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4401–4409, 2025. doi: 10.1609/aaai.v39i4.32463.

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. *arXiv preprint arXiv:2311.16922*, 2023. URL https://arxiv.org/abs/2311.16922.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding, 2025a. URL https://arxiv.org/abs/2412.03735.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024b. URL https://arxiv.org/abs/2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024c.

Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding, 2025b. URL https://arxiv.org/abs/2501.13919.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025c.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, 2024a.

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?, 2024b. URL https://arxiv.org/abs/2403.00476.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. 2025. URL https://arxiv.org/abs/2312.08870.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.679.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Daiqing Qi, Dongliang Guo, Hanzhang Yuan, Handong Zhao, Mengxuan Hu, Lehan Yang, and Sheng Li. Improve temporal reasoning in multimodal large language models via video contrastive decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2nIAtsUC27.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024.

Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1420.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Computer Vision – ECCV 2024*, pp. 256–274. Springer, 2024. doi: 10.1007/978-3-031-72943-0_15.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024.

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. Smartsight: Mitigating hallucination in video-llms without compromising video understanding via temporal attention collapse, 2025. URL https://arxiv.org/abs/2512.18671.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arxiv*, 2024.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025. doi: 10.18653/v1/2025.findings-acl.99.

Chang-Hsun Wu, Kai-Po Chang, Yu-Yang Sheng, Hung-Kai Chung, Kuei-Chun Wang, and Yu-Chiang Frank Wang. Season: Mitigating temporal hallucination in video large language models via self-diagnostic contrastive decoding, 2025. URL https://arxiv.org/abs/2512.04643.

Qixin Xiao and Kun Zhou. MACD: Model-aware contrastive decoding via counterfactual data for video-LLMs. In *International Conference on Machine Learning (ICML)*, 2026.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models, 2025. URL https://arxiv.org/abs/2506.03340.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10714–10726, 2023.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4651–4659, 2016.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, 2024.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://arxiv.org/abs/1906.02467.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77461–77486. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhao25j.html.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, pp. 9150038, December 2024. ISSN 2097-3691. doi: 10.26599/air.2024.9150038. URL http://dx.doi.org/10.26599/AIR.2024.9150038.

A Diagnostic Analyses

A.1 Overview

We provide evidence that (i) lower attention to the moving foreground is associated with more errors on temporal-order questions, and (ii) Video-LLMs recognize entities yet often answer identically when a video is reversed. *Unit* in Table 5 is the independent unit of each analysis.

TempCompass is used in two sets. **P1** holds the 126 multiple-choice direction *video files*; a clip and its released reverse (`_reverse`) count as two files. **P2** adds caption-matching questions and groups each clip with its reverse into one source video (43 sources, 317 questions).

Table 5: Diagnostic analyses.

Analysis	Backbone / data	Size	Unit
Foreground attention (A.2)	LLaVA-OV; VidHalluc TSH	600 questions	question
Recognition (A.3)	LLaVA-OV; VideoHallucer	176–200 pairs/cat.	pair
Temporal blindness (A.3)	Qwen2.5-VL; P1, P2	126 files, 317 q.	file, question

A.2 Diagnosis I: Foreground Attention

Protocol. LLaVA-OV-7B answers the 600 VidHalluc TSH questions (8 frames, $P = 196$ tokens per step; 49.5% correct). We average last-layer attention from question to visual tokens over heads and question tokens. Tokens with BiRefNet occupancy > 0.1 form the foreground; the rest form the background $\mathcal{B}_t$. For $t \geq 2$, the moving foreground $\mathcal{A}_t$ is the 20% of foreground tokens with the largest RGB change from $t-1$. The bad-step ratio ρ in equation 3 is the fraction of steps where mean attention on $\mathcal{A}_t$ is below that on $\mathcal{B}_t$; higher is worse.

Results. Correct predictions have a lower bad-step ratio (55.41% vs. 59.77%) and higher action enrichment (1.085 vs. 1.025); five of six metrics differ with $p < 0.05$, and two survive Bonferroni correction (Table 6). Effects are small ($|d| \leq 0.27$). Samples with high ρ are more error-prone (Table 7).

Table 6: Correct ($n = 297$) vs. incorrect ($n = 303$) predictions, LLaVA-OV-7B, TSH (mean $\pm$ std). Δ = Correct $-$ Wrong; d: standardized difference; p: one-sided Mann–Whitney (Bonferroni level ≈ 0.0083). Enrichment = attention share of $\mathcal{A}_t$ / token share of $\mathcal{A}_t$.

Metric	Wrong	Correct	Δ	d	p
Bad-step ratio $\rho \downarrow$	$59.77 \pm 24.07\%$	$55.41 \pm 21.73\%$	-4.36 pp	-0.19	0.011
Action enrichment $\uparrow$	1.025 ± 0.275	1.085 ± 0.288	$+0.060$	$+0.21$	0.006
Mean action attention ($\times 10^{-4}$) $\uparrow$	1.59 ± 0.50	1.68 ± 0.54	$+0.09$	$+0.17$	0.025
Log action/background attention $\uparrow$	-0.138 ± 0.319	-0.055 ± 0.283	$+0.083$	$+0.27$	0.002
Mean background attention ($\times 10^{-4}$) $\downarrow$	1.59 ± 0.23	1.55 ± 0.30	-0.04	-0.15	0.009
Action attention fraction $\uparrow$	0.083 ± 0.035	0.086 ± 0.036	$+0.002$	$+0.07$	0.180

A.3 Recognition versus Temporal Consistency

LLaVA-OV-7B answers non-temporal *basic* questions of VideoHallucer with 88.5–93.0% accuracy but only 54.0% of basic temporal questions, i.e., it recognizes entities at the video level but not their order. A TempCompass pair is *temporally blind* if a video and its frame-reversed copy receive the same option under both option orders, although the correct answer changes. Qwen2.5-VL-7B is temporally blind in 59.5% (P1) and 52.7% (P2) of cases (Table 8).

A.4 Summary

Table 9 summarizes the findings supported by these diagnostics and their scope.

Table 7: Error enrichment by bad-step ratio on the same 600 TSH samples (overall error 50.5%). High: $\rho > \tau$; Low: $\rho \leq \tau$ ($n_{\text{high}} + n_{\text{low}} = 600$). Gains in pp; Δ: error-rate difference High − Low; p: Fisher exact test, uncorrected (Bonferroni level ≈ 0.0036). Thresholds were examined post hoc.

	High ($\rho > \tau$)			Low ($\rho \leq \tau$)				
τ	n	Wrong rate	Gain	n	Correct rate	Gain	Δ (pp)	p
0.857	44	70.45	+19.95	556	51.08	+1.58	+21.53	0.007
0.833	96	63.54	+13.04	504	51.98	+2.48	+15.53	0.005
0.800	112	62.50	+12.00	488	52.25	+2.75	+14.75	0.006
0.750	117	62.39	+11.89	483	52.38	+2.88	+14.77	0.005
0.714	119	62.18	+11.68	481	52.39	+2.89	+14.58	0.006
0.667	218	59.17	+8.67	382	54.45	+4.95	+13.62	0.002
0.600	246	55.69	+5.19	354	53.11	+3.61	+8.80	0.038
0.571	253	55.73	+5.23	347	53.31	+3.81	+9.05	0.032
0.500	363	52.34	+1.84	237	52.32	+2.82	+4.66	0.278
0.429	400	50.75	+0.25	200	50.00	+0.50	+0.75	0.863
0.400	473	51.59	+1.09	127	53.54	+4.04	+5.13	0.319
0.333	483	51.55	+1.05	117	53.85	+4.35	+5.40	0.305
0.286	504	50.79	+0.29	96	51.04	+1.54	+1.84	0.824
0.250	556	50.54	+0.04	44	50.00	+0.50	+0.54	1.000

Table 8: Qwen2.5-VL-7B on TempCompass direction with temporal reversal (%). Correct: same correct option under both option orders; consistent: same option under both orders. Chance for four options: 25%.

		Correct		Consistent		
Set	n	Orig.	Rev.	Orig.	Rev.	Temporally blind
P1	126	34.1	34.1	76.2	75.4	59.5
P2	317	33.8	33.8	67.8	69.1	52.7

B Design Rationale and Ablations

Table 10 relates each term of equation 8 to the observations of Section 3 that motivated its design; these relations are motivational rather than causal, and the empirical contribution of each term is measured in Section 5.3.

Why use a first-order feature difference? The intervention modifies the projected visual representations consumed by the language model. Computing $d_{t,p}$ directly in this space requires no learned mapping from an external motion representation. The difference is measured at the same spatial position in adjacent sampled steps; it provides a proxy for inter-frame change rather than explicit object correspondence. The fused representation in equation 8 introduces no trainable parameters.

Why apply foreground gating? With uniform sampling, adjacent sampled steps may be separated by a substantial interval. The raw difference $d_{t,p}$ can therefore include changes in the background, viewpoint, and camera position as well as changes involving foreground objects. Multiplication by $F_{t,p}$ emphasizes differences at positions occupied by the estimated foreground. This gate can reduce background contributions, but it cannot remove camera-induced changes within foreground regions.

Why match residual norms? The magnitude of $d_{t,p}$ varies across tokens and backbones. Norm matching in equation 7 keeps the direction of $d_{t,p}$ and rescales it to $\|V_{t,p}\|_2$ before foreground gating, so that $\|\hat{e}_{t,p}\|_2 \approx F_{t,p}\|V_{t,p}\|_2$ and β has a consistent scale relative to each token. Because magnitude is discarded, small and large changes receive the same weight; foreground gating, not the size of the change, controls the residual strength. This does not guarantee identical behavior across backbones.

Why does the persistence map have no temporal index? As defined in equation 5, P_p averages foreground occupancy at spatial position p over sampled steps and is broadcast to all steps. It

Table 9: Findings supported by the diagnostics.

Finding	Evidence	Scope
Lower attention to the moving foreground accompanies more errors on temporal-order questions.	Bad-step ratio 55.41% vs. 59.77% ($p = 0.011$); enrichment 1.085 vs. 1.025 ($p = 0.006$); error 62.4% vs. 47.6% at $\rho > 0.75$.	LLaVA-OV, TSH; small effect.
Models recognize entities.	88.5–93.0% on non-temporal basic questions.	LLaVA-OV; video-level.
Models often ignore temporal order.	Temporally blind in 59.5% (P1) / 52.7% (P2); 54.0% on basic temporal questions.	Qwen2.5-VL; LLaVA-OV.

Table 10: Design rationale. The terms of equation 8 are linked to the diagnostic observations that motivate their inclusion.

Term	Motivating observation	Intended effect
αF (foreground gain)	Correct predictions exhibit stronger foreground evidence, although this association does not establish causality (Section 3.2).	Emphasize foreground regions
$\alpha_s P$ (persistence)	Frame-wise foreground evidence can vary across sampled steps (Section 3.3).	Stabilize spatial foreground emphasis
$\beta \hat{e}$ (temporal residual)	Direction information is decodable from visual representations but is not reliably reflected in model answers (Section 3.4).	Expose inter-frame feature changes

provides a clip-level spatial prior rather than encoding when a change occurs or tracking the same object across frames. Because it uses fixed spatial positions, its interpretation is less reliable under substantial camera motion (Section 6).

C Experimental Setup and Reproducibility

Video-LLM Backbones. We conduct experiments on three representative open-source Video-LLMs: LLaVA-OneVision-7B, LLaVA-Video-7B, and Qwen2.5-VL-7B-Instruct. The model checkpoints are kept fixed across all experiments, and no backbone parameters are updated during inference. Our intervention is applied exclusively to the intermediate visual representations before they are consumed by the language model. Consequently, Feature Enhance introduces no additional fine-tuning, gradient-based optimization, or task-specific training.

Unified Frame Sampling and Inference Protocol. To isolate the effect of inference-time feature intervention, we adopt a unified evaluation protocol across all four benchmarks. For every video, we uniformly sample exactly $T = 8$ **frames** over its temporal extent. The resulting frame indices are fixed before model inference and shared across the Base model and all mitigation methods. Thus, every method starts from the same canonical 8-frame representation of each video. Method-specific transformations required by a particular decoding strategy are constructed only from this shared frame bank and do not trigger independent resampling of the source video.

We use deterministic greedy decoding throughout the evaluation, with `do_sample=False` and `num_beams=1`. Where required by the model interface, the sampling temperature is set to zero. Within each benchmark, the prompt template, answer format, and decoding configuration are held fixed across all compared methods. This design controls for variations in frame selection and decoding settings, enabling a more direct comparison of inference-time interventions.

Table 11 summarizes the common evaluation configuration used throughout our experiments.

Comparison Methods. We compare Feature Enhance against the unmodified frozen Video-LLM Base and three inference-time hallucination mitigation methods: TCD (Zhang et al., 2024), +DINO-

Table 11: Unified inference protocol used across all four benchmarks. All compared methods start from the same canonical frame bank.

Configuration	Setting
Frames per video	8
Frame sampling	Uniform temporal sampling
Frame indices	Fixed and shared across methods
Decoding	Greedy
`do_sample`	False
`num_beams`	1
Parameter update	None
Additional training	None

HEAL Li et al. (2025a), and SEASON Wu et al. (2025). All comparison methods are executed within the same evaluation harness and use the same backbone checkpoint, canonical 8-frame input, benchmark-specific prompt, and decoding configuration. Method-specific operations, such as the construction of contrastive branches, are restricted to the prescribed input frame bank. No method is allowed to independently resample the original video.

Implementation of Feature Enhance. Our method intervenes at the projected visual-token level while keeping both the visual encoder and language model frozen. Let $V \in \mathbb{R}^{T\times P\times D}$ denote the projected visual representation, where T is the number of sampled frames, P is the number of visual tokens per frame, and D is the embedding dimension. In all experiments, $T = 8$.

For each frame, we obtain a foreground relevance map and align it with the corresponding visual-token grid. The intervention combines three complementary signals: (i) frame-wise foreground relevance, (ii) cross-frame foreground persistence, and (iii) foreground-gated temporal residuals. For a token at temporal index t and spatial index p, the enhanced representation is

$$V'_{t,p} = V_{t,p} \odot (1 + \alpha F_{t,p} + \alpha_s P_p) + \beta \hat{e}_{t,p}, \tag{9}$$

where $F_{t,p}$ denotes frame-level foreground relevance, P_p denotes foreground persistence across the sampled frames, and $\hat{e}_{t,p}$ is the normalized foreground-gated temporal residual. The coefficients α, α_s, and β control the contributions of the three enhancement terms. These coefficients are fixed during evaluation and introduce no trainable parameters.

Evaluation Metrics. We follow the task-specific evaluation protocol of each benchmark. For VidHalluc, we report the official task scores covering action, temporal-sequence, and scene-transition hallucinations. For VideoHallucer, we adopt the benchmark's strict paired evaluation: a question pair is counted as correct only when the model correctly answers both the basic question and its corresponding hallucination question. For EventHallusion, we report accuracy according to the benchmark's official evaluation split and answer-parsing protocol. For MotionBench, we use multiple-choice accuracy as the primary metric and additionally report category-wise accuracy over AO, CM, LM, MR, MO, and RC.

Reproducibility and Experiment Logging. We use a deterministic evaluation pipeline and explicitly record the configuration associated with every prediction. For each experiment, the evaluation harness stores the model checkpoint, benchmark configuration, sampled frame indices, frame-sampling policy, method-specific parameters, generation settings, raw model responses, parsed predictions, parser status, and execution logs. Runtime statistics and peak GPU memory usage are additionally recorded when available.

Frame selection is performed before method-specific inference, and the resulting frame manifest is reused by all compared methods. This prevents unintended differences in visual evidence caused by independent video decoding or frame resampling. All applicable random number generators are initialized with fixed seeds; however, the primary generation procedure itself is deterministic because stochastic sampling is disabled.

Table 12: Per-sample inference latency (seconds/sample) on EventHallusion under the unified 8-frame evaluation protocol. Macro Avg. assigns equal weight to all subset, whereas Weighted Avg. accounts for the number of evaluated samples in each subset. Lower is better.

Model	Method	Entire	Misleading	Mix	Macro Avg.	Weighted Avg.
LLaVA-OV-7B	Base	0.224	0.157	0.163	0.181	0.180
	+TCD	0.407	0.367	0.412	0.395	0.398
	+DINO-HEAL	0.255	0.247	0.259	0.254	0.254
	+SEASON	0.988	1.081	1.043	1.037	1.037
	Feature Enhance	0.520	0.514	0.533	0.522	0.524
Qwen2.5-VL-7B	Base	0.635	0.633	0.650	0.639	0.641
	+TCD	0.960	0.946	0.977	0.961	0.963
	+DINO-HEAL	0.683	0.683	0.710	0.692	0.694
	+SEASON	1.566	1.097	1.776	1.480	1.461
	Feature Enhance	0.811	0.848	0.840	0.833	0.833
LLaVA-Video-7B	Base	0.297	0.267	0.307	0.290	0.293
	+TCD	0.418	0.454	0.449	0.440	0.441
	+DINO-HEAL	0.410	0.363	0.373	0.382	0.382
	+SEASON	1.256	1.296	1.325	1.292	1.296
	Feature Enhance	1.011	0.985	0.984	0.993	0.993

The pipeline further supports resumable evaluation. A previously processed sample is reused only when its inference completed without an execution error and its prediction can be parsed successfully. Failed or incomplete samples remain explicitly recorded and are not silently treated as valid predictions. The accompanying implementation provides the benchmark configurations, frame-sampling rules, method configurations, inference scripts, and evaluation code required to reproduce the reported experiments.

Inference Efficiency and Latency Measurement. We additionally evaluate the computational overhead introduced by different inference-time interventions. All latency measurements are conducted on a single **NVIDIA H200 GPU** under the unified 8-frame evaluation protocol. All methods use the same backbone checkpoint, batch size of one video, and deterministic generation configuration. All methods are evaluated under the same unified 8-frame protocol, with batch size one at the video level, identical backbone checkpoints, and the same deterministic decoding configuration.

Latency is measured over the method-specific inference procedure after the canonical video frames have been sampled. Therefore, the reported values characterize the computational cost of model inference and the corresponding mitigation intervention, rather than video decoding or disk I/O. For EventHallusion, we report latency separately on the *Entire*, *Misleading*, and *Mix* subsets. We additionally report both the macro average across the three subsets and the sample-weighted average over the complete evaluation set.

Specifically, the macro-average latency is defined as

$$L_{\text{macro}} = \frac{L_{\text{Entire}} + L_{\text{Misleading}} + L_{\text{Mix}}}{3}, \tag{10}$$

whereas the sample-weighted average is

$$L_{\text{weighted}} = \frac{N_E L_E + N_M L_M + N_X L_X}{N_E + N_M + N_X}, \tag{11}$$

where N_E, N_M, and N_X denote the numbers of evaluated samples in the Entire, Misleading, and Mix subsets, respectively.

As shown in Table 12, all inference-time mitigation methods introduce additional computational cost relative to the unmodified backbone. However, Feature Enhance consistently requires less inference time than SEASON across all three backbones. Using the sample-weighted average, Feature Enhance reduces latency from $1.037\,\text{s}$ to $0.524\,\text{s}$ on LLaVA-OV-7B, from $1.461\,\text{s}$ to $0.833\,\text{s}$ on Qwen2.5-VL-7B, and from $1.296\,\text{s}$ to $0.993\,\text{s}$ on LLaVA-Video-7B. These correspond to latency reductions of approximately 49.5%, 43.0%, and 23.4%, respectively, relative to SEASON.

The computational overhead of Feature Enhance is nevertheless backbone-dependent. Relative to the corresponding Base model, its weighted latency is approximately $2.91\times$ on LLaVA-OV-7B, $1.30\times$ on Qwen2.5-VL-7B, and $3.39\times$ on LLaVA-Video-7B. This behavior is expected because our approach introduces an auxiliary foreground-extraction stage and visual-feature enhancement before language generation. Importantly, this additional computation does not require maintaining multiple contrastive decoding branches throughout autoregressive generation.

DINO-HEAL remains the lowest-latency mitigation method in this evaluation, indicating that Feature Enhance should not be interpreted as minimizing inference cost in isolation. Rather, the results demonstrate that the proposed positive spatiotemporal enhancement provides a substantially lighter inference profile than the more computationally intensive self-diagnostic contrastive decoding strategy while retaining a training-free intervention paradigm.

D Parameter Selection

Backbones. We evaluate Feature Enhance on LLaVA-OneVision-7B, LLaVA-Video-7B, and Qwen2.5-VL-7B-Instruct. All backbone parameters remain frozen. The intervention modifies projected visual tokens after spatial pooling and temporal merging, immediately before they enter the language model. Special tokens remain unchanged.

Foreground extraction. BiRefNet processes RGB frames resized to 1024×1024, normalized with ImageNet statistics, in batches of up to eight. Its output probabilities are converted into binary foreground masks, which are aligned with the backbone's visual preprocessing. We average these masks over the corresponding spatial token footprints and merged temporal steps to obtain $F_{t,p} \in [0, 1]$, the foreground occupancy of each visual token..

Parameter search. The coefficients α, α_s, and β control foreground weighting, persistence weighting, and the foreground-gated temporal residual, respectively. We conduct a structured grid search over $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8\}$, $\alpha_s \in \{0, 0.1, 0.2, 0.4, 0.8\}$, and $\beta \in \{0, 0.1, 0.2, 0.4, 0.6, 0.8\}$. The search evaluates 109 configurations drawn from this grid, covering individual components and their spatial and spatiotemporal combinations. Configurations are compared on development data. The full configuration reported below uses $(\alpha, \alpha_s, \beta) = (0.4, 0.4, 0.8)$. Coefficients remain fixed across samples, without per-video or per-question adaptation.

Component settings. Foreground-only enhancement uses $(0.4, 0, 0)$, persistence-only enhancement uses $(0, 0.1, 0)$, and temporal-only enhancement uses $(0, 0, 0.6)$. Persistence is computed as the temporal mean of foreground confidence at each spatial position, providing a spatial prior without tracking object identities. The temporal residual is computed from the original visual tokens, norm-matched, and then foreground-gated. Its first temporal step is zero. Since active coefficient values differ across configurations, the comparisons reflect both component inclusion and coefficient choice, rather than component removal at fixed coefficients.

Development results. Table 13 reports coefficient values and the corresponding results from Table 4. The Qwen2.5-VL development subset contains 100 samples each for TSH, MCQ, TPH, and EventHallusion. AVG is the arithmetic mean across the four tasks:

$$S_{\text{avg}} = \frac{S_{\text{TSH}} + S_{\text{MCQ}} + S_{\text{TPH}} + S_{\text{EventHallusion}}}{4}. \tag{12}$$

Because task sizes are equal, this also equals their sample-weighted accuracy.

Analysis. Individual configurations have task-dependent effects. Foreground-only enhancement improves TSH, while persistence-only enhancement improves TPH. Temporal-only enhancement improves TSH, TPH, and EventHallusion, but reduces MCQ accuracy. The full configuration achieves the highest or tied-highest accuracy on every task, averaging 64.25%. This exceeds the baseline by 3.25 percentage points and temporal-only enhancement by 1.50 points. These results favor the full configuration among those reported, although the small development subsets limit conclusions about the stability of the differences.

Table 13: Component configurations and accuracy on the Qwen2.5-VL development subset, with 100 samples per task. Scores are percentages; AVG is the arithmetic mean across tasks. Best results, including ties, are bold. Active coefficient values differ across configurations.

Configuration	α	α_s	β	TSH	MCQ	TPH	EventHallusion	AVG
Base	0.0	0.0	0.0	39	87	44	74	61.00
Foreground-only	0.4	0.0	0.0	42	85	44	72	60.75
Persistence-only	0.0	0.1	0.0	40	85	**48**	74	61.75
Temporal-only	0.0	0.0	0.6	43	85	**48**	75	62.75
Full	0.4	0.4	0.8	**44**	**88**	**48**	**77**	**64.25**

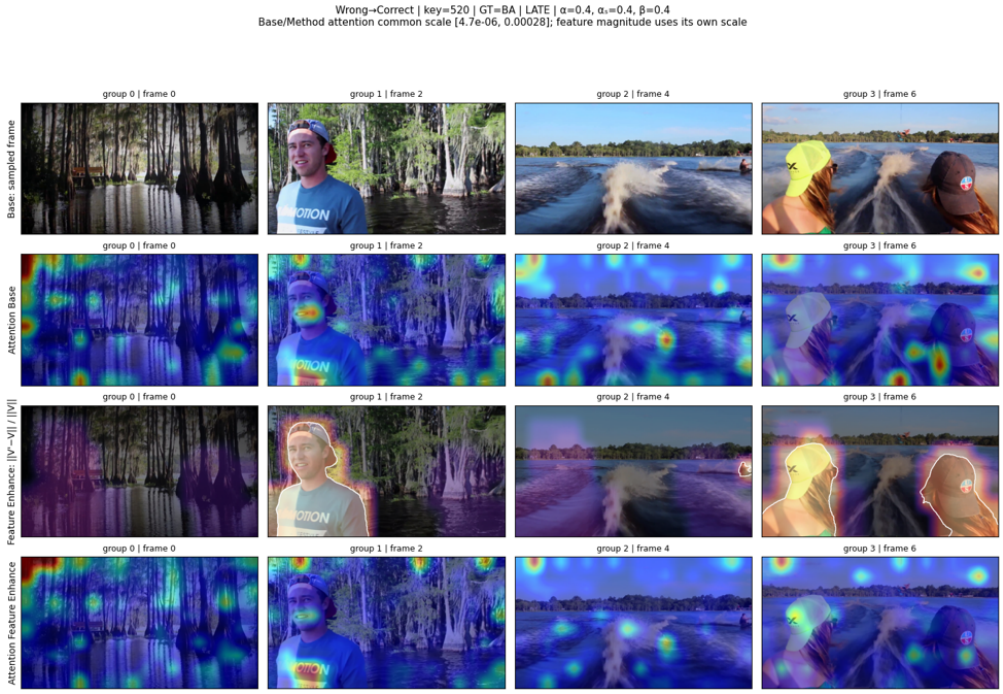

Figure 4: Case study 1: sampled frames, visual attention, and feature enhancement.

E Detailed Benchmark Results

Table 14: Ablation of positive enhancement and temporal-negative decoding on the evaluated VidHalluc subset using Qwen2.5-VL-7B. All configurations follow the same unified 8-frame inference protocol. Results are accuracy (%). Best results are shown in **bold**; ties are highlighted for all corresponding configurations.

Configuration	**TSH**	**ACH-B**	**ACH-MCQ**	**STH**
Base	64.83	63.99	77.99	48.09
Feature Enhance	65.50	75.41	78.86	**53.43**
Positive + temporal negative	**70.67**	**78.99**	**79.41**	**53.43**

F Case Studies and Paired Outcome Analysis

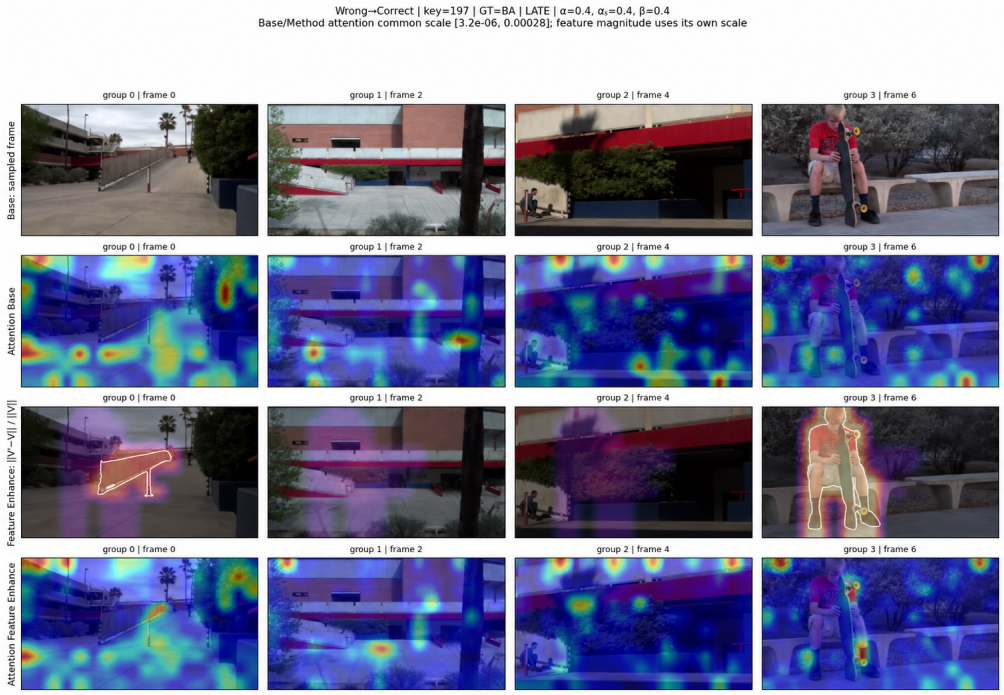

Figure 5: Case study 2: sampled frames, visual attention, and feature enhancement.



Figure 6: Case study 3: sampled frames, visual attention, and feature enhancement.

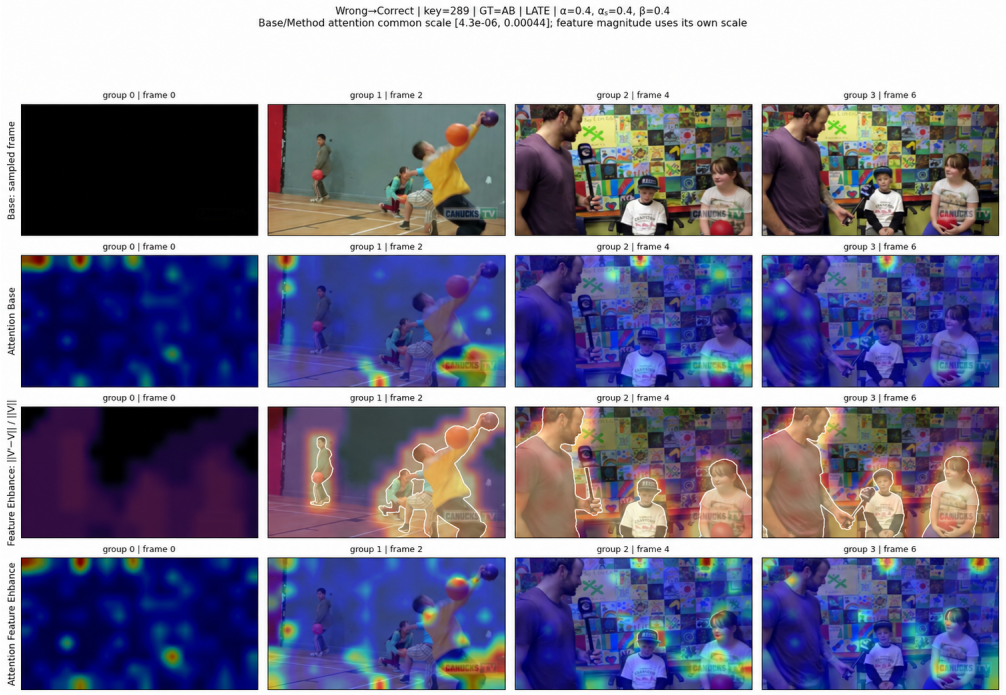

Figure 7: Case study 4: sampled frames, visual attention, and feature enhancement.

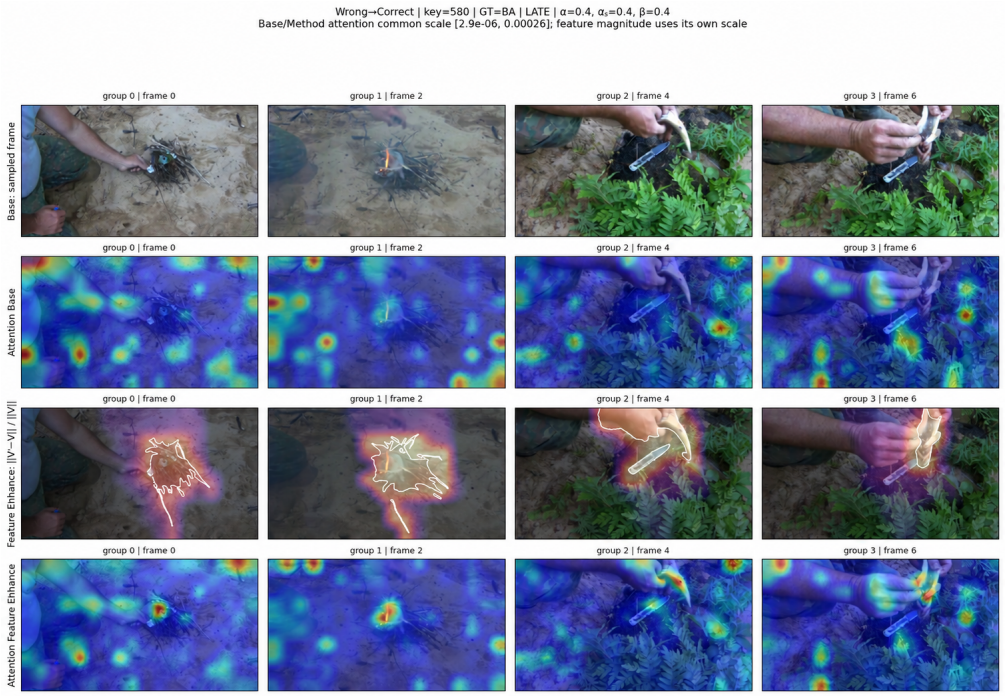

Figure 8: Case study 5: sampled frames, visual attention, and feature enhancement.