TEST-TIME PAIRED OPTION-WISE VERIFICATION TO REDUCE PRIOR-FOLLOWING ERRORS IN VIDEOLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (VideoLLMs) lose accuracy when only a few frames are available. We found that this degradation stems not only from missing visual evidence but also from prior-following errors: driven by its language prior, the model gives the same wrong answer with or without the video. Replacing the original video with an unrelated one often preserves the preferred option, indicating weak dependence of the decision on the question-specific video. However, reversing the temporal order changes the relative support for the correct and no-video preferred options, showing that sampled frames retain discriminative evidence. This mismatch between retained video evidence and prior-aligned decisions motivates Paired Option-Wise Verification (POV), a training-free correction method that independently verifies each answer option with and without the video. To remove the language prior, POV subtracts a scaled no-video score from each option's video-conditioned score. The scale is estimated from unlabeled target-task data and is reduced only when the video evidence is consistent with the prior, that is, when the verifier supports the prior-preferred option. Across several frozen backbones, POV outperforms competing training-free baselines on temporal multiple-choice and binary QA, and even surpasses dense-frame VideoLLMs while using the sparse-frame budget.

1 INTRODUCTION

Video large language models (VideoLLMs) process sampled video frames as visual tokens that condition a language model (Lin et al., 2024; Wang et al., 2024a; Li et al., 2024a). Using more frames improves temporal coverage, but also increases the visual-token budget and inference cost. Sparse-frame inference is therefore attractive for efficiency, although performance often drops on temporally demanding tasks (Nie et al., 2024; Lee et al., 2026). Existing sparse-frame methods primarily improve the visual information available to the model through query-aware frame selection, visual-token reduction, or temporal prompting (Nie et al., 2024; Chen et al., 2024; Hu et al., 2025; Tang et al., 2025; Huang et al., 2025; Lee et al., 2026). A complementary line of work has shown that multimodal predictions can also reflect question-only or language-side preferences (Goyal et al., 2017; Agrawal et al., 2018; Loginova et al., 2025). Motivated by these observations, we treat the no-video prediction of the same frozen model as an observable reference for its language-side preference and examine whether errors aligned with this preference retain discriminative temporal evidence.

We characterize a recurring error pattern in sparse-frame VideoLLMs in which the model makes an incorrect prediction that matches its prediction without the video. We refer to these cases as prior-following errors (PFEs) and examine two complementary properties of this failure mode. Replacing the original frames with an unrelated video tests whether the model's option preference depends on visual content. For PFEs, the preferred option changes substantially less than for correct predictions. However, reversing only the temporal order of the same sparse frames tests whether the retained visual content still carries discriminative temporal information. This intervention substantially changes the relative support for the correct and no-video preferred options. Thus, weak dependence of the option preference on the video does not imply an absence of informative temporal evidence. Layer-wise probing further shows that this distinction remains visible in intermediate decoder layers but weakens toward the final prediction.

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

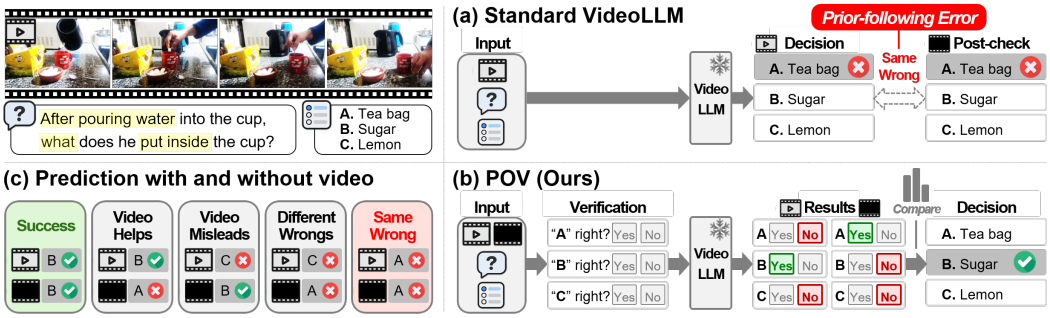


Figure 1: **Prior-following errors (PFEs) motivate option-wise verification.** (a) A standard VideoLLM selects an answer directly from the candidate options. (b) POV reformulates each option as a YES/NO prompt and compares its support under the video-conditioned and no-video conditions, enabling correction of this failure mode. (c) Predictions under the video-conditioned and no-video conditions exhibit five distinct behaviors; our method targets the prior-following case, where both are wrong and select the same option.

These results reveal a mismatch between retained video evidence and a final prediction aligned with the model's no-video preference. The mismatch motivates separating the model's response to the video from its preference without the video. We therefore introduce Paired Option-Wise Verification (POV), which evaluates each option from two complementary views—with and without the video—on a shared YES/NO score scale. As illustrated in Figure 1, this reformulation exposes the video-conditioned and no-video support for each option separately rather than collapsing them into a single multiple-choice decision. The no-video scores provide an option-wise measure of the model's prior support. POV then subtracts the component aligned with this prior through a one-sided projection, with selective use of the signed coefficient when the prior-preferred option remains supported. All coefficients are estimated from unlabeled target-task scores.

Contributions.

- **Diagnosis of prior-following errors with retained temporal evidence (Section 3).** Through controlled video replacement, temporal-order reversal, and layer-wise probing, we show that errors whose final prediction matches the no-video prediction can retain discriminative temporal information, with the correct and prior-preferred options remaining distinguishable in intermediate decoder layers.
- **Option-wise verification of video response and prior preference (Section 4.1).** We reformulate multiple-choice question answering (MCQA) as independent YES/NO verification of each option, placing the video-conditioned and prior scores in a shared score space without joint normalization over the candidate set.
- **Label-free prior correction from paired verification scores (Sections 4.2–4.3).** We estimate a correction scale from unlabeled examples within each cell while retaining each question's prior-score vector. We introduce a one-sided projection coefficient, with the signed coefficient used only when the prior-preferred option remains supported.
- **Validation across temporal reasoning tasks (Section 5).** Across eight temporal sub-tasks from four MCQA benchmarks and four frozen backbones, POV improves multiple-choice accuracy by 4.17. On EventHallusion, it improves accuracy by 8.62. Ablations separate the contributions of option-wise verification, prior correction, and coefficient estimation.

2 Related Work

Language priors and calibration. Question-only shortcuts and answer priors are well documented in VQA (Goyal et al., 2017; Agrawal et al., 2018), motivating training-time debiasing methods based on adversarial objectives, question-only branches, and bias-aware ensembles (Ramakrishnan et al., 2018; Cadene et al., 2019; Clark et al., 2019). Counterfactual VQA further removes

Table 1: **Temporal reasoning sub-tasks used in the MCQA evaluation.** We evaluate eight sub-tasks from TempCompass (Liu et al., 2024), MVBench (Li et al., 2024b), Video-MME-v2 (Fu et al., 2026), and Perception Test (Patraucean et al., 2023). The sparse-frame budget is $K = \max(4, \lfloor 0.25K_{\mathrm{ref}} \rfloor)$. The last column reports the rate of prior-following errors defined in Eq. 1, averaged over the four backbones. † denotes a metadata-filtered subset.

Benchmark	Sub-task	Avg. # Options	Ref. Frames (K_{ref})	Sparse Frames (K)	Total # Questions	Avg. # PFEs (%)
TempCompass	Direction, Speed, AttrChange	2.0	8–10	4	906	21.0
MVBench	StateChange, MovingDir, CounterInfer	3.4	5–26	4–6	600	21.5
Video-MME-v2	Change	8.0	64	16	205	21.8
Perception Test	Descriptive†	3.0	10–35	4–8	300	22.0

estimated direct language effects at inference (Niu et al., 2021). Related language-model calibration methods estimate answer preferences from content-free inputs (Zhao et al., 2021), noisy-channel scoring (Min et al., 2022), or domain-conditional priors (Holtzman et al., 2021). In VideoLLMs, BOLD analyzes and calibrates option-selection bias in multiple-choice prediction (Loginova et al., 2025). These studies establish the broader role of language-side and answer-selection biases in multimodal decision making.

Training-free multimodal correction. Training-free methods modify inference by contrasting or correcting model responses under alternative conditions. VCD contrasts original and distorted visual inputs (Leng et al., 2024), while M3ID increases visual grounding relative to language-side preference through mutual-information decoding (Favero et al., 2024). Video methods construct temporal contrasts through downsampling, reversal, or spatial-temporal negatives (Zhang et al., 2024a; Kong et al., 2025; Wu et al., 2026). DINO-HEAL reweights visual features using DINOv2 attention (Li et al., 2025a), whereas MGAP attenuates hidden-state components projected onto an estimated language-prior subspace (Zhuang et al., 2026). In contrast, we perform correction entirely in option-score space, using each question's observed no-video score vector while estimating only the removal scale from unlabeled target-task scores.

Sparse-frame video reasoning and verification. Methods for sparse-frame video understanding improve visual observation through selective frame or token processing (Lei et al., 2021; Yu et al., 2023; Nie et al., 2024; Chen et al., 2024). Complementary diagnostic studies examine video hallucination and how visual or temporal information evolves inside multimodal models (Wang et al., 2024b; Li et al., 2025a;b; Wang et al., 2025; Kim et al., 2026). Verification has been used for candidate ranking and model self-assessment (Cobbe et al., 2021; Kadavath et al., 2022; Lightman et al., 2024), while multiple-choice prompting can exhibit sensitivity to option identifiers and ordering (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024), motivating candidate-wise evaluation on a common output space. PIVOT further explores option-wise visual testing for resolving prior–evidence conflicts in image VLMs (Yu et al., 2026). We instead use the same YES/NO output pair for every candidate, enabling paired video and no-video measurements in a common verification space.

3 RETAINED TEMPORAL EVIDENCE IN PRIOR-FOLLOWING ERRORS

For an example with sparse frames V, question Q, candidate options $\mathcal{O} = \{o_1, \ldots, o_m\}$, and correct option $y \in \mathcal{O}$, let $\hat{o}^{\mathrm{video}}$ and $\hat{o}^{\mathrm{no\text{-}video}}$ denote the model's predictions with and without the video, respectively. Formally, a PFE satisfies

$$\hat{o}^{\mathrm{video}} = \hat{o}^{\mathrm{no\text{-}video}} \neq y. \tag{1}$$

The criterion is purely observational and identifies cases in which an incorrect video-conditioned prediction agrees with the prediction obtained without video. For the analyses below, each candidate $o \in \mathcal{O}$ is independently evaluated as a YES/NO claim using the same frozen model. Let $\ell_{\mathrm{Yes}}(X, Q, o)$ and $\ell_{\mathrm{No}}(X, Q, o)$ denote the model's output logits for the "Yes" and "No" tokens under visual input X. We define the paired verification scores as

$$v(o) = \ell_{\mathrm{Yes}}(V, Q, o) - \ell_{\mathrm{No}}(V, Q, o), \qquad p(o) = \ell_{\mathrm{Yes}}(\varnothing, Q, o) - \ell_{\mathrm{No}}(\varnothing, Q, o), \tag{2}$$

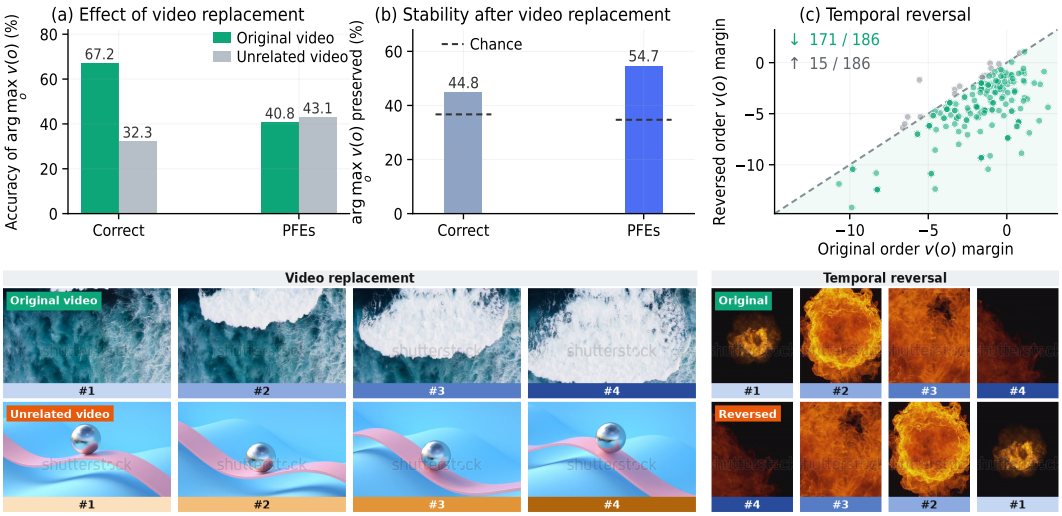


Figure 2: **Verification-based video interventions on PFEs.** (a) Accuracy of $\arg\max_o v(o)$ with the original and an unrelated replacement video for correct predictions and PFEs. (b) Fraction of examples for which $\arg\max_o v(o)$ is unchanged after video replacement; dashed lines denote subset-specific chance levels. (c) For the 186 TempCompass-AttrChange PFEs, the $v(o)$ margin is the score difference between the correct and prior-preferred options. Reversal lowers this margin in 171 cases, which fall below the diagonal. The frame strips below illustrate the interventions.

where $\varnothing$ denotes the identical textual prompt with the visual input removed. We refer to $v(o)$ and $p(o)$ as the video-conditioned and no-video verification scores, respectively. The prior-preferred option is the candidate with the highest $p(o)$. The correct-prediction and PFE subsets in the following analyses are defined by the model's standard multiple-choice predictions, whereas the intervention metrics are computed from these option-wise verification scores.

Video-replacement intervention. We first examine whether the option preference under verification depends on question-specific visual content. We replace the original video with an unrelated video from the same sub-task and recompute the option-wise verification scores $v(o)$. Because this intervention changes the visual content itself, we evaluate whether the top-ranked option under $v(o)$ remains stable rather than attributing individual score changes to a specific visual factor. The replacement video is selected deterministically by question identifier so that the pairing remains fixed across analyzed subsets (Figure 2(a)). For correct predictions, the original–unrelated gap in the accuracy of $\arg\max_o v(o)$ is $+34.9$, whereas the corresponding gap for PFEs is only -2.3. Consistently, the top-ranked option under $v(o)$ is preserved more often after replacement for PFEs than for correct predictions (Figure 2(b)). These results indicate substantially weaker dependence of the verification preference on the question-specific video for PFEs. A control on the remaining incorrect predictions shows a similar tendency, indicating that the PFE criterion identifies agreement with the no-video prediction rather than establishing causal prior influence (Appendix B).

Temporal-order reversal. Weak dependence under video replacement does not imply an absence of useful temporal evidence. We therefore preserve the same sparse frames and reverse only their temporal order on TempCompass-AttrChange (Figure 2(c)). With visual content fixed, we measure the verification margin between the correct and prior-preferred options rather than whether the top-ranked option changes. Across the 186 PFEs, reversal reduces this margin by 2.47 on average, with decreases in 171 cases. Thus, the sampled frames retain discriminative temporal information even when the final prediction agrees with the no-video prediction.

Layer-wise diagnosis. We further examine whether the correct and prior-preferred options remain distinguishable within the decoder. For the layer-wise analysis, we define the paired scores $v^{\text{layer}}(o)$ and $p^{\text{layer}}(o)$ by projecting the last-position hidden state through the model's final normalization and LM head. Figure 3 tracks the resulting video-induced change $v^{\text{layer}}(o) - p^{\text{layer}}(o)$ across decoder

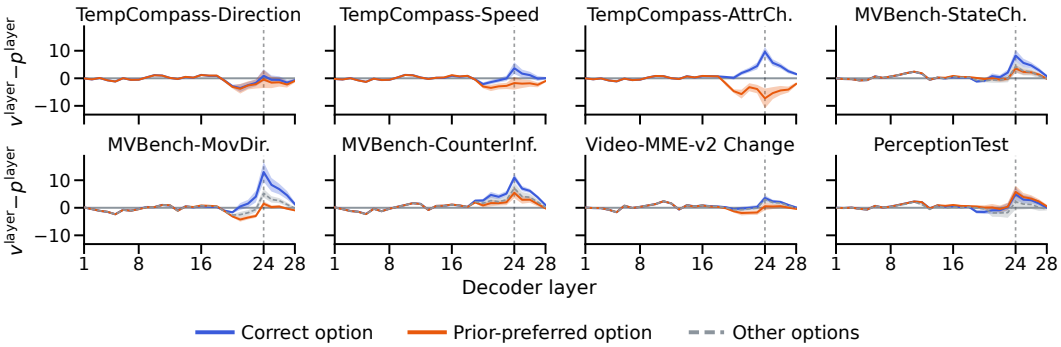


Figure 3: **Layer-wise video-induced changes in option verification.** For Qwen2-VL, $v^{layer}(o) - p^{layer}(o)$ denotes the difference between the video-conditioned and no-video verification scores computed at each decoder layer using the model's final normalization and LM head. We evaluate questions for which the prior-preferred option is incorrect. Shaded regions show 95% bootstrap confidence intervals over questions.

layers. The correct option is favored over the prior-preferred option in six of eight sub-tasks, with a peak advantage of $+2.98$ to $+16.87$. This separation peaks at layer 24 of 28 and then weakens toward the output; on MVBench-MovingDir, for example, it decreases from $+11.52$ to $+1.46$. These findings indicate that discriminative temporal information remains accessible in intermediate representations even when the final decision aligns with the no-video preference, motivating the paired correction introduced in the following section.

4 POV: Paired Option-Wise Verification

4.1 Option-wise verification

POV uses the paired verification scores in Eq. 2 as its measurement primitive. Under the same YES/NO normalization, each score is the binary log-odds of the claim, and their difference measures the video-induced change in verification for that claim. Figure 4 shows the paired evaluations, which differ only in whether the video is present. The claim follows each task's native answer format. General multiple-choice QA (e.g., Caption-based options (Robinson et al., 2022; Cai et al., 2025)) are verified independently, while native binary QA (Zhang et al., 2024a) uses the original YES/NO question directly.

This option-wise construction differs from ordinary multiple-choice scoring in that each candidate is evaluated independently rather than through a joint score over the full candidate set, so its verification score does not directly depend on the other candidate options. Moreover, using the same YES/NO output pair places the scores in a shared verification space (Figure 5(a)), making the video-conditioned and no-video scores directly comparable.

4.2 Estimating the prior-removal coefficients

We estimate the coefficients independently within each model–sub-task pair, which we refer to as a *cell*. Consider a cell containing N unlabeled questions. For each question $i \in \{1, \dots, N\}$, let $M_i = |\mathcal{O}_i|$, and denote by $\mathbf{v}_i, \mathbf{p}_i \in \mathbb{R}^{M_i}$ its video-conditioned and no-video option-score vectors, respectively. We center these vectors within each question, $\tilde{\mathbf{v}}_i = \mathbf{v}_i - \bar{v}_i \mathbf{1}$ and $\tilde{\mathbf{p}}_i = \mathbf{p}_i - \bar{p}_i \mathbf{1}$, and stack the centered vectors over the N unlabeled questions into $\mathbf{Z}_v$ and $\mathbf{Z}_p$ (Figure 4). Centering removes option-invariant score offsets, while the final decision is evaluated on the original scores.

$$\lambda_{\text{proj}} = \arg\min_{\lambda} \|\mathbf{Z}_v - \lambda \mathbf{Z}_p\|_2^2 = \frac{\langle \mathbf{Z}_v, \mathbf{Z}_p \rangle}{\|\mathbf{Z}_p\|_2^2}. \tag{3}$$

The residual $\mathbf{Z}_v - \lambda_{\text{proj}} \mathbf{Z}_p$ is orthogonal to $\mathbf{Z}_p$.

An alternative is to fit a separate projection coefficient from each question. In that case, the same small set of option scores determines both the question-specific prior vector $\mathbf{p}_i$ and how much of that vector is removed, so question-specific alignment between $\mathbf{v}_i$ and $\mathbf{p}_i$ is absorbed directly into

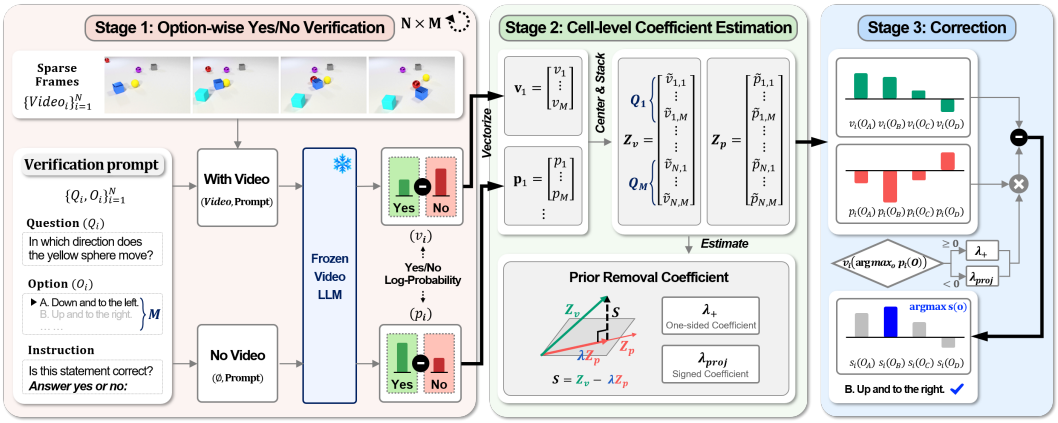


Figure 4: **Overview of our method.** Each answer option is converted into a YES/NO claim and scored by the same frozen VideoLLM under two inputs: the video-conditioned input and the no-video input. The corresponding yes–no logit differences define the video-conditioned verification score $v(o)$ and the prior score $p(o)$. POV forms the corrected score $s(o) = v(o) - \lambda p(o)$ and predicts the option with the largest $s(o)$. The correction coefficients are estimated from unlabeled option scores at the cell level, with selective use of the signed coefficient when the prior-preferred option remains supported.

the fitted coefficient. We instead estimate only the removal scale from multiple questions in the cell. The actual correction remains question-specific because question i subtracts its own prior-score vector, $\lambda \mathbf{p}_i$. The coefficient is computed entirely from unlabeled scores within each cell and is not interpreted as latent prior strength. Its value varies substantially across cells, from 0.10 to 1.28. Figure 5(b) further shows that λ_{proj} achieves a higher mean accuracy gain than coefficients selected with labels by cross-validation within each cell, while a label-aware search over its global scale selects 1.05, close to the default scale of 1. At inference, we score each option as

$$s_i(o) \;=\; v_i(o) - \lambda_i\, p_i(o), \qquad \hat{o}_i \;=\; \arg\max_{o \in \mathcal{O}_i} s_i(o). \tag{4}$$

When positive and negative question-wise alignments are aggregated with sign, opposing contributions can reduce the estimated prior-aligned support. We therefore form a one-sided estimate using only positive alignments:

$$\tilde{\lambda}_+ = \frac{\sum_i [\langle \tilde{\mathbf{v}}_i, \tilde{\mathbf{p}}_i \rangle]_+}{\sum_i \|\tilde{\mathbf{p}}_i\|_2^2}. \tag{5}$$

where $[x]_+ = \max(x, 0)$ denotes the positive part. This preserves prior-aligned contributions while excluding opposing ones. To cap the one-sided estimate at unit subtraction while preserving a larger signed coefficient, we use

$$\lambda_+ = \max\{\lambda_{\text{proj}},\ \min(1, \tilde{\lambda}_+)\}, \tag{6}$$

which serves as the default value of λ_i in Eq. 4.

4.3 SELECTIVE USE OF THE SIGNED COEFFICIENT

We use the one-sided coefficient λ_+ by default and switch to the signed coefficient λ_{proj} only when the prior-preferred option remains supported by the video-conditioned verifier. Let $o_i^p = \arg\max_o p_i(o)$ denote the option preferred by the no-video verification scores. For the general multiple-choice QA tasks (e.g., (Robinson et al., 2022)), we define:

$$\lambda_i = \begin{cases} \lambda_{\text{proj}}, & \text{if } v_i(o_i^p) \geq 0, \\ \lambda_+, & \text{otherwise.} \end{cases} \tag{7}$$

Because $v_i(o_i^p)$ is a YES $-$ NO logit difference, a nonnegative value indicates support rather than rejection. Otherwise, the one-sided coefficient is retained.

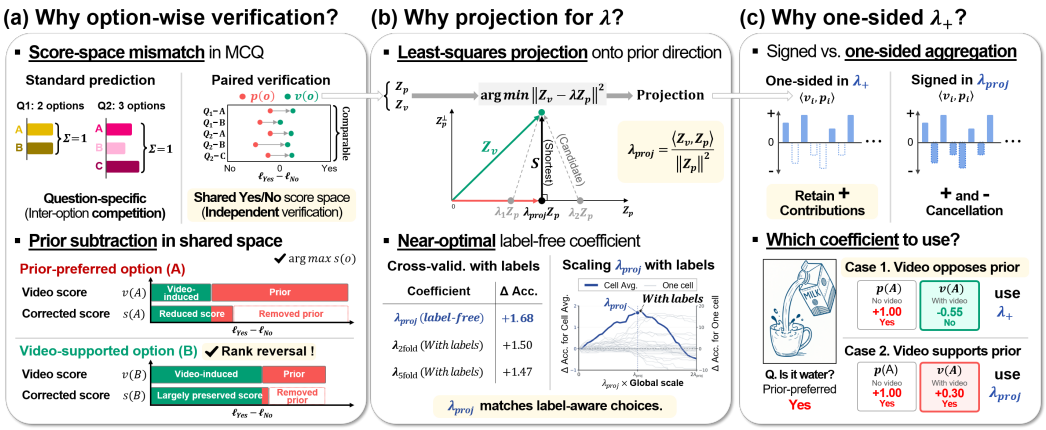


Figure 5: **Design principles of POV.** (a) Paired option-wise verification maps options from different questions onto a shared verification score space, making prior subtraction comparable across questions. (b) Least-squares projection estimates a shared label-free coefficient λ_{proj}, which matches choices from 2- and 5-fold cross-validation with labels within each cell, while its original scale is near the label-aware optimum. (c) One-sided aggregation retains positive question-wise alignments, yielding the default coefficient λ_+. Signed aggregation includes negative alignments and can cancel positive contributions, while λ_{proj} is applied when the prior-preferred option remains supported.

4.4 The POV decision

With λ_i determined by Eq. 7, POV applies Eq. 4 directly in the verification score space. The cell-level coefficients λ_{proj} and λ_+ are computed once from the unlabeled paired scores, while Eq. 7 selects the coefficient for each question using the same scores as shown in the Figure 5(c).

For the general multiple-choice QA (Robinson et al., 2022; Cai et al., 2025), paired verification requires $2|\mathcal{O}_i|$ option evaluations, whereas native binary QA (Zhang et al., 2024a) requires one video-conditioned and one no-video evaluation. Once these scores are obtained, coefficient selection and correction are purely score-level operations and require no additional model evaluations, ground-truth labels, or gradient-based optimization.

5 Experiments

Protocol. We evaluate four frozen VideoLLMs: LLaVA-Video-7B (Zhang et al., 2024b), Qwen2-VL-7B (Wang et al., 2024a), Qwen2.5-VL-7B (Bai et al., 2025), and LLaVA-OneVision-7B (Li et al., 2024a). For MCQA (Robinson et al., 2022), we use the eight temporal sub-tasks in Table 1, comprising $2,011$ questions and 32 combinations of backbone and sub-task. We evaluate the three EventHallusion BinaryQA (Zhang et al., 2024a) sub-tasks on the same backbones using 8 frames, one quarter of the 32-frame reference budget. The coefficients λ_{proj} and $\tilde{\lambda}_+$ are estimated separately for each evaluation cell from unlabeled scores. Appendix A provides the prompts, frame-sampling procedure, and additional implementation details.

5.1 Temporal multiple-choice reasoning

Table 2 summarizes the results on the eight *temporal MCQA sub-tasks*. The VideoLLM baseline, averaged over the four backbones (Zhang et al., 2024b; Wang et al., 2024a; Bai et al., 2025; Li et al., 2024a), improves from 50.95 to 51.83 as the frame budget increases from K to K_{ref}, whereas POV reaches $\mathbf{55.12}$ at the sparse-frame budget, improving the baseline by 4.17. Competing training-free methods show less consistent gains across sub-tasks; for example, MotionCD improves TempCompass-AttrChange but degrades Speed and CounterInfer, leaving its overall mean nearly unchanged. In comparison, POV improves six of eight sub-tasks, all four backbones, and 26 of the 32 cells (Table 2b).

(a) Mean results across backbones. Sparse and dense-frame VideoLLM use K and K_{ref} frames, respectively.

	MCQA										EventHallusion				
	TempCompass			MVBench			Video-MME-v2	Perc.Test	Total					Total	
Method	Direction	Speed	AttrCh.	StateCh.	MovDir.	CountInf.	Change	Descr.	Acc.	Δ	Entire	Mix	Misl.	Acc.	Δ
Sparse-frame VideoLLM (K)	**61.93**	63.92	76.22	49.25	34.75	53.62	17.44	50.50	50.95	–	51.32	56.35	83.58	61.74	–
Dense-frame VideoLLM (K_{ref})	61.24	64.69	78.39	51.12	35.88	54.88	17.20	**51.25**	51.83	1.7% ▲	51.97	66.06	**87.99**	67.60	9.5% ▲
VCD (Leng et al., 2024)	60.55	63.92	74.48	51.75	35.38	55.88	18.78	48.50	51.15	0.4% ▲	51.10	65.54	82.11	65.65	6.3% ▲
TCD (Zhang et al., 2024a)	61.54	61.86	79.25	50.00	35.75	53.62	18.05	49.00	51.13	0.4% ▲	50.66	63.21	82.11	64.43	4.4% ▲
DINO-HEAL (Li et al., 2025a)	61.01	62.29	75.35	50.00	34.75	52.62	18.66	50.67	50.67	0.5% ▼	53.73	52.20	83.09	60.33	2.3% ▼
MotionCD (Kong et al., 2025)	63.99	55.24	**89.24**	47.38	39.88	44.88	18.78	48.17	50.94	0.0% –	**55.26**	70.08	73.77	66.87	8.3% ▲
POV **(ours)**	60.40	**67.27**	80.30	**60.38**	**41.88**	**58.75**	**22.56**	49.42	**55.12**	8.2% ▲	51.75	**75.13**	82.11	**70.35**	**13.9%** ▲

(b) Per-backbone results for sparse-frame VideoLLM, VCD, and POV. Bold denotes the best per column.

	MCQA										EventHallusion				
	TempCompass			MVBench			Video-MME-v2	Perc.Test	Total					Total	
Backbone	Direction	Speed	AttrCh.	StateCh.	MovDir.	CountInf.	Change	Descr.	Acc.	Δ	Entire	Mix	Misl.	Acc.	Δ
Sparse-frame VideoLLM															
LLaVA-Video	62.08	65.64	76.39	48.00	25.50	45.50	16.10	53.00	49.03	–	54.39	54.40	93.14	64.06	–
Qwen2-VL	62.69	62.20	77.43	46.50	41.00	57.50	18.54	49.67	51.94	–	43.86	67.36	61.76	59.41	–
Qwen2.5-VL	63.91	65.98	78.82	53.00	51.00	65.50	20.00	54.00	56.53	–	57.89	50.78	89.22	62.35	–
LLaVA-OneVision	59.02	61.86	72.22	49.50	21.50	46.00	15.12	45.33	46.32	–	49.12	52.85	90.20	61.12	–
Average	**61.93**	63.92	76.22	49.25	34.75	53.62	17.44	**50.50**	50.95	–	51.32	56.35	**83.58**	61.74	–
VCD (Leng et al., 2024)															
LLaVA-Video	63.61	64.60	75.35	55.50	32.50	47.00	20.00	47.67	50.78	3.6% ▲	54.39	69.43	90.20	70.42	9.9% ▲
Qwen2-VL	59.33	61.17	78.12	47.00	40.00	59.00	19.51	48.00	51.52	0.8% ▼	44.74	68.39	66.67	61.37	3.3% ▲
Qwen2.5-VL	61.77	65.98	74.65	55.00	46.50	63.50	19.51	52.33	54.91	2.9% ▼	54.39	58.55	80.39	62.84	0.8% ▲
LLaVA-OneVision	57.49	63.92	69.79	49.50	22.50	54.00	16.10	46.00	47.41	2.4% ▲	50.88	65.80	91.18	67.97	11.2% ▲
Average	60.55▼	63.92–	74.48▼	51.75▲	35.38▲	55.88▲	18.78▲	48.50▼	51.15	0.4% ▲	51.10▼	65.54▲	82.11▼	65.65	6.3% ▲
POV **(ours)**															
LLaVA-Video	58.72	66.67	77.08	65.00	34.50	52.50	22.44	52.00	53.61	9.3% ▲	57.89	75.65	88.24	73.84	15.3% ▲
Qwen2-VL	63.30	65.29	82.29	62.00	41.50	62.50	25.37	44.00	55.78	7.4% ▲	53.51	77.72	73.53	69.93	17.7% ▲
Qwen2.5-VL	63.00	69.76	82.64	55.00	53.50	70.50	23.90	51.33	58.70	3.8% ▲	50.00	70.98	81.37	67.73	8.6% ▲
LLaVA-OneVision	56.57	67.35	79.17	59.50	38.00	49.50	18.54	50.33	52.37	13.1% ▲	45.61	76.17	85.29	69.93	14.4% ▲
Average	60.40▼	**67.27**▲	**80.30**▲	**60.38**▲	**41.88**▲	**58.75**▲	**22.56**▲	49.42▼	**55.12**	8.2% ▲	**51.75**▲	**75.13**▲	82.11▼	**70.35**	**13.9%** ▲

Table 2: **Quantitative results on temporal MCQA and EventHallusion.** Accuracy (%) is reported for each sub-task and evaluation setting. MCQA totals are unweighted means over the eight temporal sub-tasks. EventHallusion totals aggregate the Entire, Mix, and Misleading subsets. ▲ / ▼ indicate improvement or degradation relative to the corresponding sparse-frame VideoLLM.

TempCompass-Direction and Perception Test are the only sub-tasks with negative mean gains. Notably, they are also the only cases without clear separation between the correct and prior-preferred options in the layer-wise analysis ($+0.25$ and -0.85; Table 8).

5.2 Binary Verification on EventHallusion

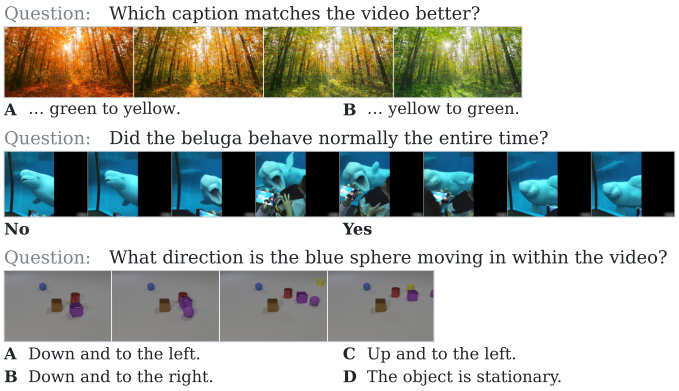
Unlike general multiple-choice QA tasks (Robinson et al., 2022; Cai et al., 2025), EventHallusion (Zhang et al., 2024a) has a native binary answer space (i.e., $o \in \{\text{Yes}, \text{No}\}$). Since the two answer branches directly define the decision axis, we retain the one-sided coefficient λ_+ throughout to prevent opposing video–prior preferences from cancelling the prior-aligned component (Further implementation details are provided in Eq. 11 and Table 11 in the Appendix).

As shown in Table 2, POV improves the 8-frame VideoLLM by 8.61, reaching 70.35 and exceeding the 32-frame VideoLLM of 67.60. Under the same frame budget, VCD and TCD improve the baseline by 3.91 and 2.69, respectively. Because the native binary score already coincides with the verification score, the improvement from 61.74 to 70.35 isolates the effect of prior correction from the option-wise reformulation used for MCQA.

5.3 Method Analysis and Controls

Figure 6 shows how the correction changes the option ranking in representative binary and multiple-choice examples. In each example, the incorrect option receives strong no-video support and remains highly scored under video-conditioned verification. Subtracting the measured prior contribution reduces this support and changes the final ranking in favor of the correct option, illustrating the role of prior correction beyond option-wise verification alone.

Table 3 further examines the contribution of each stage and the granularity of coefficient estimation. On MCQA, option-wise verification alone contributes a gain of 2.53 over the sparse-frame VideoLLM. Estimating λ_+ from each question separately is markedly unstable, whereas pooling



$\lambda_i = 1.18$	A ✗	B ✓
v(option)	+2.12	+2.88
p(option)	+2.06	-1.38
s(option)	-0.32	**+4.50**

$\lambda_i = 0.85$	Yes ✗	No ✓
v(option)	-0.71	-0.82
p(option)	-0.21	-2.05
s(option)	-0.54	**+0.93**

$\lambda_i = 0.40$	A	B ✗	C	D ✓
v(option)	+0.50	+0.56	+0.56	+0.69
p(option)	+0.56	+0.69	+0.50	-0.94
s(option)	+0.28	+0.29	+0.36	**+1.06**

Figure 6: **Option-wise score correction on examples recovered by POV.** We show representative examples from TempCompass, EventHallusion, and MVBench. For each option, we report the video-conditioned score $v(o)$, no-video prior score $p(o)$, and corrected score $s(o) = v(o) - \lambda_i p(o)$. A ✗ marks the base VideoLLM prediction and a ✓ the correct option selected by POV. The base prediction uses the model's standard multiple-choice output and need not equal $\arg\max_o v(o)$.

Table 3: **Granularity of coefficient estimation.** The default one-sided coefficient is estimated either per example or from unlabeled questions within each cell. POV selectively applies the signed coefficient at the question level.

		MCQA		EventHallusion	
Method	Frames	Acc.	Δ	Acc.	Δ
VideoLLM	K	50.95	–	61.74	–
Verification only	K	53.48	5.0% ▲	61.74	–
+ Per-example λ_+	K	46.11	9.5% ▼	51.89	15.9% ▼
+ Per-cell λ_+	K	54.72	7.4% ▲	**70.35**	**13.9%** ▲
+ Selective λ_{proj} **(ours)**	K	**55.12**	**8.2%** ▲	**70.35**	**13.9%** ▲

Table 4: **Inference cost.** Median latency in seconds per sample, averaged over backbones. Speed-up is measured relative to the dense-frame VideoLLM. POV is reported with and without shared video-prefix KV caching.

		MVBench-SC		Video-MME-v2 Ch.	
Method	Frames	Acc.	Latency	Acc.	Latency
VideoLLM	K_{ref}	51.12	0.76s (–)	17.20	9.17s (–)
VideoLLM	K	49.25	**0.17s (4.5×)**	17.44	**2.42s (3.8×)**
VCD	K	51.75	0.36s (2.1×)	18.78	2.77s (3.3×)
POV-Uncached	K	60.38	0.77s (1.0×)	**22.56**	21.73s (0.4×)
POV-Cached	K	**61.25**	0.34s (2.3×)	21.59	3.14s (2.9×)

unlabeled questions within a cell improves accuracy by 8.61 on MCQA and 18.46 on EventHallusion relative to the per-example estimate. This contrast shows that the correction benefits from sharing only the scalar coefficient across questions, while each question retains its own prior-score direction. Selective use of the signed coefficient then recovers an additional 0.40 on MCQA without reducing the EventHallusion gain.

The paired evaluations introduce additional inference cost because each candidate is scored independently. However, the video-conditioned evaluations can share the video–question prefix, allowing the corresponding key–value states to be reused across candidates. Shared-prefix KV reuse changes the aggregate MCQA only marginally while substantially reducing inference latency. As shown in Table 4, it reduces latency from 0.77 to 0.34 seconds on the 3-option MVBench-StateChange and from 21.73 to 3.14 seconds on the 8-option Video-MME-v2 Change. The cached implementation is 2.3× and 2.9× faster than the corresponding dense-frame VideoLLM. Once the paired scores are obtained, coefficient estimation and score correction require no additional model evaluations.

6 Conclusion and Limitations

Conclusion. In this work, we identify prior-following errors in VideoLLMs, where the final prediction agrees with the no-video preference despite retained temporal evidence in the sampled frames by its language priors. To overcome this error, we introduce Paired Option-Wise Verification (POV), a training-free method that measures each option with and without video and corrects prior-aligned support using coefficients estimated from unlabeled scores. POV consistently improves temporal reasoning on both multiple-choice and binary QA, demonstrating the effectiveness of decision-stage correction as a complement to improving visual observation.

Limitations. The gains of POV are not uniform across all temporal tasks, and the current formulation assumes a scorable candidate set or binary answer space. Furthermore, prior correction can attenuate useful video evidence when it is aligned with the no-video preference. Extending POV to open-ended generation remains an important direction for future work.

AI USE STATEMENT

In this work, LLMs were used for sentence-level polishing and to support literature retrieval and discovery. Initial manuscript writing was carried out by the authors, and references identified with LLMs were independently verified. All core contributions, methodology, experiments, and findings are the authors' original work. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The method and coefficient estimators are specified in Section 4. Mathematical derivations, implementation details, evaluation protocols, and prompts are provided in Appendix A. Diagnostic controls, robustness analyses, and additional evaluations are reported in Appendices B–E. All reported results are produced from the stated model outputs and evaluation procedures.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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.

Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Yao Feng, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. In *Workshop on Video-Language Models @ NeurIPS 2024*, 2025. URL https://openreview.net/forum?id=bhuZn4URBH.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language Processing (EMNLP-IJCNLP)*, pp. 4069–4082, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-Modal Hallucination Control by Visual Information Grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu

Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding, 2026. URL https://arxiv.org/abs/2604.05015.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7038–7051, 2021.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-LLM Based Video Frame Selection for Efficient Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712. IEEE, 2025.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual Token Pruning for Efficient Video Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Minji Kim, Taekyung Kim, and Bohyung Han. Map the Flow: Revealing Hidden Pathways of Information in VideoLLMs. In *International Conference on Learning Representations*, volume 2026, pp. 50839–50870, 2026.

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.

Yeonkyung Lee, Dayun Ju, Youngmin Kim, Seil Kang, and Seong Jae Hwang. Vikey: Enhancing temporal understanding in videos via visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38880–38890, June 2026.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7331–7341, June 2021.

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.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13723–13733. IEEE, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVbench: A comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024b.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal Reasoning Transfer from Text to Video. In *International Conference on Learning Representations*, volume 2025, pp. 74070–74102, 2025b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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.

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, 2024.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing Blind Guessing: Calibration of Selection Bias in Multiple-Choice Question Answering by Video Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3216–3246, 2025.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy Channel Language Model Prompting for Few-Shot Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5316–5330, 2022.

Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, Hang Xu, and Li Zhang. Slow-Focus: Enhancing Fine-grained Temporal Understanding in Video LLM. *Advances in Neural Information Processing Systems*, 37:81808–81835, 2024.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12700–12710, June 2021.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. *Advances in neural information processing systems*, 31, 2018.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging Large Language Models for Multiple Choice Question Answering. *arXiv preprint arXiv:2210.12353*, 2022.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive Keyframe Sampling for Long Video Understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. VideoHallucer: Evaluating Intrinsic and Extrinsic Hallucinations in Large Video-Language Models. *arXiv preprint arXiv:2406.16338*, 2024b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11096–11105, June 2026.

Hang Yu, Hong Sun, Xiaochen Liu, Jiahao Wen, Ruihan Wang, Shilv Cai, Haoning Alice Xi, and Zhenbo Li. Pivot: Prior-interrupting visual option testing for counterintuitive image understanding with frozen vlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3027–3033, June 2026.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-Chained Image-Language Model for Video Localization and Question Answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. EventHallusion: Diagnosing Event Hallucinations in Video LLMs. *arXiv preprint arXiv:2409.16597*, 2024a.

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*, 2024b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Yingxuan Zhuang, Jingxiao Yang, Miao Pan, Cheng Tan, Yuxiang Cai, Siwei Tan, Chen Zhi, Xuhong Zhang, Jianwei Yin, and Jintao Chen. Mitigating Manifold Departure: Uncertainty-Aware Subspace Rectification for Trustworthy MLLM Decoding. *arXiv preprint arXiv:2606.09859*, 2026.

APPENDIX CONTENTS

A Method Derivation and Implementation Details **14**

B Further Analysis of Prior-Following Errors **17**

C Analysis of Option-Wise Verification **18**

D Robustness of Coefficient Estimation **19**

E Generalization and Efficiency **20**

F Additional Qualitative Visualization **22**

A METHOD DERIVATION AND IMPLEMENTATION DETAILS

Paired-score interpretation. For each option o, we associate a binary correctness variable $T_o \in \{0, 1\}$ and define the induced verifier by renormalizing the model scores over the two answer branches. Let Y and N denote $T_o{=}1$ and $T_o{=}0$, respectively, and let X denote the visual-input condition:

$$\pi_X(Y) = \frac{e^{\ell_{\text{Yes}}(X,Q,o)}}{e^{\ell_{\text{Yes}}(X,Q,o)} + e^{\ell_{\text{No}}(X,Q,o)}}, \qquad v(o) = \log \frac{\pi_{V_K}(Y)}{\pi_{V_K}(N)}, \qquad p(o) = \log \frac{\pi_{\varnothing}(Y)}{\pi_{\varnothing}(N)}. \tag{8}$$

Because the two branches share the same normalizer, $\log[\pi_X(Y)/\pi_X(N)] = \ell_{\text{Yes}}(X, Q, o) - \ell_{\text{No}}(X, Q, o)$, so Eq. 8 is equivalent to Eq. 2. Consequently, $v(o) - p(o)$ is exactly the video-induced change in the Yes–No logit difference.

The same identity admits a probabilistic interpretation. If the renormalized YES/NO scores are interpreted as probabilities over T_o, the video-conditioned and no-video scores correspond to posterior and prior log-odds, respectively. By Bayes' rule,

$$v(o) - p(o) = \log \frac{P(V_K \mid T_o = 1, Q, o)}{P(V_K \mid T_o = 0, Q, o)},$$

so unit prior subtraction isolates the log-likelihood ratio contributed by the video. More generally,

$$v(o) - \lambda p(o) = \log \frac{P(V_K \mid T_o = 1, Q, o)}{P(V_K \mid T_o = 0, Q, o)} + (1 - \lambda) \log \frac{P(T_o = 1 \mid Q, o)}{P(T_o = 0 \mid Q, o)}.$$

Thus, λ controls how much of the measured no-video preference is retained in the corrected score. The following projection provides a label-free estimate of this correction strength directly from the paired score geometry.

Coefficient estimation. Using the centered cell-level score vectors $\mathbf{Z}_v$ and $\mathbf{Z}_p$ defined in Section 4.2, let $\mathbf{\Delta}_Z = \mathbf{Z}_v - \mathbf{Z}_p$. We model the paired score difference as

$$\mathbf{\Delta}_Z = \mathbf{E} + c\,\mathbf{Z}_p, \tag{9}$$

where $\mathbf{E}$ denotes video-dependent evidence and the second term lies along the measured prior direction. This decomposition is not identifiable along $\mathbf{Z}_p$: for any $t \in \mathbb{R}$, $(\mathbf{E} + t\mathbf{Z}_p, c - t)$ gives the same $\mathbf{\Delta}_Z$. The component invariant to this ambiguity is its orthogonal projection,

$$\mathbf{R} = \Pi_{\mathbf{Z}_p^{\perp}} \mathbf{\Delta}_Z = \mathbf{Z}_v - \lambda_{\text{proj}}\,\mathbf{Z}_p, \qquad \lambda_{\text{proj}} = \frac{\langle \mathbf{Z}_v, \mathbf{Z}_p \rangle}{\|\mathbf{Z}_p\|_2^2}. \tag{10}$$

The coefficient λ_{proj} is therefore an operational projection coefficient rather than an estimate of latent prior strength. In particular, $\lambda_{\text{proj}} = 1$ indicates that the paired score difference has no component

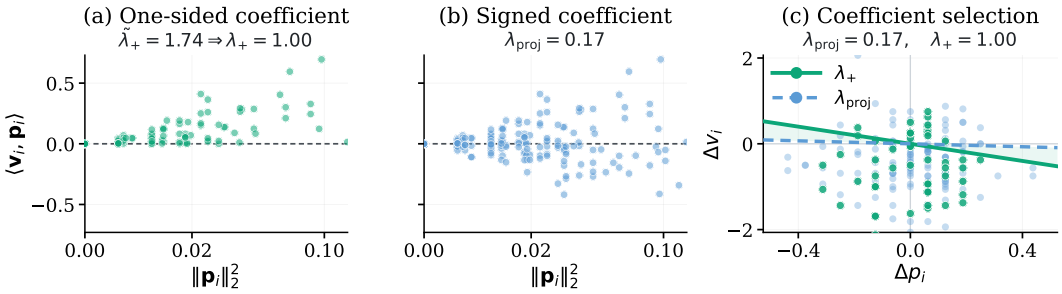


Figure 7: **Geometry of coefficient estimation.** One MVBench-CounterInfer/Qwen2-VL cell at $K = 4$. (a) One-sided aggregation excludes negative alignments. (b) Signed aggregation includes both signs and can cancel positive contributions. (c) λ_+ is the default, while λ_{proj} is selected when $v_i(o_i^p) \geq 0$. Axes in (a,b) use a square-root scale.

along $\mathbf{Z}_p$, whereas deviations from 1 quantify the corresponding alignment. Because the projection removes the component parallel to $\mathbf{Z}_p$ irrespective of its origin, video-dependent evidence aligned with the same direction may also be attenuated. The projection is therefore justified by the identifiable score geometry rather than by an assumption that every removed component is uninformative. Figure 7 illustrates the signed and one-sided coefficient geometry together with the question-level selection rule. The one-sided estimate follows from the corresponding question-wise nonnegative least-squares fit. For each question,

$$\arg\min_{c \geq 0} \|\tilde{\mathbf{v}}_i - c\tilde{\mathbf{p}}_i\|_2^2 = \frac{[\langle \tilde{\mathbf{v}}_i, \tilde{\mathbf{p}}_i \rangle]_+}{\|\tilde{\mathbf{p}}_i\|_2^2}.$$

Averaging these solutions with weights $\|\tilde{\mathbf{p}}_i\|_2^2$ recovers $\tilde{\lambda}_+$ in Eq. 5. Thus, negative question-wise projections are excluded before aggregation rather than clipping λ_{proj} afterward.

Finally, question-wise centering does not affect the final decision: centering $v_i(o) - \lambda p_i(o)$ subtracts the same quantity $\bar{v}_i - \lambda \bar{p}_i$ from every option. We therefore estimate the coefficient from centered vectors while applying the correction to the original scores.

Evaluation protocol. We uniformly sample K normalized frame positions over $[0.02, 0.98]$ and resize frames to 384×384; TempCompass uses $[0.05, 0.95]$. Reference budgets sample one frame per second up to 64 frames. The sparse/reference budgets are $4/10$ for Direction and AttrChange, $4/8$ for Speed, $6/26$ for StateChange, $4/5$ for MovingDir and CounterInfer, and $16/64$ for Video-MME-v2 Change. Perception Test uses $K = \max(4, \lfloor 0.25 K_{\text{ref}} \rfloor)$ (4–8, median 8), with median $K_{\text{ref}} = 35$.

For MCQA, $v_i(o)$ and $p_i(o)$ use the first-token *Yes–No* logit difference, taking the maximum over corresponding surface variants when needed. Coefficients are estimated within each cell from centered question-level scores, with $\epsilon = 10^{-9}$ for numerical stability. We use three prompt forms according to the benchmark answer format. For MVBench, Video-MME-v2, and Perception Test, each candidate is inserted into `{question}\nStatement: "{option}"\nBased on the video, is this statement correct? Answer yes or no:`. TempCompass captions are verified directly without the question stem. EventHallusion uses `{question}\nPlease answer yes or no:`. The no-video branch uses the same textual prompt with the visual input removed.

For EventHallusion, we form:

$$\mathbf{v}_i = [\ell_i(\textit{Yes}), \ell_i(\textit{No})], \tag{11}$$

and $\mathbf{p}_i$ analogously without video, where ℓ_i is the first-token log-probability maximized over the corresponding surface variants. The vectors are centered and processed by the same coefficient estimation and correction; the common log-normalization offset cancels under centering and $\arg\max$, making logits and log-probabilities equivalent here.

External baselines. We adapt VCD (Leng et al., 2024) using the released contrastive rule and hyperparameters: a sigmoid β schedule over 1000 steps, $t = 500$ ($\sqrt{\bar{\alpha}_t} = 0.863$), $\alpha = 1$, and a plausibility

Algorithm 1 Paired Option-Wise Verification (POV)

Require: Frozen VideoLLM; sparse frames $\{V_i\}_{i=1}^{N}$; unlabeled questions and option sets $\{(Q_i, \mathcal{O}_i)\}_{i=1}^{N}$
Ensure: Predictions $\{\hat{o}_i\}_{i=1}^{N}$

Paired option-wise verification

for each question i and option $o \in \mathcal{O}_i$ **do**
  $v_i(o) \leftarrow \ell_{\mathrm{Yes}}(V_i, Q_i, o) - \ell_{\mathrm{No}}(V_i, Q_i, o)$
  $p_i(o) \leftarrow \ell_{\mathrm{Yes}}(\varnothing, Q_i, o) - \ell_{\mathrm{No}}(\varnothing, Q_i, o)$
end for

Cell-level coefficient estimation

Form $\mathbf{v}_i = [v_i(o)]_{o\in\mathcal{O}_i}$ and $\mathbf{p}_i = [p_i(o)]_{o\in\mathcal{O}_i}$
$\tilde{\mathbf{v}}_i \leftarrow \mathbf{v}_i - \bar{v}_i\mathbf{1}$ and $\tilde{\mathbf{p}}_i \leftarrow \mathbf{p}_i - \bar{p}_i\mathbf{1}$ for each question i
Stack $\{\tilde{\mathbf{v}}_i\}_{i=1}^{N}$ and $\{\tilde{\mathbf{p}}_i\}_{i=1}^{N}$ into $\mathbf{Z}_v$ and $\mathbf{Z}_p$
$\lambda_{\text{proj}} \leftarrow \langle \mathbf{Z}_v, \mathbf{Z}_p \rangle / \|\mathbf{Z}_p\|_2^2$ (Eq. 3)
$\tilde{\lambda}_+ \leftarrow \sum_i [\langle \tilde{\mathbf{v}}_i, \tilde{\mathbf{p}}_i \rangle]_+ / \sum_i \|\tilde{\mathbf{p}}_i\|_2^2$ (Eq. 5)
$\lambda_+ \leftarrow \max\{\lambda_{\text{proj}}, \min(1, \tilde{\lambda}_+)\}$ (Eq. 6)

Question-level correction

for each question i **do**
  $o_i^p \leftarrow \arg\max_{o\in\mathcal{O}_i} p_i(o)$
  $\lambda_i \leftarrow \begin{cases} \lambda_{\text{proj}}, & v_i(o_i^p) \geq 0, \\ \lambda_+, & \text{otherwise.} \end{cases}$ (Eq. 7)
  $\hat{o}_i \leftarrow \arg\max_{o\in\mathcal{O}_i} \big(v_i(o) - \lambda_i p_i(o)\big)$ (Eq. 4)
end for
return $\{\hat{o}_i\}_{i=1}^{N}$

threshold of $\beta = 0.1$ relative to the whole-vocabulary maximum of the clean branch. The diffusion perturbation is applied to the preprocessed tensor at the same stage as in the released implementation, with identical frames, questions, options, and prompt templates in the clean and corrupted branches. We implement *MotionCD* (Kong et al., 2025) from the released decision rule. The contrast $(1+\alpha)z_{\text{ori}} - \alpha z_{\text{rev}}$ is masked below $\log\beta + \max z_{\text{ori}}$, where the maximum is taken over the full vocabulary. We follow the repository settings $\alpha = 20$ (halved internally) and $\beta = 0.1$. Because the released implementation does not support the evaluated backbones, we port its decision rule while keeping the frames, questions, options, and prompt templates fixed across comparisons. For *TCD* (Zhang et al., 2024a), the contrastive branch uses $\max(1, \lfloor K/4 \rfloor)$ uniformly sampled frames, following the proposed temporal downsampling, with $\alpha = \beta = 0.5$.

For VCD, MotionCD, and TCD, we apply the contrastive rule and plausibility constraint at the first answer-token position and select among the valid answer tokens.

For *DINO-HEAL* (Li et al., 2025a), we obtain saliency by averaging the attention from the CLS token to patch tokens across heads in the final DINOv2-large layer and resize the resulting map to the backbone's visual token grid. The saliency map is normalized by its per-frame mean and used to reweight the patch features before projection.

Table 5: Effect of replacing the original video with an unrelated clip from the same model–sub-task cell. Verification accuracy is computed from $\arg\max_o v(o)$. The final column reports whether the top-ranked option is preserved after replacement, with chance agreement in parentheses.

		Verification accuracy (%)			Top-1
Question group	n	Original	Unrelated	Gap [95% CI]	Unchanged
Correct predictions	1497	67.2	32.3	$+34.9$ $[+32.1, +37.9]$	44.8% *(36.7)*
Prior-following errors	1019	40.8	43.1	-2.3 $[-5.9, +1.3]$	54.7% *(34.7)*
Other incorrect predictions	1156	18.9	26.5	-7.6 $[-10.7, -4.5]$	38.1% *(27.1)*

Table 6: **Decision recovery and video-source control.** (a) compares the sparse-frame baseline, option-wise verification, and prior correction. (b) evaluates the correction using the original or an unrelated video, with the coefficient either re-estimated or fixed to its original-video value. Results are reported on the complete cells of Table 5 and weighted by cell size.

(a) Decision recovery

	Accuracy (%)	
Decision rule	All	PFE
Baseline	33.15	0.0 *(by constr.)*
Verification only	—	22.1
Projection (λ_{proj})	**41.17**	**38.3**

(b) Video-source control

Configuration		
Video	λ	Accuracy (%)
Original	Fitted	**41.17**
Unrelated	Re-fitted	28.83
Unrelated	Fixed	29.51

Table 7: **Temporal-order reversal on TempCompass-AttrChange PFEs.** The verification margin is the score difference between the correct and prior-preferred options. Δ_{temp} denotes the original-order margin minus the reversed-order margin. Brackets denote 95% bootstrap confidence intervals from 20,000 draws; "Pos." counts examples with $\Delta_{\text{temp}} > 0$. One exact tie is excluded.

Subset	n	Original margin	Reversed margin	Δ_{temp}	Pos.
All PFEs	186	-1.76	-4.24	$+\mathbf{2.47}$ $[+2.16, +2.81]$	$171/186$
POV does not recover	100	-3.08	-4.89	$+1.81$ $[+1.47, +2.16]$	$87/100$
POV recovers	86	-0.24	-3.48	$+3.24$ $[+2.71, +3.79]$	$84/86$
Original margin > 0	43	$+0.82$	-3.25	$+4.07$ $[+3.34, +4.82]$	$42/43$
Original margin ≤ 0	143	-2.54	-4.53	$+1.99$ $[+1.68, +2.32]$	$129/143$

B Further Analysis of Prior-Following Errors

Dependence on the input video. We replace each original video with a deterministically paired unrelated clip from the same cell, keeping the question, options, prompt, and frame budget fixed. As shown in Table 5, verification accuracy decreases substantially for correct predictions but not for either error group, and PFEs preserve the top-ranked option more often than expected by chance. Other incorrect predictions show the same qualitative pattern, indicating that weak video dependence is not unique to PFEs. The PFE criterion instead isolates errors whose final decision agrees with the no-video prediction; it does not imply that the sampled frames contain no useful information.

Table 6 provides two complementary controls. On PFEs, verification alone recovers 22.1%, whereas prior correction reaches 38.3%, showing that the verification ordering is informative but insufficient by itself. Replacing the original video with an unrelated one reduces corrected accuracy from 41.17 to 28.83 with a re-estimated coefficient and 29.51 with the original coefficient, both below the sparse-frame baseline. Thus, the correction depends on question-specific video-conditioned measurements rather than prior subtraction alone.

Temporal reversal. We reverse the fixed frame order for all 186 TempCompass-AttrChange PFEs across the four backbones. Table 7 shows that reversal shifts the verification margin away from the correct option in every analyzed subset, including the 100 PFEs that POV does not recover. Temporal-order information can therefore remain available even when it does not change the final decision. Applying the same intervention to physically reversed TempCompass clips produces consistent score changes (correlation ≥ 0.950 across backbones), supporting that the effect reflects temporal ordering rather than visual-content changes.

Layer-wise decodability. We project each decoder layer's last-position hidden state through the final normalization and LM head, using the resulting Yes–No logit difference as a layer-wise verification score. We analyze Qwen2-VL on the first 40 questions per cell for which the prior-preferred option is incorrect; remaining-option scores are averaged within each question and confidence intervals use question-level bootstrap resampling.

Figure 3 and Table 8 show that separation between the correct and prior-preferred options emerges in intermediate layers and generally attenuates toward the output. This logit-lens analysis measures linear decodability rather than mechanistically localizing video reasoning, and is diagnostic only.

Table 8: Layer-wise separation between the correct and prior-preferred options at the peak decoder layer of each cell. The table reports the analyzed population and 95% bootstrap confidence interval for the separation.

Sub-task	n/N	Peak	True	Prior-pref.	Rem.	Separation [95% CI]
TempCompass-Direction	15/40	21	-3.56	-3.81	–	$+0.25$ $[-2.03, +2.59]$
TempCompass-Speed	26/40	24	$+3.67$	-1.73	–	$+5.40$ $[+2.27, +8.56]$
TempCompass-AttrChange	14/40	24	$+9.61$	-7.27	–	$+16.87$ $[+12.97, +20.88]$
MVBench-StateChange	21/40	24	$+8.26$	$+3.59$	$+4.25$	$+4.67$ $[+2.17, +7.20]$
MVBench-MovingDir	33/40	24	$+12.98$	$+1.46$	$+5.25$	$+11.52$ $[+7.92, +15.18]$
MVBench-CounterInfer	28/40	24	$+10.97$	$+5.45$	$+7.11$	$+5.53$ $[+3.28, +7.85]$
Video-MME-v2 Change	31/40	24	$+3.53$	$+0.55$	$+2.98$	$+2.98$ $[+1.33, +4.75]$
Perception Test	24/40	24	$+4.92$	$+5.77$	$+2.19$	-0.85 $[-3.83, +2.12]$

Table 9: Option-wise verification compared with joint-softmax scoring on 14 MCQA cells with no-video joint scores. Both methods use the same backbone, frames, prompt, and label-free coefficient estimation; only the score primitive differs. Matched-budget rows constrain the same fraction of base decisions to be overridden. Here, $\hat{h}$ and h_{gt} denote the predicted and ground-truth answer-position distributions.

Method	Joint softmax	Option-wise verification
Net accuracy change at matched override budget		
5% overridden	+0.53	**+1.50**
10% overridden	+0.44	**+2.33**
20% overridden	+0.74	**+2.86**
30% overridden	+1.47	**+2.77**
Accuracy on baseline errors (%)		
PFEs ($n = 801$)	31.71	**35.58**
Remaining errors ($n = 1012$)	11.96	**18.48**
No-video answer-position bias $\frac{1}{2}\|\hat{h} - h_{\mathrm{gt}}\|_1$		
Score used for prior subtraction	0.153	**0.063**

C Analysis of Option-Wise Verification

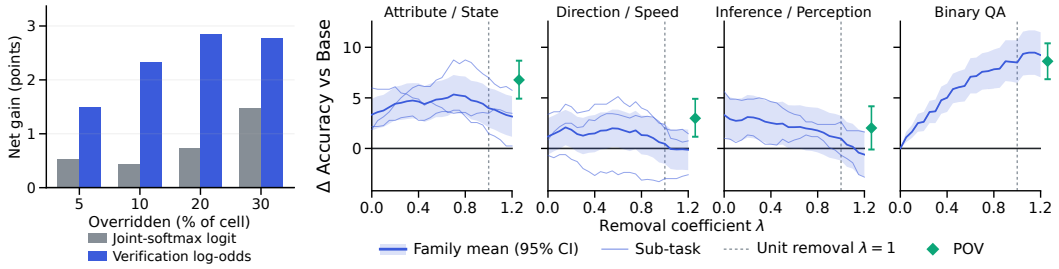
Option-wise versus joint scoring. As an alternative to option-wise verification, we assign each candidate its forced-choice log-probability, $\ell(o) = \log P(o \mid V_K, Q, \mathcal{O})$, together with the corresponding no-video score $\ell_{\mathrm{txt}}(o)$. By Bayes' rule, $\ell(o) = \log P(V_K \mid o, Q, \mathcal{O}) + \log P(o \mid Q, \mathcal{O}) - \log P(V_K \mid Q, \mathcal{O})$, making the score dependent on the full candidate set. Three distinctions are relevant.

i) Choice-set dependence. Modifying the distractor set changes the normalized joint score, whereas $v(o)$ evaluates each candidate independently.

ii) Fixed output space. Joint scoring over option strings or identifiers can inherit length, frequency, or position effects, whereas option-wise verification scores the same *Yes*/*No* output pair for every candidate.

iii) Cross-question comparability. The paired scores $(v(o), p(o))$ share a common verification space, allowing questions with different numbers of options to use a cell-level coefficient. For native binary QA, the two scoring forms coincide by construction.

We further compare the two scoring rules at matched rates of overriding the base prediction. Figure 8(a) and Table 9 show that option-wise verification produces a larger net correction at every matched rate, including separately on PFEs and the remaining baseline errors. Table 9 also compares the no-video arg-max position distribution with the correct distribution at the same option count. The joint probe shows substantially greater position bias than option-wise verification, consistent with known sensitivity of multiple-choice scoring to option identifiers and ordering (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024). We treat this as a diagnostic difference between the score spaces rather than a complete explanation of the accuracy gap.



(a) Matched override budgets (b) Coefficient sensitivity

Figure 8: **(a)** Net correction from the two scoring rules under matched override budgets, with examples ranked by each rule's own margin. **(b)** Accuracy relative to the sparse-frame VideoLLM when a single coefficient is imposed on an entire task family. Thin curves denote the eight sub-tasks and the bold curve denotes the corresponding family mean.

Table 10: Coefficient estimation without self-inclusion. We re-estimate λ_{proj} and $\tilde{\lambda}_{+}$ from the full unlabeled cell, under leave-one-out estimation, or from the opposite fold in two-fold cross-fitting. Δ is relative to the sparse-frame baseline.

	MCQA		EventHallusion	
Coefficient source	Accuracy (%)	Δ	Accuracy (%)	Δ
Sparse-frame VideoLLM	50.95	–	61.74	–
Full unlabeled cell	55.12	+4.17	70.35	+8.62
Leave-one-out	55.17	+4.22	70.35	+8.62
Two-fold cross-fit	55.06	+4.11	70.22	+8.49

Shuffled-prior control. To test whether the correction depends on the question-specific no-video direction, we permute prior-score vectors across questions within each option-count stratum, using 200 permutations per cell while keeping the model scores and estimator fixed. This preserves the marginal score distribution and vector norms while breaking the question–prior pairing. The mean fitted coefficient decreases from 0.574 to 0.007, and MCQA accuracy decreases from 55.15 to 53.30, close to the verification-only result of 53.48 in Table 3. Thus, the gain depends on the measured question-specific prior direction rather than an arbitrary score-space perturbation.

D Robustness of Coefficient Estimation

Coefficient estimation. The coefficients in POV are estimated from unlabeled questions within each model–sub-task cell. We re-estimate the cell-level quantities under leave-one-out estimation and two-fold cross-fitting averaged over 20 random partitions. Table 10 shows comparable accuracy, indicating that the gains do not depend on the evaluated question contributing to its own coefficient. We therefore use the transductive estimator in the main experiments.

Figure 5(b) additionally compares λ_{proj} with label-aware references. Within each MCQA cell, we select λ from $\{0, 0.05, \ldots, 2.0\}$ by two- or five-fold cross-validation and average over 40 random partitions. Relative to verification only ($\lambda = 0$), these references yield $+1.50$ and $+1.47$, respectively, compared with $+1.68$ for the label-free λ_{proj}. A label-aware global scaling of the cell-wise λ_{proj} selects 1.05 and yields $+1.76$, close to the default scale of 1 and its $+1.68$ gain.

Coefficient choice and robustness. Table 11 compares the estimated coefficients with fixed references: $\lambda = 0$ isolates option-wise verification, while $\lambda = 1$ applies unit prior subtraction. On MCQA, the signed coefficient λ_{proj} outperforms unit subtraction, whereas the one-sided coefficient λ_{+} yields higher EventHallusion accuracy. Figure 8(b) shows corresponding heterogeneity across task families: attribute/state tasks favor an intermediate coefficient, inference/perception degrades as the coefficient increases, whereas binary QA continues to improve at larger coefficients.

Table 11: Ablation of score correction and coefficient selection. Δ is measured relative to the sparse-frame baseline.

		MCQA		EventHallusion	
Decision rule	Coefficient	Accuracy (%)	Δ	Accuracy (%)	Δ
Sparse-frame VideoLLM	Joint logits	50.95	—	61.74	—
Verification only	$\lambda = 0$	53.48	+2.52	61.74	+0.00
Unit subtraction	$\lambda = 1$	52.86	+1.91	70.23	+8.50
Signed coefficient throughout	λ_{proj}	55.15	+4.20	66.69	+4.95
One-sided coefficient throughout	λ_+	54.72	+3.77	70.35	+8.62
Selective signed coefficient **(ours)**	$\lambda_i \in \{\lambda_{\text{proj}}, \lambda_+\}$	**55.12**	+4.17	**70.35**	+8.62

Table 12: Effect of frame budget on the VideoLLM baseline and POV across all 32 MCQA cells. Δ is computed relative to the VideoLLM baseline at the corresponding frame budget.

	Sparse budget					Reference budget				
Sub-task	K	VideoLLM	POV **(ours)**	Δ	PFEs (%)	K_{ref}	VideoLLM	POV **(ours)**	Δ	PFEs (%)
TempC-Direction	4	61.9	**60.4**	−1.5	23.5	10	61.2	**60.5**	−0.7	23.8
TempC-Speed	4	63.9	**67.3**	+3.4	22.9	8	64.7	**66.8**	+2.1	22.3
TempC-AttrCh	4	76.2	**80.3**	+4.1	16.1	10	78.4	**80.7**	+2.3	14.7
MVB-StateCh	6	49.2	**60.4**	+11.2	33.6	26	51.1	**61.0**	+9.9	33.4
MVB-MovDir	4	34.8	**41.9**	+7.1	18.0	5	35.9	**41.8**	+5.9	18.1
MVB-CountInf	4	53.6	**58.8**	+5.2	13.0	5	54.9	**57.8**	+2.9	11.5
Video-MME-v2 Ch	16	17.4	**22.6**	+5.2	21.8	64	17.2	**23.2**	+6.0	19.8
Perception Test	4–8	50.5	**49.4**	−1.1	22.0	30	51.2	**49.7**	−1.5	22.2
MCQA		**50.95**	**55.12**	**+4.17**	**21.4**		**51.83**	**55.16**	**+3.33**	**20.7**

Table 13: Class-wise EventHallusion performance pooled over the four backbones (1,636 questions). In addition to accuracy, we report recall for both answer classes and the predicted-*yes* rate to distinguish improved discrimination from a global class-prior shift.

K	Method	Accuracy	Recall (*Yes*)	Recall (*No*)	Pred.-*Yes* rate
8	VideoLLM	61.74	57.85	63.69	43.52
	λ_{proj}	66.69	72.08	63.97	48.11
	POV **(ours)**	**70.35**	**80.11**	**65.44**	**49.82**
32	VideoLLM	67.60	68.61	**67.10**	44.87
	λ_{proj}	68.70	75.36	65.35	48.29
	POV **(ours)**	**72.62**	**85.22**	66.27	**50.98**

Applying λ_+ throughout improves EventHallusion from 66.69 to 70.35 relative to λ_{proj} but reduces MCQA accuracy to 54.72. Selectively using λ_{proj} when the prior-preferred option remains supported preserves the EventHallusion result and increases MCQA accuracy to 55.12. Thus, the selective rule approaches the signed-coefficient result on MCQA while retaining the one-sided gain on EventHallusion. Removing the unit cap changes aggregate accuracy by less than 0.2 on either benchmark.

E Generalization and Efficiency

Reference-frame evaluation. Table 12 applies the same decision rule to the reference frame budget of each benchmark using the identical question set, with both coefficients re-estimated from the corresponding unlabeled scores. Because the no-video branch contains no visual input, its prior score is unchanged across frame budgets. Increasing the frame budget improves the VideoLLM baseline by less than one point on average and changes the PFE rate only modestly, whereas the POV gain remains substantial. The sparse-budget result is within 0.04 of the reference-budget POV accu-

Table 14: **Results on four TemporalBench categories.** Accuracy (%) is reported for each backbone and category. The total is the unweighted mean over the four evaluated categories. ▲ / ▼ indicate improvement or degradation relative to the corresponding sparse-frame VideoLLM.

	TemporalBench					
	Category				Total	
Backbone	Event Order	Motion Dir.	Action Freq.	Motion Mag.	Acc.	Δ
Sparse-frame VideoLLM						
LLaVA-Video	66.75	50.20	43.96	59.69	55.15	–
Qwen2-VL	67.65	51.17	48.11	60.62	56.89	–
Qwen2.5-VL	68.37	60.68	60.19	60.31	62.39	–
LLaVA-OneVision	64.27	52.86	43.58	55.94	54.16	–
Average	**66.76**	53.73	48.96	59.14	57.15	–
VCD (Leng et al., 2024)						
LLaVA-Video	66.22	52.47	49.06	60.94	57.17	3.7% ▲
Qwen2-VL	67.32	50.07	46.42	60.31	56.03	1.5% ▼
Qwen2.5-VL	66.17	57.62	58.87	59.69	60.59	2.9% ▼
LLaVA-OneVision	63.55	53.84	47.55	56.56	55.38	2.2% ▲
Average	65.82 ▼	53.50 ▼	50.47 ▲	59.38 ▲	57.29	0.3% ▲
POV **(ours)**						
LLaVA-Video	58.41	52.80	64.15	66.88	60.56	9.8% ▲
Qwen2-VL	61.17	57.68	57.92	66.56	60.84	6.9% ▲
Qwen2.5-VL	62.22	59.70	65.28	64.38	62.89	0.8% ▲
LLaVA-OneVision	63.36	56.25	57.74	63.12	60.12	11.0% ▲
Average	61.29 ▼	**56.61** ▲	**61.27** ▲	**65.23** ▲	**61.10**	**6.9%** ▲

Table 15: Inference latency with shared-prefix KV reuse. **(a)** Median seconds per sample across four backbones for MCQA cells. **(b)** Seconds per sample on MVBench-StateChange for sparse-frame VideoLLM (K=6), dense-frame VideoLLM (K_{ref}=26), and POV at the sparse budget.

(a) Across tasks

Benchmark	# opt.	# cells	Latency (s)
MVB-StateChange	3.0	4	0.412
Perception Test	3.0	4	0.473
MVB-CounterInfer	3.2	4	0.368
MVB-MovingDir	4.0	4	0.381
VideoMME-v2 Change	8.0	4	3.136

(b) Across backbones

Backbone	Sparse	Dense	POV **(ours)**
LLaVA-Video	0.209	0.899	**0.303**
Qwen2-VL	0.160	0.719	**0.386**
Qwen2.5-VL	0.199	0.894	**0.430**
LLaVA-OneVision	0.117	0.533	**0.233**
Average	0.171	0.761	**0.338**

racy and exceeds the reference-budget VideoLLM while using the prescribed sparse-frame budgets. These results indicate that the correction is not specific to a single frame budget and do not support attributing prior-following errors to sparsity alone.

Class-wise EventHallusion analysis. To distinguish improved discrimination from a simple shift in the predicted class distribution, Table 13 reports class-wise recall together with the predicted-*yes* rate. At eight frames, POV improves recall for both classes, by 22.26 for *Yes* and 1.75 for *No*, while the predicted-*yes* rate increases from 43.52 to 49.82. The gain therefore cannot be attributed solely to a global shift toward *Yes*. The aggregate gain is concentrated on the Mix split, while Entire changes little and Misleading decreases slightly.

Generalization to TemporalBench. To examine whether the gains of POV are specific to the temporal sub-tasks used in our main evaluation, we additionally evaluate four categories from TemporalBench (Cai et al., 2025): Event Order, Motion Direction, Action Frequency, and Motion Magnitude. These categories were not used in the development of POV, and we retain the same method formulation and sparse-frame evaluation protocol. As shown in Table 14, POV improves the mean accuracy across the four categories from 57.15 to 61.10, corresponding to a 6.9% relative improvement over the sparse-frame VideoLLM, compared with 0.3% for VCD. The gain is consistent across all four frozen backbones and is observed in three of the four categories, with the largest improvement on Action Frequency.

Inference cost. After the paired option scores are computed, coefficient estimation and score correction require no additional model evaluations. The option evaluations share the same video–question

Table 16: **Accuracy with shared-prefix KV reuse.** MCQA accuracy of POV per backbone without and with KV reuse. For the cached rows, Δ is the accuracy change relative to the corresponding uncached row.

	TempCompass			MVBench			Video-MME-v2	Perc.Test	**Total**	
Backbone	Direction	Speed	AttrCh.	StateCh.	MovDir.	CountInf.	Change	Descr.	Acc.	Δ
POV-Uncached										
LLaVA-Video	58.72	66.67	77.08	65.00	34.50	52.50	22.44	52.00	53.61	–
Qwen2-VL	63.30	65.29	82.29	62.00	41.50	62.50	25.37	44.00	55.78	–
Qwen2.5-VL	63.00	69.76	82.64	55.00	53.50	70.50	23.90	51.33	58.70	–
LLaVA-OneVision	56.57	67.35	79.17	59.50	38.00	49.50	18.54	50.33	52.37	–
Average	60.40	67.27	80.30	60.38	41.88	58.75	22.56	49.42	55.12	–
POV-Cached										
LLaVA-Video	59.33	65.64	77.43	65.00	34.00	50.50	20.00	51.67	52.95	−0.66 ▼
Qwen2-VL	62.39	65.98	82.29	62.50	42.50	63.00	23.41	44.00	55.76	−0.02 ▼
Qwen2.5-VL	63.61	70.10	82.29	56.50	53.00	70.50	23.90	51.33	58.90	+0.20 ▲
LLaVA-OneVision	59.02	65.98	78.12	61.00	39.50	48.00	19.02	49.67	52.54	+0.17 ▲
Average	61.09	66.92	80.03	61.25	42.25	58.00	21.59	49.17	55.04	−0.08 ▼

prefix, allowing its key–value states to be reused across candidates. Table 15 reports cached latency across representative tasks and backbones.

Latency is measured at batch size one on a single NVIDIA RTX PRO 6000 GPU in bfloat16, with CUDA synchronization around each measurement. Frame decoding is excluded. All methods use the same samples after ten disjoint warm-up samples: 25 questions for MVBench-StateChange and 20 for Video-MME-v2 Change. We record one pass per sample and report the median latency for each backbone, averaged over the four backbones.

Table 16 additionally compares POV accuracy with and without shared-prefix KV reuse. Reusing the prefix changes the execution path and can introduce small numerical differences under bfloat16 computation, occasionally affecting near-tied predictions. The MCQA average changes only from 55.12 to 55.04, with no consistent degradation across backbones.

F Additional Qualitative Visualization

Score-level examples. Table 17 presents eight PFEs corrected by POV and reports the option text, video-conditioned verification score v, no-video score p, and corrected score $s_i(o) = v_i(o) - \lambda_i p_i(o)$. The examples are drawn from 195 corrected PFEs across the two MVBench subtasks and four backbones for which option text is available, with at most two examples per backbone and no repeated questions. In each case, the VideoLLM prediction agrees with the prior-preferred option, while the corrected score favors the correct option.

Across these examples, the no-video scores are often more strongly separated than the video-conditioned scores, while the corrected ranking is determined by weaker within-question differences. The magnitude and sign of these differences vary across questions, making a fixed threshold on the verification score insufficient to reproduce the same corrections.

Frame-level examples. Figures 9 and 10 show the exact sampled frames for examples in which the relevant visual evidence is directly interpretable. This selection criterion is used only for visualization and does not affect the quantitative evaluation. For EventHallusion, the annotated event presupposition receives strong no-video support, whereas the video-conditioned verification score is comparatively less separated; applying Eq. 4 removes the corresponding prior-aligned contribution and changes the final decision.

Table 17: **Qualitative PFE recoveries with POV.** For each option, v, p, and s denote the video-conditioned, no-video, and corrected scores. $\star$ marks the correct option, and boldface the POV prediction.

EXAMPLE 1 LLAVA-VIDEO *If the person tried to turn on the electric device and it was not working, what was the reason for that?*

	option	v	p	s
	The electric device did not work because the person unplugged the device during the video.	+0.44	+0.25	+0.36
$\star$ *POV*	The device worked when the person tried to turn it on.	**+0.94**	-1.88	**+1.50**
VideoLLM	The electric device did not work most likely because it was already unplugged at the beginning of the video.	+0.19	+0.91	-0.09

EXAMPLE 2 QWEN2.5-VL *Can you identify the direction of the gray cube's movement in the video?*

	option	v	p	s
	Up and to the left.	-0.25	-2.19	-0.56
$\star$ *POV*	Down and to the left.	**+0.19**	-1.50	**-0.03**
	Down and to the right.	+0.00	-1.19	-0.17
VideoLLM	The object is stationary.	-1.75	-1.00	-1.89

EXAMPLE 3 QWEN2-VL *Is the bag empty at the end?*

	option	v	p	s
$\star$ *POV*	No	**-0.12**	+0.19	**-0.27**
VideoLLM	Yes	-1.12	+0.81	-1.73
	The person doesn't interact with a bag.	-2.62	+0.50	-3.00

EXAMPLE 4 LLAVA-ONEVISION *What direction is the blue sphere moving in within the video?*

	option	v	p	s
	Up and to the left.	+0.44	+0.50	+0.24
	Down and to the left.	+0.50	+0.56	+0.28
VideoLLM	Up and to the right.	+0.50	+0.62	+0.25
$\star$ *POV*	The object is stationary.	**+0.88**	-0.94	**+1.25**

EXAMPLE 5 LLAVA-ONEVISION *Can you identify the direction of the yellow sphere's movement in the video?*

	option	v	p	s
	Down and to the left.	+0.38	+0.56	+0.15
$\star$ *POV*	The object is stationary.	**+1.19**	+0.06	**+1.16**
	Up and to the right.	+0.31	+0.62	+0.06
VideoLLM	Down and to the right.	+0.50	+0.75	+0.20

EXAMPLE 6 QWEN2-VL *The person interacts with a lighting device among other objects. Is the lighting device on at any point?*

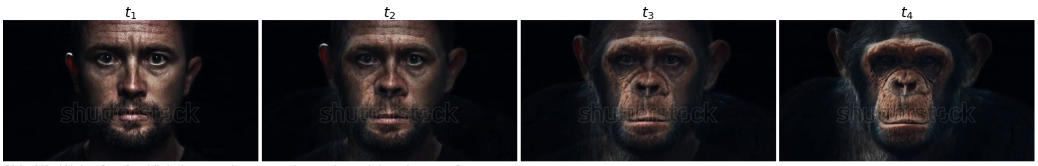
	option	v	p	s
$\star$ *POV*	no	**+1.12**	+0.19	**+0.98**
VideoLLM	yes	+0.06	+0.62	-0.41
	I don't know	+0.12	+0.00	+0.12

EXAMPLE 7 QWEN2.5-VL *In which direction does the blue sphere move in the video?*

	option	v	p	s
	Up and to the left.	+0.19	-2.44	-0.16
$\star$ *POV*	Down and to the left.	**+0.44**	-1.94	**+0.16**
	The object is stationary.	-2.31	-3.75	-2.85
VideoLLM	Up and to the right.	+0.00	-1.69	-0.24

EXAMPLE 8 LLAVA-VIDEO *What can you say about the temperature of the water being poured?*

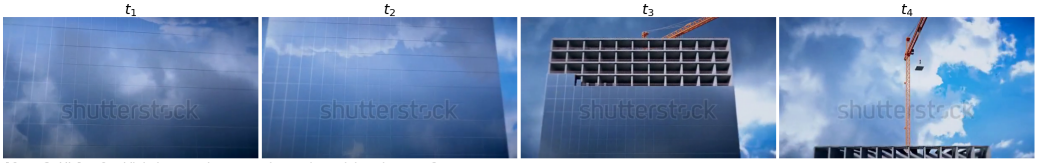
	option	v	p	s
	No water was poured.	-0.56	-0.69	-0.35
$\star$ *POV*	The water seems hot.	**+0.88**	+0.19	**+0.82**
VideoLLM	The water seems cold.	+0.69	+0.62	+0.50



[LLaVA-Video] Q: Which caption matches the video better?
(A) The face is morphing from a chimpanzee into a human. ✗ VideoLLM
(B) The face is morphing from a human into a chimpanzee. ✓ POV (correct)
v(A,B)=(+1.44,+2.94) p=(-1.22,-1.12) s = v - 0.971 p = (+2.62,+4.03) margin=1.409



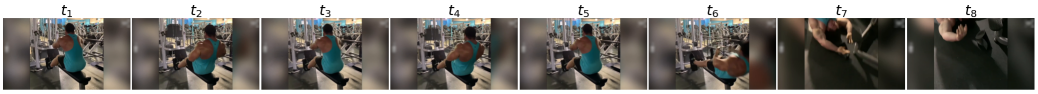
[Qwen2-VL] Q: Which caption matches the video better?
(A) The glass is going from full of water to empty. ✓ POV (correct)
(B) The glass is going from empty to full of water. ✗ VideoLLM
v(A,B)=(+2.31,+2.25) p=(+1.69,+2.06) s = v - 1.184 p = (+0.31,-0.19) margin=0.506



[Qwen2-VL] Q: Which caption matches the video better?
(A) The building is staying the same. ✗ VideoLLM
(B) The building is demolishing. ✓ POV (correct)
v(A,B)=(-0.25,-1.88) p=(+0.62,-0.94) s = v - 1.184 p = (-0.99,-0.77) margin=0.225



[LLaVA-Video] Q: Which caption matches the video better?
(A) The black ink is expanding from the center. ✗ VideoLLM
(B) The black ink is shrinking into the center. ✓ POV (correct)
v(A,B)=(+1.56,+1.75) p=(+0.00,-0.03) s = v - 0.971 p = (+1.56,+1.78) margin=0.218

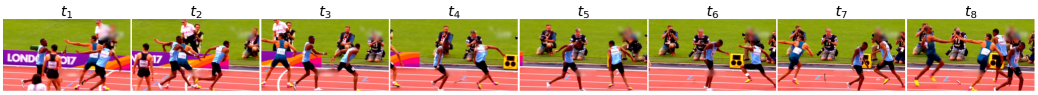


[Qwen2-VL | mix] Q: Did the man complete his rowing practice in this video?
the question presupposes: the clip proceeds normally throughout
what the video shows: The connecting rope of the fitness equipment suddenly broke, causing the man to fall to the ground.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-0.66,-0.79) p=(-0.49,-0.99) s = v - 0.870 p = (-0.24,+0.07) margin=0.310

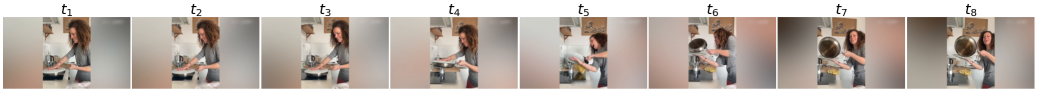


[Qwen2-VL | entire] Q: Is the woman enjoying eating the food in this video?
the question presupposes: A woman enjoys food very much.
what the video shows: A woman spat out food.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-1.09,-1.59) p=(-0.29,-1.54) s = v - 0.870 p = (-0.84,-0.25) margin=0.588

Figure 9: **Recoveries in both answer formats.** Exact sampled frames are shown in playback order. Top: four TempCompass attribute-change examples at $K = 4$ with opposing temporal descriptions. Bottom: two EventHallusion examples at $K = 8$, where the observed frames contradict an event favored by the no-video score.



[LLaVA-Video | mix] Q: Did the athletes successfully run the relay race in the video?
the question presupposes: the clip proceeds normally throughout
what the video shows: The relay baton slipped and fell to the ground.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-0.65,-0.78) p=(-0.24,-1.62) s = v - 0.821 p = (-0.46,+0.55) margin=1.004



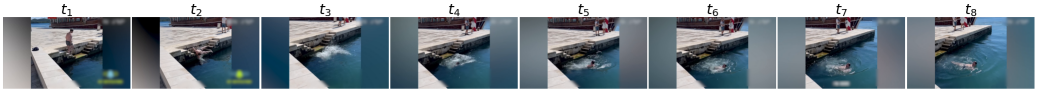
[LLaVA-Video | mix] Q: Was the pouring process smooth and controlled throughout the video?
the question presupposes: the clip proceeds normally throughout
what the video shows: Food fell onto the table and stove.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-0.49,-0.99) p=(-0.24,-1.61) s = v - 0.821 p = (-0.29,+0.34) margin=0.629



[Qwen2-VL | misleading] Q: Is this person engaging in a badminton competition in this video?
the question presupposes: This athlete is currently playing badminton.
what the video shows: This athlete is performing a spinning racket trick.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-0.70,-0.70) p=(-0.29,-1.41) s = v - 0.870 p = (-0.45,+0.53) margin=0.979



[Qwen2-VL | entire] Q: Is the dog playing with the girl in this video?
the question presupposes: The dog is playing with the girl.
what the video shows: The dog is pulling a cart to transport a little girl.
(A) Yes ✗ VideoLLM (= the no-video prior)
(B) No ✓ POV (correct)
v(A,B)=(-0.60,-0.85) p=(-0.29,-1.41) s = v - 0.870 p = (-0.35,+0.38) margin=0.729



[Qwen2.5-VL | mix] Q: Did any accident or anything unexpected happen in the video?
the question presupposes: the clip proceeds normally throughout
what the video shows: He slipped and fell.
(A) Yes ✓ POV (correct)
(B) No ✗ VideoLLM (= the no-video prior)
v(A,B)=(-0.83,-0.58) p=(-4.52,-0.02) s = v - 0.495 p = (+1.41,-0.57) margin=1.976



[LLaVA-OneVision | mix] Q: Did any accident or anything unexpected happen in the video?
the question presupposes: the clip proceeds normally throughout
what the video shows: The pizza accidentally fell to the ground.
(A) Yes ✓ POV (correct)
(B) No ✗ VideoLLM (= the no-video prior)
v(A,B)=(-0.77,-0.65) p=(-1.42,-0.29) s = v - 1.000 p = (+0.65,-0.35) margin=1.000

Figure 10: **EventHallusion recoveries across backbones.** Six examples spanning all four backbones and three splits are shown at $K = 8$. In each case, the no-video score supports the presupposed event, the video-conditioned scores provide little separation, and prior correction shifts the decision toward the answer supported by the frames.