Video Finetuning Improves Reasoning Between Frames

Ruiqi Yang
Brown University
ruiqi_yang1@brown.edu

Tian Yun
Brown University
tian_yun@brown.edu

Zihan Wang
Brown University
zihan_wang3@brown.edu

Ellie Pavlick
Brown University
ellie_pavlick@brown.edu

Abstract

Multimodal large language models (LLMs) have made rapid progress in visual understanding, yet their extension from images to videos often reduces to a naive concatenation of frame tokens. In this work, we investigate what video finetuning brings to multimodal LLMs. We propose Visual Chain-of-Thought (vCoT), an explicit reasoning process that generates transitional event descriptions between consecutive frames. Using vCoT, we systematically compare image-only LVLMs with their video-finetuned counterparts, both with and without access to these transitional cues. Our experiments show that vCoT significantly improves the performance of image-only models on long-form video question answering, while yielding only marginal gains for video-finetuned models—suggesting that the latter already capture frame-to-frame transitions implicitly. Moreover, we find that video models transfer this temporal reasoning ability to purely static settings, outperforming image models' baselines on relational visual reasoning tasks.

1 Introduction

Multimodal large language models (LLMs) have reached remarkable progress in understanding visual contents through the integration of pretrained LLMs with pretrained image or video encoders. Image LLMs, such as [Li et al., 2023] and [Liu et al., 2023], have demonstrated strong capabilities on various downstream tasks, such as image captioning [Chen et al., 2015, Plummer et al., 2015], tabular data understanding [Chen et al., 2019], visual question answering [Bigham et al., 2010, Goyal et al., 2017, Hudson and Manning, 2019]. However, their naive extension to the video domain often involves frame-by-frame tokenization without true temporal understanding. Consequently, these models tend to rely on superficial visual cues and struggle when a task requires reasoning over implicit transitions in between multiple video frames.

In contrast, video LLMs are designed to reach better video understanding by additional finetuning on video data and more inductive biases, such as additional temporal positional encoding with RoPE [Bai et al., 2025, Hong et al., 2025]. Despite their architectural advantages, pretrained video LLMs often underperform on tasks requiring deep temporal reasoning unless explicitly fine-tuned on temporally grounded datasets [Gao et al., 2017, Liu et al., 2024b, Ren et al., 2024]. This raises a critical question: To what extent does video-based fine-tuning enhance the reasoning capabilities of LLMs beyond what image-based models can achieve?

In this work, we investigate whether the video finetuning of video LLMs improves their understanding and reasoning between frames. To approach this question, we systematically comparing video LLMs

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Interpreting Cognition in Deep Learning Models.

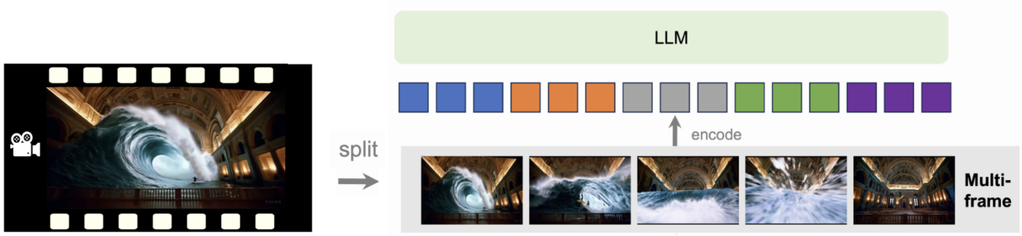


Figure 1: Video LLMs typically sample a number of frames from a video and digest them as a sequence of concatenated visual tokens [Zhang et al., 2024].

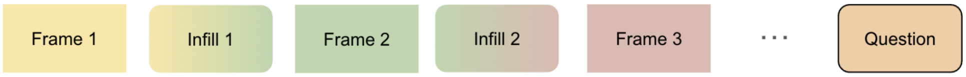


Figure 2: Visual CoT prompting pipeline. For each clip, transitional text infills are generated between every adjacent frame pair, interleaved with corresponding frames, and followed by the downstream question.

with their image LLM counterparts, which share the same architecture and image training data, and differ with the video LLMs mainly in terms of the absense of video finetuning. We introduce a visual Chain-of-Though (vCoT) to obtain explicit descriptions in between each pair of frames to measure models' performance with and without these descriptions on EgoSchema [Mangalam et al., 2023], which is a long-form video understanding benchmark and cannot be simply solved by observing static images. Specifically, we make the following contributions:

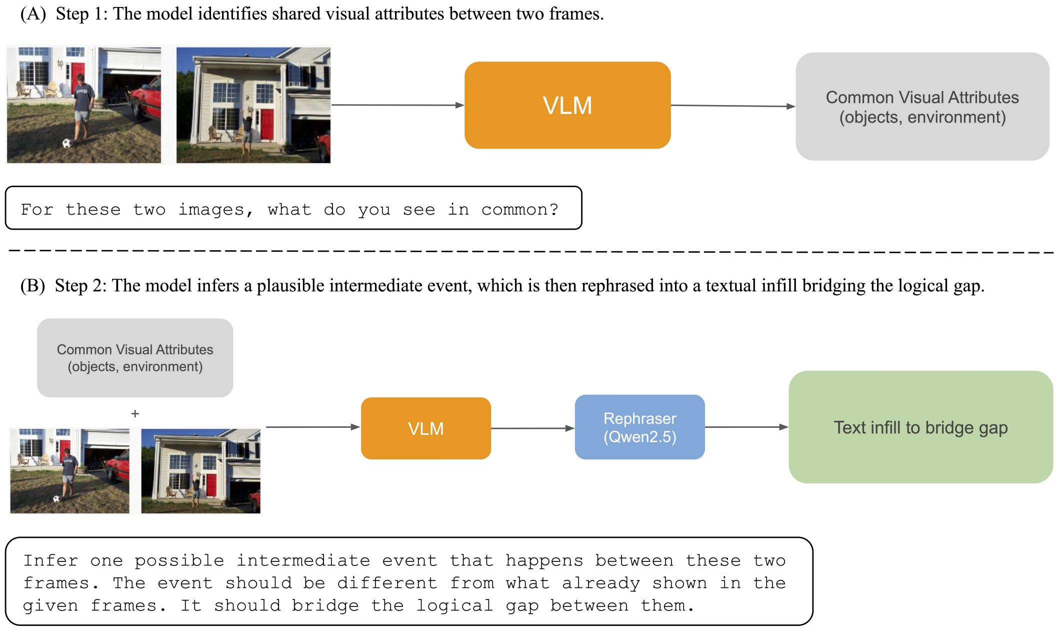
1. We consider pairs of image LLMs and video LLMs and measure models' performance with and without vCoT. We demonstrate that video LLMs gain marginal benefits with explicit in-between frame descriptions, while image LLMs gain substantial improvement, reflecting that video LLMs may learn to infer the frame transition implicitly.
2. We randomly sample videos from other samples and replace the video stream or text infill stream to examine the robustness to noise of video LLMs and image LLMs. We observe that video LLMs are more robust noise from either modality.
3. We study whether the reasoning capabilities between frames can be extended from video domain to static images. We focus on RAVEN [Zhang et al., 2019], and observe video LLMs outperform image LLMs on all sub-tasks in RAVEN, showing that video LLMs achieve better reasoning between not only the contiguous image frames but also the static images.

2 Method

To study whether video LLMs achieve stronger inter-frame reasoning than image-based counterparts, we introduce **Visual Chain-of-Thought (vCoT)** infills—explicit textual reasoning inserted between consecutive frames. We investigate whether these infills improve the performance of (i) an image-only vision–language model (VLM) and (ii) its video-finetuned counterpart. If adding vCoT substantially boosts accuracy, it suggests that the model benefits from explicit reasoning and lacks implicit temporal understanding. Conversely, minimal or negative gains imply that the model already reasons over frames without additional cues.

2.1 Visual Chain-of-Thought (vCoT) Infills

Motivated by chain-of-thought prompting in NLP [Wei et al., 2022], we introduce **visual Chain-of-Thought** (VCOT)—a two-step reasoning process that first captures the static context shared by two frames and then infers a plausible event connecting them. Each one-sentence *transitional caption* explicitly describes how the scene evolves between consecutive frames (F_i, F_{i+1}), serving as a *text*



(A) Step 1: The model identifies shared visual attributes between two frames.

VLM
Common Visual Attributes (objects, environment)

```
For these two images, what do you see in common?
```

(B) Step 2: The model infers a plausible intermediate event, which is then rephrased into a textual infill bridging the logical gap.

Common Visual Attributes (objects, environment)
+
VLM
Rephraser (Qwen2.5)
Text infill to bridge gap

```
Infer one possible intermediate event that happens between these two
frames. The event should be different from what already shown in the
given frames. It should bridge the logical gap between them.
```

Figure 3: Visual CoT for inter-frame reasoning. Step 1: The model identifies shared visual attributes between two frames. Step 2: It infers a plausible intermediate event, which is then rephrased into a textual infill bridging the logical gap.

Table 1: Controlled experiment results on the EgoSchema benchmark. (a) LLaVA models and (b) InternVL models evaluated with and without vCoT.

(a) LLaVA variants. Accuracy (%).

Model	**#F**	**Base**	**+vCoT**
LLaVA-NeXT	5	44.0	51.4 (+7.4)
LLaVA-NeXT-Video	5	47.0	48.6 (+1.6)
LLaVA-NeXT	10	49.2	55.4 (+6.2)
LLaVA-NeXT-Video	10	49.0	51.4 (+2.4)

(b) InternVL variants. Accuracy (%).

Model	**#F**	**Base**	**+vCoT**
InternVL-Image	5	38.4	40.4 (+2.0)
InternVL-Video	5	44.6	42.4 (–2.2)
InternVL-Image	10	37.4	42.6 (+5.2)
InternVL-Video	10	45.8	49.0 (+3.2)

infill interleaved with the original video sequence. These infills provide interpretable intermediate steps that make temporal reasoning explicit and enhance the model's understanding of frame-to-frame continuity. The vCoT generation process consists of two sequential prompts (Figure 3):

Step 1: Common Visual Attributes. We first query the model: *"For these two images, what do you see in common?"* This encourages the model to identify shared scene elements—such as objects, background, or spatial configuration—providing a stable context across frames.

Step 2: Bridging Event Inference. Next, using both frames and the identified context, we prompt: *"Infer one possible intermediate event that happens between these two frames. The event should be different from what is already shown and should bridge the logical gap between them."* This step yields a plausible, one-sentence description of the temporal transition (e.g., "the person kicks the ball toward the house"). To maintain clarity and conciseness, each event description is then rephrased using a lightweight Qwen-2.5 model [Yang et al., 2024].

For each video clip, the above process is applied iteratively to every consecutive frame pair, producing a chain of inferred events. The resulting interleaved sequence—frames and infills—is then combined with the task question (multiple-choice or open-ended) and passed to the VLM for final prediction.

Table 2: vCoT accuracy and degradation on the EgoSchema benchmark under two perturbation strategies: visual shuffle and text shuffle.

Model ID	#F	vCoT	visual shuffle	text shuffle
LLaVA-NeXT	5	51.4	39.8 (–11.6)	42.0 (–9.4)
LLaVA-NeXT-Video	5	48.6	41.6 (–7.0)	47.0 (–1.6)
LLaVA-NeXT	10	55.4	51.8 (–3.6)	45.0 (–10.4)
LLaVA-NeXT-Video	10	51.4	46.4 (–5.0)	45.4 (–6.0)

3 Results

We evaluate vCoT on two representative LVLMs, LLAVA-NEXT [Liu et al., 2024a] and INTERNVL [Chen et al., 2024b], along with their video-finetuned counterparts, LLAVA-NEXT-VIDEO [Zhang et al., 2024] and INTERNVL2 [Chen et al., 2024a]. Within each pair, the vision encoder, language backbone, and cross-modal projector remain identical; the only difference lies in whether video finetuning is applied. This controlled setup ensures that any observed performance variation can be attributed specifically to the effect of video finetuning.

Building on this evaluation design, we report three sets of findings. First, we conduct a controlled study on the EgoSchema long-form video-QA benchmark [Mangalam et al., 2023], highlighting the impact of vCoT under matched model capacity and data conditions. Second, to examine modality sensitivity, we introduce a shuffling diagnostic that deliberately introduces conflicting visual or textual evidence. Finally, we assess the transferability of vCoT to a different reasoning domain, the I-RAVEN logical-reasoning suite.

3.1 Controlled experiment on EgoSchema

EgoSchema subset comprises 500 three-minute egocentric videos paired with one multiple-choice question each. Following VLMEvalKit [Duan et al., 2024], we prompt the model to output an option index (“A–D”) and use QWEN2.5-7B-CHAT [Yang et al., 2024] as a judge. All prompts replicate the official VLMEvalKit template.

Table 1 shows LLaVA and InternVL models’ performance on EgoSchema with and without vCoT. Our experiments reveal that explicit temporal reasoning through vCoT consistently enhances image models’ performance across all tested configurations, with improvements ranging from $+1.6\%$ to $+7.4\%$ over the frame-only baseline. This underscores the importance of structured temporal understanding for long-form video comprehension.

Notably, the benefits of vCoT are most pronounced in models that lack prior video supervision. For instance, when using dense temporal sampling (#F=5), the image-only LLAVA-NEXT model sees a substantial performance gain of $+7.4\%$, compared to only $+1.6\%$ for its video-finetuned counterpart. A similar trend is observed across the InternVL variants. This discrepancy suggests that vCoT infills serve as a crucial supplementary reasoning mechanism for models not pretrained on video data, while models already exposed to video supervision may implicitly capture transitional dynamics, thereby reducing—or even negating—the marginal utility of vCoT.

3.2 Modality-shuffling ablations

To disentangle reliance on visual versus textual cues, we construct two perturbations (Figure 4): (1) *Visual shuffle*: replace every video frame with a frame from an unrelated clip while keeping the text infills intact. (2) *Text shuffle*: keep the original frames but swap the text infills with those from a different video.

Table 2 shows the results. Both image-based and video-based LLaVA models are sensitive to visual perturbations; however, the video variant exhibits significantly greater robustness to textual noise, suggesting a stronger reliance on the visual modality.

Table 3: Accuracy (%) on the i-RAVEN benchmark, comparing model performance across different reasoning rule types. The left block summarizes accuracy for position-based rules (object alignment and spatial distribution), while the right block evaluates relational and directional rules (object interactions and spatial reasoning).

Model ID	**center**	**dist_4**	**dist_9**	**in/out**	**indist4/out**	**L/R**	**U/D**	**Avg.**
InternVL-Image	14.8	14.4	15.2	11.6	13.2	**15.2**	**14.4**	14.1
InternVL-Video	**15.6**	**16.0**	**15.8**	**13.8**	**17.0**	14.0	14.2	**15.2**
LLaVA-Image	7.0	8.0	15.0	7.0	9.0	12.0	14.0	10.3
LLaVA-Video	7.0	**14.0**	**16.0**	**8.0**	**13.0**	**14.0**	**21.0**	**13.3**

3.3 Transferability to Relational Visual Reasoning

A natural question is whether the ability to reason across video frames also transfers to reasoning across static image frames. To examine this, we turn to the i-Raven benchmark Hu et al. [2021], a relational visual reasoning suite derived from Progressive Matrices. In this task, models must infer abstract visual rules from a set of panels and apply them to identify the correct completion, making it a challenging test of non-temporal relational inference.

Despite the absence of explicit temporal signals in the task, models finetuned on video data consistently outperform their image-only counterparts, indicating that temporal reasoning can transfer to relational reasoning. Specifically, video-finetuned InternVL and LLaVA models yield overall gains of $+1.1\%$ and $+3.0\%$, respectively, compared to their non-video baselines. The most notable improvements occur in categories involving spatial layouts and relative positioning. For example, InternVL shows a $+3.8\%$ gain on the `indist4/out` rule, while LLaVA achieves a $+7.0\%$ improvement on `U/D`. These results suggest that video finetuning may induce a stronger inductive bias toward relational structures, enabling models to better capture abstract spatial dependencies even in tasks devoid of temporal context.

4 Conclusion

In this paper, we investigate the unique benefits that video finetuning brings to multimodal LLMs through a controlled comparison between image-based models and their video-finetuned counterparts. To this end, we introduce Visual Chain-of-Thought (vCoT), which generates explicit textual infills between consecutive frames to represent intermediate reasoning steps. Through controlled experiments between image LLMs and their video counterparts, we demonstrate that video-finetuned models already perform implicit inter-frame reasoning, and that this capability naturally transfers to static visual relational reasoning tasks.

5 Acknowledgement

We would like to thank all reviewers and the area chair for their valuable feedback. We would like to thank Zitian Tang, Shijie Wang, and other members of the SuperLab at Brown University for their discussions and insights. The project depicted is sponsored in part by a Young Faculty Award from the Defense Advanced Research Projects Agency, Grant #D24AP00261. The content of the information does not necessarily reflect the position, or the policy of the government and no official endorsement of this work should be inferred.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024b.

Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2024. URL https://sharegpt4o.github.io/.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

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*, pages 6904–6913, 2017.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1567–1574, 2021.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

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

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? *arXiv preprint arXiv:2403.00476*, 2024b.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

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

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

A Appendix

A.1 Modality shuffling demonstration

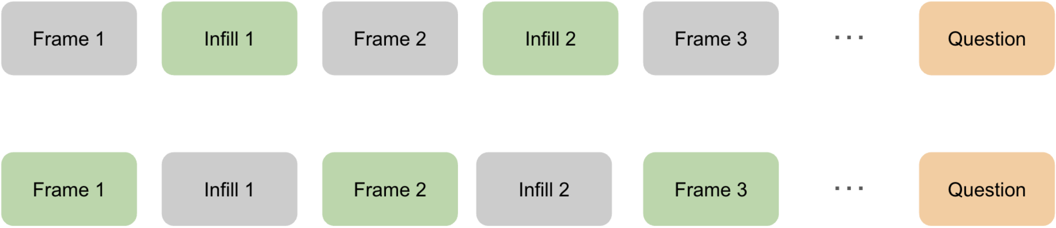


Figure 4: Shuffling to create conflicting modalities. Grey = retained modality; green = shuffled modality. Top row = text shuffle; bottom row = visual shuffle.

A.2 Frame captions as text infills

To distinguish vCoT infills from simple frame captions, we evaluate LLaVA models on EgoSchema using image captions as inter-frame text. Results show that while captions provide some benefit, vCoT infills consistently yield greater improvements, offering stronger guidance for temporal reasoning (see Table 4).

Table 4: Accuracy (%) with frame captions as vCoT on LLaVA variants.

Model ID	**#F**	**Baseline**	**Captions**	**vCoT**
LLaVA-NeXT	5	44.0	48.2	51.4 (+3.2)
LLaVA-NeXT-Video	5	47.0	45.0	48.6 (+3.6)
LLaVA-NeXT	10	49.2	54.6	55.4 (+0.8)
LLaVA-NeXT-Video	10	49.0	50.0	51.4 (+1.4)

A.3 LoRA fine-tuning to neutralise data-scale effects.

The video-finetuned variant of LLaVA-NeXT is trained on approximately 100k additional video-instruction pairs beyond the 600k image prompts used for the base model [Zhang et al., 2024]. To disentangle the effect of data scaling, we perform parameter-efficient LoRA fine-tuning on image LLaVA-NeXT. Specifically, we train rank-128 LoRA [Hu et al., 2022] adapters on 100k high-quality image instructions from `ShareGPT-4V` [Cui et al., 2024] and `ShareGPT-4o`, so that the total amount of training data matches that of the video variant.

Table 5: Accuracy (%) on the EgoSchema subset (n=500) *after* LoRA fine-tuning the image-only backbone.

Model ID	**#F**	**Stride**	**Baseline**	**vCoT**
LLaVA-NeXT (LoRA)	5	1	45.8	45.0 (−0.8)
LLaVA-NeXT-Video	5	1	47.0	48.6 (+1.6)
LLaVA-NeXT (LoRA)	5	2	36.2	37.4 (+1.2)
LLaVA-NeXT-Video	5	2	36.8	42.2 (+5.4)
LLaVA-NeXT (LoRA)	10	1	49.2	55.3 (+6.1)
LLaVA-NeXT-Video	10	1	49.0	51.4 (+2.4)
LLaVA-NeXT (LoRA)	10	2	39.8	43.0 (+3.2)
LLaVA-NeXT-Video	10	2	39.8	45.4 (+5.6)

After LoRA tuning, the earlier trend—where vCoT consistently enhanced performance, particularly for image-only models—begins to shift significantly (see Table 5). In particular, with dense temporal sampling (#F=5, stride=1), the vCoT module no longer provides gains for the image-only model—in fact, performance drops by 0.8%. This reversal suggests that the earlier advantage of vCoT in image-based models, relative to video-finetuned models, may have been driven by differences in training data scale, rather than by an inherent deficiency in temporal understanding. Thus, the emergence of implicit temporal reasoning might not be exclusive to video finetuning.

Additionally, a second phenomenon emerges: potential degradation in textual modeling. Prior work has shown that EgoSchema questions often rely on shortcut textual cues, such as those derived from image captions [Zhang et al., 2023, Wang et al., 2024]. LoRA adaptation on 100k image-instruction pairs may induce forgetting, weakening the model's capacity to reason with such cues. This degradation in textual understanding likely contributes to the reduced effectiveness of vCoT-generated text infills in post-LoRA settings.

A.4 Limitations

The main limitation of this project is the inability to train control models from scratch. Ideally, we would use a small LLM (e.g., 0.5B parameters) and train two variants: one on image data and the other on video data. This setup would allow us to fully control the training pipeline and ablate the effects of various inductive biases—such as the use of temporal encoders, temporal positional embeddings, and other architectural components.

Additionally, greater care must be taken in selecting which inductive biases to control in such experiments. Interestingly, our results show that LLaVA-NeXT often outperforms LLaVA-NeXT-Video, which is unexpected and suggests that the latter may not serve as a strong representative of video LLMs due to its limited performance. We argue that all inductive biases—aside from those inherent to the base LLM architecture—should be incorporated if they enhance the model's ability to learn from image or video data. This approach would allow us to obtain upper-bound variants of both image- and video-trained LLMs starting from the same initialization, enabling a fair and meaningful comparison.