Unstitching the Chimera: Frame-Level Risk and Train-Free Mitigation for Video Hallucination

Supplementary Material

1. Related Works

1.1. MLLMs for Video

Building on the success of LLMs [1, 14, 31, 43, 48], multimodal large language models (MLLMs) have advanced visual language understanding rapidly [21]. LLaVA [27–29] popularized instruction tuning on GPT-4 [1] curated data and established a widely adopted recipe for constructing MLLMs, inspiring many follow up systems [5, 7, 55]. Fueled by short-video applications and video generation, video understanding [46, 50, 53] has emerged as a focal area, yet spatio-temporal joint modeling remains challenging, demanding fine-grained spatial grounding and multi-hop reasoning, especially in long-video settings [9, 24, 32, 52]. To improve efficiency and scalability, Vamba [33] introduces a hybrid Mamba–Transformer architecture with near-linear token complexity, while recent efforts [17, 25, 42] explore GRPO-based reinforcement fine-tuning for VideoLLMs and agentic pipelines for planning and tool use. In practice, closed-source commercial systems (e.g., GPT-4o [19], Gemini [37]) currently lead on video understanding, whereas open-source families such as Qwen-VL Series [2–4, 38], InternVL Series [10, 11, 39, 56], and VideoLLaMA Series [13, 49, 50] provide widely used, reproducible research baselines.

1.2. Hallucinations in MLLMs

Research on multimodal hallucination has been predominantly developed in the image setting, where taxonomies and diagnostic protocols coalesce around three major types: object hallucination [15, 18, 34, 44], relationship hallucination [12, 41, 47], and decoding-related hallucination [8, 35, 36]. Within this landscape, two mitigation families have become standard. Retraining-based methods [16, 20, 22] introduce additional supervision or alignment objectives, while train-free techniques [8, 35, 36, 45, 54] act at inference time and are attractive for deployment due to their low integration cost.

In contrast, hallucinations in video remain less mechanistically characterized [6, 23, 51]. Temporal reasoning introduces cross-frame dependencies and causal chains [30, 40], making image-style token-level analyses ill-suited for multi-frame structures. Moreover, many VideoLLMs [4, 26, 49, 50] inherit image–text pretraining, creating a static-to-dynamic distribution mismatch that weakens modeling of true frame order and cross-frame causal structure, thereby elevating hallucination risk in temporal settings.

2. Event-Segment Induction

We operate on the uniformly sampled frame set $V = \{f_t\}_{t=1}^{T}$ used by the model at inference. Event segments $\mathcal{S} = \{S_m\}_{m=1}^{M}$ are induced in a single forward pass by fusing two complementary boundary cues that are both available without extra supervision. The first cue is an appearance change signal computed from early visual features. For each frame f_t, we extract an L_2-normalized feature vector ϕ_t by average-pooling the visual tokens at an early encoder layer (before heavy cross-frame fusion); in backbones with explicit per-frame embeddings, we use the per-frame token directly. We then compute adjacent-frame cosine similarity $s_t = \langle \phi_t, \phi_{t+1} \rangle$ and define an appearance-drop score $d_t^{\text{app}} = 1 - s_t$. The second cue is a cross-frame attention cut extracted from the model's attention tensors. For each middle fusion layer and head, we form the inter-frame attention matrix $A^{(\ell,h)} \in \mathbb{R}^{T\times T}$ by summing attention mass from visual tokens in frame t to visual tokens in frame u and normalizing rows to sum to one. We then average across heads and selected fusion layers to obtain $\bar{A}$. A boundary between t and $t{+}1$ corresponds to a drop in mutual attention, captured by $d_t^{\text{att}} = 1 - \frac{1}{2}(\bar{A}_{t,t+1} + \bar{A}_{t+1,t})$. The fused boundary score is

$$b_t \;=\; w_{\text{app}}\, \tilde{d}_t^{\text{app}} \;+\; w_{\text{att}}\, \tilde{d}_t^{\text{att}}, \tag{1}$$

where tildes denote per-video z-score normalization and $w_{\text{app}}, w_{\text{att}} \in [0,1]$ with $w_{\text{app}}{+}w_{\text{att}}{=}1$ (default $0.5/0.5$). We smooth $\{b_t\}$ with a length-3 median filter and detect local peaks subject to a minimum distance of three frames. To obtain a robust threshold, we use the larger of two criteria per video: an absolute threshold $\mu_b{+}\kappa\sigma_b$ with $\kappa{=}1.0$, or a top-R rule with $R = \lfloor T/L_{\min} \rfloor$ and $L_{\min}{=}4$ frames. Peaks closer than two frames are merged, and segments shorter than $L_{\min}$ are absorbed into the neighbor with higher peak prominence. This procedure is deterministic, runs in $O(T)$, and requires only tensors already available during the single forward pass. In all experiments, the same defaults are used for all models and datasets; sensitivity analyses show that varying $\kappa \in [0.8, 1.2]$ or $L_{\min} \in [3, 5]$ changes the number of segments marginally and does not alter conclusions. For very short clips ($T \le 8$), we disable peak selection and treat the clip as a single segment to avoid over-segmentation. All boundaries are computed on the sampled frames and are therefore aligned with attention extraction, ensuring consistency with the main text.

3. Layerwise Attention Extraction

We instrument the attention modules during a single forward pass to read the tensors required by CH-Risk and CH-M without re-running the model. For models with explicit cross-attention from text queries to visual keys/values (e.g., LLaVA-style adapters, Q-Former variants), we register lightweight hooks to capture the pre-softmax logits and softmax probabilities at each layer and head. For unified-token transformers that mix modalities in self-attention, we use the token-to-frame index map emitted by the video tokenizer to aggregate attention mass per frame. The depth is normalized by layer index so that "early", "middle", and "late" refer to the first 30%, middle 40%, and last 30% of layers within the fusion stack, respectively; this percentile grouping makes the procedure architecture-agnostic across backbones with different depths.

The frame-level support $p(f)$ is extracted from a designated middle layer near answer formation. We consider two query sources: (i) the token immediately preceding the first generated answer token (pre-answer position), and (ii) an automatically detected subset of temporal keywords in the prompt. Temporal keywords are matched from a fixed lexicon (e.g., *before, after, then, next, start, end, while, during, until, first, last, soon*) at the tokenizer level, and their attentions are averaged with the pre-answer query using equal weights. For each selected query, we sum attention mass over visual tokens within frame f and then average over heads to obtain $p(f) \in [0,1]$ with $\sum_f p(f)=1$. To reduce spurious spikes, we apply a within-query temperature rescale to the pre-softmax logits such that the entropy of $p(\cdot)$ falls within a fixed band; the rescale is used only for measurement and does not modify the model state.

The early-to-middle pathway centrality $c(f)$ is computed from inter-frame attention inflow across the early and middle groups. For each such layer and head, we form the $T\times T$ inter-frame matrix $A^{(\ell,h)}$ as above, but zero the diagonal to remove within-frame mass. The inflow to frame f is $I^{(\ell,h)}(f) = \sum_{u\neq f} A^{(\ell,h)}_{u,f}$. We aggregate across heads and layers by averaging inflows and then compute $c(f) = I(f)/\sum_g I(g) \in [0,1]$. This centrality measures how much temporal information other frames route into f along the early–middle fusion pathway. For backbones where attention tensors are not exposed by default, we rely on standard PyTorch hooks or framework-specific flags to enable attention logging; we do not modify parameters or control flow. All quantities $p(f)$, $c(f)$, and the segment map $\mathcal{S}$ are aligned on the same T frames and the same pass, ensuring that the SCR@α, AETP, and mitigation operations are computed consistently. The computations are linear in T for segment scores and near-linear for attention aggregation with small constants; memory overhead is dominated by temporary per-head statistics that are released before decoding continues. We verified that early/middle/late grouping by percentiles yields stable results across all six 7B-class backbones; alternative groupings (fixed layer IDs) lead to statistically similar centrality rankings and do not change any main conclusions.

4. Naturally Multi-segment Tasks

This section details how we handle task families that legitimately aggregate evidence across multiple segments, such as Counting, Navigation, and Repetition. Our goal is to prevent systematic over-flagging by the risk gate without weakening the CH-Risk signal for genuinely risky cases. The procedure is confined to the gating path; it does not alter the core metric definitions, the global threshold τ, model parameters, or the main results reporting.

Task family detection. We conservatively identify naturally multi-segment queries using two sources that are available without extra supervision: (i) dataset-provided category tags (where present) and (ii) a tokenizer-level lexicon match in the user/query text. The lexicon includes cues for multi-segment reasoning such as "how many times," "count," "number of," "repeat(ed)," "again," "navigate," "go to," "from ... to," "route," "path," "across clips," "over the video". A sample is treated as naturally multi-segment if either the category indicates one of these families or the lexicon match is positive with no conflicting negation (e.g., "do not count"). This AND/OR hybrid is intentionally conservative: if neither signal fires, the sample follows the default pipeline; if the signals disagree, the category tag wins. We found this rule stable across all datasets and six 7B backbones.

Prior-aware segment coverage. Let K_α be the minimal number of segments needed to cover α mass of the mid-layer text→frame support. For naturally multi-segment families we introduce a small, family-level prior $K_{\text{prior}} \in \{2,3\}$ determined on a development split, and we compute a prior-aware coverage cost

$$\tilde{K}_\alpha \;=\; \max\big(K_\alpha - K_{\text{prior}},\, 0\big), \tag{2}$$

$$\widetilde{\text{SCR}}@\alpha \;=\; \frac{\tilde{K}_\alpha}{\max(1,\, M - K_{\text{prior}})}. \tag{3}$$

This subtraction-and-renormalization preserves the scale and guarantees $0 \le \widetilde{\text{SCR}}@\alpha \le 1$. It is monotone in K_α and reduces to the original SCR@α when $K_{\text{prior}}=0$. Edge cases are safe by construction: if $M \le K_{\text{prior}}$ or $K_\alpha \le K_{\text{prior}}$, then $\widetilde{\text{SCR}}@\alpha=0$ and the gate cannot spuriously flag the sample solely due to legitimate multi-segment aggregation.

Table 1. **Ablation on Naturally Multi-segment Tasks.** We evaluate LLaVA-Video-7B to analyze gate behavior on queries annotated as inherently multi-segment. We report Accuracy (%), CH-Risk (lower is better), HighRisk@τ (%, τ=0.28), and the false-positive rate (FPR, %) of the risk gate. "+ Baseline correction" applies the light prior $K_{\text{prior}} \in \{2, 3\}$ when computing $\text{SCR@}\alpha$ on these task types.

Subset	Variant	ACC↑	CH-Risk↓	HighRisk@τ↓	FPR↓
Counting	w/o correction	66.8	0.34	39	22
	+ Baseline correction	**68.2**	**0.27**	**26**	**10**
Navigation	w/o correction	61.5	0.33	36	19
	+ Baseline correction	**63.0**	**0.26**	**24**	**9**
Repetition	w/o correction	64.1	0.35	38	21
	+ Baseline correction	**65.6**	**0.27**	**25**	**11**
Re-editing	w/o correction	62.7	0.32	34	18
	+ Baseline correction	**64.5**	**0.25**	**22**	**8**

Risk and gate. For these families we form a prior-aware risk

$$\widetilde{\text{CH-Risk}} \;=\; \widetilde{\text{SCR}}@\alpha \cdot \big(1 - \text{AETP}\big) \in [0, 1], \quad (4)$$

and apply the same global threshold τ for gating: enable sSAFR+RTC if and only if $\widetilde{\text{CH-Risk}} \geq \tau$. Outside these families, we use the unmodified CH-Risk. In the main results, we report the standard CH-Risk to keep the metric definition uniform; the prior-aware variant is used only to decide whether to intervene on naturally multi-segment queries. In the dedicated ablation, we additionally display the prior-aware values to analyze gate behavior on these subsets.

Choice of K_{prior} and robustness. We fix K_{prior} per family on a development split using a simple criterion: the smallest value in $\{2, 3\}$ that keeps the risk-gate false-positive rate below a target band (10–12%) while preserving recall of genuinely risky cases under time-shuffle perturbations. This coarse prior avoids overfitting to any dataset and transfers across models. We do not tune K_{prior} per model or benchmark. The correction is computationally negligible (constant-time update after computing K_α), does not change α, AETP, or τ, and leaves sSAFR/RTC unchanged.

Failure modes and safeguards. The prior never increases risk and cannot mask misaligned evidence: if AETP is low (stage mismatch), $\widetilde{\text{CH-Risk}}$ remains high even when $\tilde{K}_\alpha = 0$; if segment usage is excessive beyond what is legitimate for the family, $\tilde{K}_\alpha$ grows and the gate triggers. If the family detector abstains or is ambiguous, we revert to the standard pipeline. These safeguards, together with conservative calibration, ensure that the gate is selective rather than permissive, aligning with the empirical reductions in false positives and the accuracy gains reported in the ablation.

5. Ablation on Naturally Multi-segment Tasks

Table 1 evaluates the light baseline correction K_{prior} on LLaVA-Video-7B for naturally multi-segment tasks, where correct reasoning legitimately aggregates evidence across multiple segments (e.g., Counting, Navigation, Repetition, Re-editing). Without correction, the greedy nature of $\text{SCR@}\alpha$ over-penalizes such legitimate aggregation and inflates the baseline CH-Risk (reaching $0.32 \sim 0.35$, well above the global average of 0.28). This causes the gate to erroneously flag benign multi-evidence cases as "high risk," resulting in an elevated False Positive Rate (FPR). Introducing a small prior ($K_{\text{prior}} \in \{2, 3\}$) successfully recalibrates the metric. It reduces CH-Risk by $0.07 \sim 0.08$ on average and essentially halves the false-positive rate (e.g., Counting: FPR $22\% \to 10\%$; Navigation: $19\% \to 9\%$). Downstream accuracy improves by $+1.4\% \sim +1.8\%$, indicating that the risk gating becomes highly discriminative rather than overly sensitive, intervening only when necessary without disrupting normal multi-hop aggregation.

Justification for the Residual FPR. It is worth noting that the corrected FPR does not drop to absolute zero, settling around $8\% \sim 11\%$. This is an intended and desirable behavior rather than a flaw. Even in multi-segment tasks, if a model's text-to-frame support scatters excessively far beyond the legitimate prior K_{prior}, or if it exhibits severe stage mismatch (captured by a low AETP score), it strongly indicates a genuine Chimera Hallucination risk. The corrected CH-Risk correctly retains its sensitivity to these extreme outliers, proving that the prior K_{prior} effectively prevents systematic over-flagging without masking true risks.

6. Ablation on Intervention Layer

Table 2 examines the optimal location for applying CH-M intervention within the model architecture.

Table 2. **Ablation on Intervention Layer.** We report Accuracy (%) on NExT-QA and VidHalluc using LLaVA-Video-7B. Risk entries are deltas averaged over the two benchmarks (↓ desirable for ΔCH-Risk / ΔSCR@0.8 / ΔHighRisk@τ, ↑ desirable for ΔAETP). All variants use a single forward pass without finetuning. **Bold** indicates the optimal single-hook variant, and underline indicates the global optimum.

Variant	Accuracy (%)↑		Risk Deltas (avg)			
	NExT-QA	VidHalluc	ΔCH-Risk↓	ΔSCR@0.8↓	ΔAETP↑	ΔHighRisk@τ (%)↓
Baseline (no mitigation)	73.2	76.6	0.00	0.00	0.00	0
Early layer hook	74.1	78.2	−0.04	−0.03	+0.02	−8
Middle layer hook (Ours)	**75.1**	**80.2**	−0.07	−0.06	+0.05	−11
Late layer hook	74.4	79.1	−0.05	−0.04	+0.03	−9
Two-hook (Early+Middle)	75.2	80.4	−0.08	−0.06	+0.05	−12
Two-hook (Middle+Late)	75.0	80.0	−0.07	−0.06	+0.05	−11

Table 3. **Ablation on Temporal Perturbation Stress Test.** We report Accuracy (%) and risk statistics before/after CH-M (single pass, gate $\tau = 0.28$, $\alpha = 0.8$). HighRisk@ τ is the percentage of samples with CH-Risk $\geq \tau$.

Perturbation	Accuracy (%)↑		CH-Risk↓		HighRisk@τ (%)↓	
	w/o CH-M	w/ CH-M	w/o CH-M	w/ CH-M	w/o CH-M	w/ CH-M
None (original)	73.2	**75.1**	0.28	**0.21**	30	**20**
Frame reversal	62.0	**64.2**	0.38	**0.30**	48	**35**
Segment shuffle (light)	69.4	**71.2**	0.33	**0.25**	38	**27**
Segment shuffle (medium)	66.1	**68.2**	0.36	**0.28**	43	**31**
Segment shuffle (heavy)	62.8	**64.9**	0.39	**0.30**	49	**36**
Temporal downsample ($\times\frac{1}{2}$ fps)	68.0	**70.0**	0.34	**0.26**	41	**29**

The "Sweet Spot" of Single-Hook Mitigation. Among all single-hook variants, Middle-layer hooking achieves the best performance trade-off on both NExT-QA and VidHalluc. It yields the largest single-layer Accuracy gains and the strongest joint risk reduction (ΔCH-Risk↓, ΔSCR@0.8↓, ΔAETP↑). This strongly aligns with our mechanistic hypothesis: temporal pathways are formed in early–middle layers and integrated mid-stream. Aligning and consolidating attention exactly at this stage proves most effective. In contrast, Early hooks help but underperform, likely because the temporal causal structure is not yet fully assembled; Late hooks arrive after stronger language decoding priors begin to dominate, limiting the capacity for visual temporal consolidation.

Pushing the Limits with Two-Hook Variants. We additionally explore the potential of applying intervention across multiple layers. A lightweight two-hook variant (Early+Middle) marginally surpasses the default Middle-hook configuration, achieving the global optimum in both Accuracy (75.2% on NExT-QA) and absolute risk reduction (-0.08 ΔCH-Risk). While the two-hook configuration captures a slightly more robust temporal trajectory, the performance headroom it provides is relatively small compared to the substantial gains already achieved by the single Middle layer hook.

Design Choice Justification. We intentionally adopt the single Middle layer hook as our default CH-M setting in the main paper. It offers a near Pareto-optimal point, delivering $> 95\%$ of the maximum possible accuracy improvement and risk reduction, while minimizing the engineering complexity, transient memory footprint, and latency overhead associated with maintaining multiple hooks during inference. The Two-hook option is reserved as an advanced configuration for users seeking maximal hallucination suppression with less stringent latency constraints.

7. Ablation on Temporal Perturbation Stress Test

Table 3 probes whether CH-Risk reliably tracks temporal integrity and whether CH-M can recover degraded performance under counterfactual perturbations. We use LLaVA-Video-7B to ensure alignment with our main diagnostics. As perturbation severity increases (from "None" to heavy shuffling or reversal), Accuracy drops monotonically, while the baseline CH-Risk rises significantly from 0.28 to 0.39, and the HighRisk@ τ population expands from 30% to 49%. This demonstrates the strong sensitivity of our risk estimate to disrupted temporal structure. Applying CH-M

consistently mitigates these structural disruptions. It reduces the risk score by roughly -0.07 to -0.09 (absolute) and shrinks the high-risk population by 10–13%, while recovering a solid fraction of the lost Accuracy: $+1.9\%$ on the clean set, $+1.8$–$+2.1\%$ under shuffles, and $+2.2\%$ under frame reversal.

Mechanistic Insights under Extreme Disruption. A seemingly counterintuitive observation is that the Frame reversal case still benefits from CH-M, even though it completely destroys the global causal direction. Specifically, risk drops from $0.38 \rightarrow 0.30$ and accuracy improves from 62.0% to 64.2%. Rather than magically restoring the corrupted global causality, CH-M gracefully degrades into a local semantic consolidator. Under extreme structural collapse, standard attention tends to scatter uniformly across the randomized timeline, exacerbating severe hallucination. By enforcing segment-level routing (sSAFR) and residual token calibration (RTC), CH-M prevents this uniform scattering and forces the model to ground its reasoning on the most salient, internally coherent local event segments. This localized grounding rescues basic object and action recognition, providing a measured but reliable performance recovery, even though the complex multi-hop causal reasoning remains fundamentally impaired compared to the clean baseline.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

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

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[5] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024. 1

[6] Jianfeng Cai, Wengang Zhou, Zongmeng Zhang, Jiale Hong, Nianji Zhan, and Houqiang Li. Mitigating hallucination in videollms via temporal-aware activation engineering. *arXiv preprint arXiv:2505.12826*, 2025. 1

[7] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 239–249, 2025. 1

[8] Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21635–21644, 2025. 1

[9] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[12] Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *arXiv preprint arXiv:2507.19024*, 2025. 1

[13] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[15] Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*, 2025. 1

[16] 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 *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14303–14312, 2024. 1

[17] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 1

[18] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[20] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 1

[21] Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. From specific-MLLMs to omni-MLLMs: A survey on MLLMs aligned with multi-modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. 1

[22] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238, 2024. 1

[23] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, 2024. 1

[24] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 1

[25] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spatial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025. 1

[26] 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*, pages 5971–5984, 2024. 1

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[30] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160, 2024. 1

[31] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4 (7):2025, 2025. 1

[32] Ziqi Pang and Yu-Xiong Wang. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*, 2025. 1

[33] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. *arXiv preprint arXiv:2503.11579*, 2025. 1

[34] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1

[35] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[36] Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159, 2025. 1

[37] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[38] 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. *arXiv preprint arXiv:2409.12191*, 2024. 1

[39] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

[40] 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*, 2024. 1

[41] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. *arXiv preprint arXiv:2406.16449*, 2024. 1

[42] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025. 1

[43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1

[44] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14635–14645, 2025. 1

[45] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025. 1

[46] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1

[47] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer, 2024. 1

[48] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 1

[49] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1

[50] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1

[51] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024. 1

[52] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025. 1

[53] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1

[54] Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*, 2024. 1

[55] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 1

[56] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1