Appendix for ACT2SEE: Emergent Active Visual Perception for Video Reasoning

A. Impact statement

The ACT2SEE framework significantly advances complex video reasoning by empowering models with active visual perception, enabling them to dynamically retrieve or synthesize visual information. However, we acknowledge the potential negative societal impacts associated with these capabilities. A significant concern arises from the framework's ability to generate novel visual frames to illustrate hypothetical or counterfactual scenarios. This generative function could be exploited by malicious actors to synthesize convincing deceptive media (deepfakes) or fabricate visual "evidence", thereby facilitating fraud, harassment, or the manipulation of public discourse. Furthermore, the enhanced efficiency and granularity of video analysis, particularly the active retrieval mechanism, could be misused to develop or extend harmful forms of surveillance. This risks intensifying the monitoring and profiling of individuals without consent, raising substantial privacy and human rights concerns. Additionally, the framework may inherit and perpetuate societal biases present in the underlying foundational models used for data generation and frame synthesis. To mitigate these risks, we strongly advocate for the development of robust forensic tools and digital watermarking to ensure the traceability of synthesized content. We also recommend implementing strict ethical guidelines for deployment, including rigorous bias evaluations and potentially gated access to the generative components, to prevent misuse in sensitive domains.

B. SFT synthesis prompt

Below we include the full prompts for round 1 (Table 1) and round 2 (Table 2) each with an example included. Round 1 prompts Gemini [4] to output reasoning traces with retrieval or generation requests along with the corresponding queries. Then the query will be used to retrieve or generate with offline retrieval and generation models. The text after `</retrieve>` or `</generate>` will be removed, the retrieved or generated frames will be added, and then altogether these will be input for the second round. The second round is to let Gemini finish the reasoning and output the answer based on the extra visual information. No ground truth answer or reasoning is included to encourage diversity of the response.

C. Dataset

Below we include more details of the three datasets we used to create the SFT data:

MINERVA (Multimodal INterpretablE Reasoning Video Annotations) [14] is a challenging benchmark designed to rigorously evaluate the intermediate reasoning capabilities of multimodal models. Traditional video benchmarks often rely solely on outcome supervision, obscuring whether models genuinely integrate perceptual and temporal information or exploit biases. MINERVA addresses this by providing 1,515 complex, hand-crafted, multi-step questions across diverse video domains and lengths (up to 100 minutes), each accompanied by detailed, manually annotated reasoning traces. These traces facilitate interpretable assessment beyond final answer accuracy by outlining the necessary localization, perception, and logical steps required for the solution. Extensive evaluation demonstrates that MINERVA poses a significant challenge to frontier models (best performance 66.2% vs. 92.5% human accuracy), and an accompanying taxonomy of errors reveals that temporal localization and visual perception remain primary failure modes. Type of the videos include Short Films, Sports and Board Games, Educational, and Lifestyle. Type of skills required for the QA include Temporal Reasoning, Counting, Cause and Effect, Goal Reasoning, Situational Awareness, Event Occurrence, State Changes, Reading (OCR), Listening (identifying a detail in the audio track), Spatial Perception, Numerical Reasoning (all math operations other than counting), Object Recognition, Counterfactual Reasoning ('what if', but with an objective outcome).

Social Genome [13] is the first benchmark designed to evaluate the fine-grained, grounded social reasoning capabilities of multimodal models. The dataset comprises 272 videos of real-world social interactions, accompanied by 1,486 human-annotated reasoning traces totaling 5,777 steps. Each reasoning step is explicitly grounded in multimodal evidence, referencing visual, verbal, and vocal cues extracted from the interactions. Uniquely, SOCIAL GENOME is the first benchmark to systematically incorporate external knowledge—contextual information not

Provide a step-by-step answer for the question below. You will be provided a set of initial video frames. Your task is to generate the reasoning traces, which include texts and images, to correctly answer the question. You must conduct reasoning inside <think> and </think>. When reasoning, if you need any visual knowledge, you can call a frame retrieval module by <retrieve> *retrieval prompt* </retrieve>, or a frame generation module by <generate> *generation prompt* </generate>, and either will return the video frame between <frame> and </frame>. The retrieval or generation prompt is the text description of the frame you want to retrieve/generate and you should provide it with specificity and conciseness about the exact content of the image, instead of timestamps (e.g., 2 seconds) in the query. Please do not include any timestamps in the query. Example query: "*a bison baby chased by a wolf pack*". Try calling the module whenever needed. Finally, give the answer in free-form text in the end, with <answer> and </answer> tags.
Please following the format below:
<think>
reasoning steps
<retrieve> *retrieval prompt* </retrieve> (or <generate> *generation prompt* </generate>)
reasoning step
</think>
<answer> *Answer to the question in free-form text.* </answer>
Question: "What is directly to the left of the black phone booth at the station where the train stops in the evening?"
Answers: "0": "A red mailbox.",
"1": "A couple of steps.",
"2": "A few white vans.",
"3": "A telephone pole.",
"4": "A display of flowers.",
Video: [video file]

Table 1. An example of Gemini input to collect SFT data for ACT2SEE in the first round. This is to get the start of the reasoning along with the retrieval or generation queries. Anything after the retrieval prompt or generation prompt will be discarded.

present in the video stimuli—which accounts for 51% of the reasoning steps. This dense annotation structure, featuring over $11,000$ entities and $2,900$ external knowledge observations, facilitates a holistic evaluation of the semantic and structural validity of model-generated social reasoning traces.

CausalVQA [6] is a benchmark dataset designed to evaluate multimodal models' capacity for physically grounded causal reasoning in real-world scenarios. Existing Video Question Answering (VQA) benchmarks typically focus on surface-level perception or rely on narrow synthetic simulations; CausalVQA bridges this gap by utilizing egocentric videos sourced from the EgoExo4D dataset. The benchmark comprises $1,586$ items (793 paired questions) across five categories designed to probe causal understanding: counterfactual, hypothetical, anticipation, planning, and descriptive. To ensure robustness, the dataset was curated using a rigorous hybrid human-and-model pipeline specifically engineered to enforce visual grounding and mitigate reliance on linguistic shortcuts, incorporating mechanisms such as immunization against "blind" LLMs and the use of paired, perturbed distractor sets. Baseline evaluations reveal a substantial gap between state-of-the-art models (61.66%) and human performance (84.78%), particularly on anticipation and hypothetical questions, underscoring the challenge of applying spatial-temporal reasoning and physical principles in complex, real-world settings. The types of questions in CausalVQA include Counterfactual, Hypothetical, Anticipation, Planning, and Descriptive.

C.1. Details of the retrieval and generation tool

Below we provide the brief introductions to the retrieval and generation tools we use. In ACT2SEE, we simply use the original video as the input, and use the retrieval model to retrieve a series of relevant frames based on the retrieval or generation query provided by Gemini. For retrieval requests, we use the middle frame of retrieved frames. For generation requests, we use the middle frame of retrieved frames as conditional, and use the generation query to generate the frame.

Provide a step by step answer for the question below. You will be given a half-finished reasoning steps you provided last round, which asked for new additional frames. You will then be provided the set of initial frames and the additional frames you asked for. Your task is to generate the rest of the reasoning traces in text correctly answer the question. Do NOT include any <retrieve> ... </retrieve> or <generate> ... </generate> this time, as these modules are not available now. You must finish the reasoning started from the given half-finished response and finish the reasoning with </think>. Start the reasoning directly without <think>. Finally, give the answer in free-form text in the end, with <answer> and </answer> tags. Please following the format below:
reasoning step
</think>
<answer> *Answer to the question in free-form text.* </answer>
Question: "What is directly to the left of the black phone booth at the station where the train stops in the evening?"
Answers: "0": "A red mailbox.",
"1": "A couple of steps.",
"2": "A few white vans.",
"3": "A telephone pole.",
"4": "A display of flowers.",
Half-completed reasoning:
Let me answer this question. <think> The question asks what is directly to the left of the black phone booth where the train stops in the evening. The video is a vlog based on a trip of the train, where the train arrives when it is dark. To know the answer, I need to retrieve the visual details when the train stops in the evening. <retrieve> train stopped at station platform in the evening with black phone booth </retrieve>
<frame> [retrieved video frame] </frame>

Table 2. An example of Gemini input to collect SFT data for ACT2SEE in the second round. This is to complete the reasoning from the first round.

TFVTG (Training-Free Video Temporal Grounding) is an approach that synergizes the reasoning capabilities of Large Language Models (LLMs) with the alignment strengths of Vision-Language Models (VLMs), eliminating the reliance on annotated training data. To handle complex, multi-event queries, TFVTG first employs an LLM (BLIP-2 [12]) to decompose the query into constituent sub-events and infer their temporal order and relationships. Crucially, to overcome the tendency of VLMs to overlook dynamic transitions, TFVTG propose a novel localization mechanism that explicitly models both the dynamic transition and static status phases of an event. This mechanism utilizes distinct dynamic and static scoring functions to measure the rate of similarity change and comparative relevance, respectively. Finally, the localized proposals for each sub-event are filtered and integrated based on the LLM-derived temporal constraints. This framework ensures more comprehensive event boundary localization and demonstrates superior generalization capabilities across diverse datasets and out-of-distribution scenarios.

Stable Diffusion 3.5 is an open family of latent diffusion models that introduce architectural and training refinements to improve prompt adherence, image quality, and controllability while remaining efficient on consumer-class hardware. The models incorporate Query-Key Normalization within transformer blocks and an enhanced MMDiT-X backbone, which together stabilize training and make the base models more amenable to downstream fine-tuning and multi-resolution generation. The model emphasizes customizability and diversity of outputs, deliberately allowing higher intra-prompt variability to preserve a broader style and knowledge distribution in the base models.

Distilled and medium-capacity variants extend the design to latency- and resource-constrained settings, targeting competitive text–image alignment and visual fidelity within a more accessible computational and licensing regime.

C.2. Data quality control and filtering

For all the generated CoTs, we perform quality check below. In the first step, we filter out all CoTs that lead to wrong answers, and re-generate the CoTs from the corresponding input video-QA pairs until the correct answers are produced. We cap the re-generation to two times. To make sure that the generated CoTs are high-quality, we measure the text embedding similarities of the generated interleaved

CoTs with the ground-truth CoTs provided by MINERVA, CausalVQA and Social Genome. Specifically, we compare the generated CoTs with ground-truths and only retain them if the BGE M3-Embedding [3] similarity is greater than 80%, following the practice from Han et al. [10]. Lastly, we perform a format check to ensure all the CoTs contain necessary thinking and answering tokens, and remove CoTs without the thinking tokens or answer tokens. Lastly, we perform a manual check of 100 to validate the CoT quality, where each CoT is inspected by checking whether the retrieval and generation formats are followed, if the reasoning in the CoT and the answer are consistent and coherent, and if the automatic format check and answer check is correct. All 100 samples pass the manual checks.

C.3. Metadata

After quality control and filtering, we have $3,373$ samples in the SFT dataset. There are $1,334$ from CausalVQA, $1,307$ from MINERVA, and 732 from Social Genome. To be specific, there are 64 samples of Counterfactual, 50 of Hypothetical, 152 of Anticipation, 106 of Planning, and 962 of Descriptive from CausalVQA, and 418 of Sports, 112 of STEM, 418 of Short Films, 359 of Lifesyle from MINERVA.

D. Experimental details

Below we include additional details of the experiments including benchmarks and baselines.

D.1. Benchmark details

Below we provide brief introductions to the benchmarking datasets we used in this paper.

Video-MME [7] is a comprehensive video benchmark designed to assess MLLMs in video analysis. It comprises 900 manually curated videos totaling 254 hours and $2,700$ multiple-choice questions. The dataset contains diverse content spanning 6 visual domains and 30 fine-grained categories, with video durations ranging widely from 11 seconds to 1 hour.

VideoEspresso [10] is a large-scale dataset focused on fine-grained video reasoning via Chain-of-Thought (CoT). Constructed using an automatic pipeline, it features VideoQA pairs enriched with multimodal CoT annotations, including intermediate reasoning steps, spatial bounding boxes, and temporal grounding. The benchmark evaluates models across 14 diverse tasks, emphasizing complex semantic reasoning while preserving spatial details and temporal coherence.

Egonormia [17] is a benchmark designed to evaluate the understanding of physical-social norms (PSNs) in Vision-Language Models. It comprises $1,853$ multiple-choice questions grounded in $1,077$ unique egocentric videos sourced from Ego4D. The benchmark spans seven norm categories (e.g., safety, privacy, proxemics, politeness, cooperation, coordination/proactivity, and communication/legibility) and focuses on evaluating normative decision-making, particularly in situations where norms conflict.

VCR-Bench [15] is a comprehensive evaluation framework dedicated to Video Chain-of-Thought (CoT) Reasoning. It consists of 859 videos and 1,034 question-answer pairs (including both multiple-choice and open-ended formats), featuring 4,078 manually annotated reference reasoning steps. The benchmark assesses the entire reasoning process across seven distinct task dimensions, distinguishing between models' perception and logical reasoning capabilities.

ViTIB [20], the Video-Text Interleaved Benchmark, is constructed to support the Video-Text Interleaved CoT (ViTCoT) paradigm. Sourced from VideoEspresso, it contains $1,382$ videos across 14 categories. The benchmark features $5,051$ MLLM-selected and manually verified keyframes (averaging 3.7 per key-video), designed to be integrated directly into the reasoning process to facilitate intuitive video comprehension. This is the benchmark that directly assess the effectiveness and generalizability of the video-text interleaved CoT paradigm, the key focus of this paper.

D.2. Baseline details

Below we provide brief introductions of the baselines we used in this paper.

Qwen2.5-VL-7B [1] integrates a dynamic-resolution Vision Transformer with the Qwen2.5 language model via an MLP-based merger to facilitate efficient token compression. The vision encoder utilizes 2D/3D patching, window attention, and SwiGLU to process inputs at near-native resolution while utilizing absolute image coordinates for precise spatial grounding. Temporal reasoning is achieved through Multimodal Rotary Position Embeddings (M-RoPE) aligned to absolute time, enabling robust handling of variable resolutions and video lengths. Finally, a staged training regimen involving large-scale multimodal pre-training and preference optimization yields advanced capabilities in document parsing, object grounding, and agentic interaction.

Intern 2.5-8B [1] implements an optimized scaling methodology within a ViT-MLP-LLM framework, utilizing a Progressive Scaling Strategy to align large-scale vision encoders with language models via incremental learning. To efficiently manage diverse inputs such as multi-image and video data, the architecture incorporates dynamic high-resolution processing and multimodal data packing. Furthermore, a rigorous data filtering pipeline is employed to mitigate noise and repetition, thereby enhancing training stability and the efficacy of test-time scaling techniques like CoT reasoning.

Video-LLaMA3-7B [18] establishes a vision-centric multimodal foundation by extending robust image understanding to the video domain through a four-stage training pipeline that leverages large-scale, re-captioned image–text corpora and curated temporal data. The architecture utilizes Any-resolution Vision Tokenization with a ViT-based encoder and 2D RoPE to process inputs at arbitrary resolutions and aspect ratios, ensuring the preservation of fine-grained spatial details. To optimize temporal efficiency, a Differential Frame Pruner compresses video content by eliminating redundant patches between consecutive frames, yielding compact and informative tokens. This unified framework effectively handles diverse inputs, including documents, charts, and both short and long videos, within a single instruction-tuned model.

Qwen3-VL-8B [16] integrates a Vision Transformer-based visual encoder with a Qwen3 text backbone, forming a unified autoregressive transformer designed to process interleaved text, images, and videos. The architecture employs Interleaved-MRoPE for spatio-temporal positional encoding, DeepStack fusion for multi-level feature integration, and text–timestamp alignment to facilitate precise temporal grounding and long-context reasoning. Following large-scale multimodal pretraining, instruction tuning yields both dense and Mixture-of-Experts (MoE) variants, offering specialized modes for instruction following and complex reasoning. This design ensures robust performance across diverse tasks, including multilingual OCR, complex visual question answering, and agentic GUI control, while supporting scalable deployment from edge to cloud environments.

Video-R1 [5] adapts rule-based reinforcement learning to video reasoning tasks by combining a temporal-aware optimization algorithm with mixed image–video training data. Central to this approach is Temporal Group Relative Policy Optimization (T-GRPO), which isolates and rewards explicit temporal reasoning by contrasting model performance on ordered versus shuffled video frames. The training protocol proceeds through supervised fine-tuning on a Chain-of-Thought dataset followed by reinforcement learning on a verifiable-answer dataset, employing a length-based reward to balance reasoning depth and succinctness. This methodology results in a model exhibiting enhanced temporal understanding and robust generalization across various video understanding benchmarks.

Chain-of-Shot [11] serves as a training-free, test-time mechanism for video MLLMs that enhances long-video comprehension through adaptive visual input optimization relative to a specific query. The method employs a binary video summarization technique wherein mosaiced shots are evaluated by a general MLLM to generate binary codes that pseudo-temporally ground task-relevant segments. Subsequently, balanced sequences of positive and negative sub-shots are processed alongside original shots via a co-reasoning module that dynamically reweights token logits based on the sparsity of relevant content. This design effectively concentrates computation on informative features while suppressing noise, enabling robust reasoning across varying video lengths without requiring architectural modifications.

FrameMind [8] is proposed as an end-to-end reinforcement learning framework that facilitates dynamic video reasoning by interleaving textual generation with active visual perception. Through a Frame-Interleaved Chain-of-Thought (FiCOT), the model iteratively detects informational deficits and actively queries targeted visual inputs, such as high-resolution frames, rather than processing static sequences. This adaptive capability is supported by Dynamic Resolution Frame Sampling (DRFS) and optimized via DRFS-GRPO, a group-relative policy optimization algorithm that derives sampling strategies from sparse rewards without frame-level supervision. Ultimately, this methodology enables the flexible balancing of temporal coverage and spatial detail to achieve efficient video understanding.

ViTCoT [20] ViTCoT introduces a Video-Text Interleaved Chain-of-Thought (CoT) paradigm in which a multimodal LLM first generates an initial textual reasoning trace from the original video, question, and options, and then refines this reasoning by interleaving it with a key-video composed of task-relevant frames. The key-video is obtained by automatic key-frame selection with a powerful MLLM and subsequent multi-annotator human verification. This two-stage prompting scheme is model-agnostic and can be plugged into various CoT variants (e.g., standard CoT, Desp-CoT, Plan-and-Solve), yielding more human-like, visually grounded reasoning, better exploitation of critical temporal cues, and richer neuron activation patterns in complex video understanding scenarios.

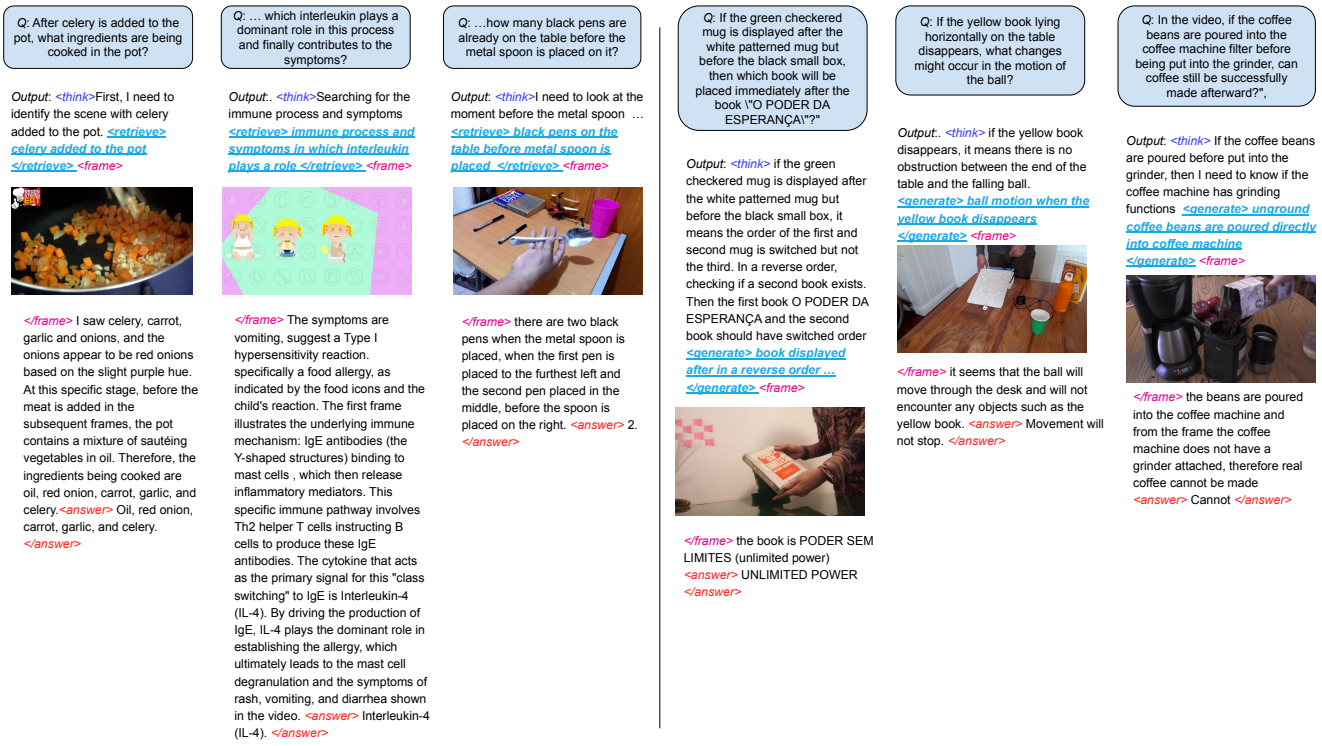

Figure 1. Qualitative examples (full CoT) showing the emergent capability of ACT2SEE outputting retrieved (left half) or generated frames (right) inside the CoTs at inference time, dynamically adding requested visual information (factual original videos via retrieval or counterfactual synthesis via conditional image generation) to the reasoning process.

Chain-of-Frames [9] advances video understanding in Multimodal Large Language Models (LLMs) by introducing temporally grounded, step-by-step reasoning. This approach involves fine-tuning video LLMs on COF-DATA, a large-scale dataset of diverse, frame-aware reasoning traces generated efficiently from both real and synthetic videos. By explicitly referencing relevant frames within the reasoning process, CoF integrates temporal information directly into the chain-of-thought structure. This method is self-contained, eliminating the need for the auxiliary networks or complex inference frameworks required by existing approaches. Consequently, CoF enhances model interpretability through explicit temporal grounding and significantly improves performance across various video understanding tasks while reducing hallucinations.

ReWatch-R1 [19] advances complex video reasoning in Large Vision-Language Models through an integrated approach combining agentic data synthesis and process-oriented reinforcement learning. It introduces a multi-stage pipeline to generate the ReWatch dataset, featuring temporally dense captions, challenging multi-hop questions, and video-grounded CoT data. A core innovation is the use of a Multi-Agent ReAct framework for CoT synthesis, which simulates iterative information retrieval and verification against video content. The model is subsequently trained via Supervised Fine-Tuning and Reinforcement Learning with Verifiable Reward (RLVR), incorporating a novel Observation and Reasoning (O&R) reward. This mechanism evaluates both the final answer accuracy and the factual grounding of intermediate reasoning steps, thereby explicitly penalizing hallucinations and enhancing the model's capacity for verifiable temporal reasoning.

E. Emergent full example

We include the full CoTs of Figure 4 in the main text in Figure 1. As the examples show, emergent capabilities of actively requesting to retrieve or generate frames is enabled by ACT2SEE, and demonstrating that the reasoning process is enhanced by the additional visual information.

F. Further Analyses

Below we include further analyses including latency and FLOPs, when image generation will help reasoning, categorization of image generation failure inside CoTs and their impact, and model's robustness towards generation failure.

Latency and FLOPs. We report latency (excluding data-loading) and FLOPs (counting all GPU operations, includ-

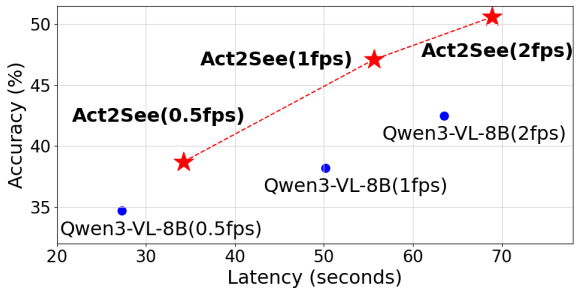

Figure 2. ACT2SEE pushes the Pareto frontier beyond Qwen3-VL-3B baseline by achieving higher accuracy with lower latency.

ing retrieval and generation) amortized over the full VCR-Bench, and varying the number of frames to the VLMs. Figure 2 and Table 3 show that ACT2SEE achieves higher accuracy, while having lower latency and lower FLOPs than the Qwen3-VL-8B baseline when using half as many frames, demonstrating Pareto dominance.

Table 3. ACT2SEE also pushes the Pareto frontier on FLOPs.

Method	Avg. input frames	Amortized TFLOPs	Accuracy (%)
Qwen3-VL-8B-Thinking	1fps (210 frames)	1025	38.2
ACT2SEE	0.5fps (106 frames)	**877**	**38.7**

When will generation help. We analyze this using counterfactual questions in VCR-Bench. Because VCR-Bench does not provide ground-truth counterfactual labels, we use GPT-5.2 to classify the questions, identifying 107 as counterfactual questions. We then split VCR-Bench into a counterfactual set and a factual set. Table 4 shows that: 1) Generation significantly outperforms retrieval-only and no-retrieval-or-generation on counterfactual questions (29% and 58% relative gains); and 2) Combining retrieval and generation yields the best overall performance.

Table 4. Accuracy (%) by question type, showing that generation significantly boosts performance on counterfactual questions, and combining retrieval and generation yields best performance.

CoT type	**Counterfactual set**	**Factual set**	**Full set**
No-retrieval-or-generation	35.52	38.51	38.20
Retrieval-only	43.41	45.80	45.55
Generation-only	56.01	37.70	39.60
Retrieval + Generation (ACT2SEE)	**56.22**	**46.05**	**47.10**

Categorize when generation fails. Following Borji [2], we categorize failures as: geometry (disproportionate size or shape), physics (physically unrealistic), and semantic (wrong spatial relationship, attributes, and composition). Among 127 generation calls on VCR-Bench, we found 13 failures (10.2%): 7 semantic, 4 physics, and 2 geometry. Among the CoTs with the generation failures, semantic failures yield lowest accuracy (28.6%) and predominantly affect counterfactual questions (4 out of 5 cases).

Robustness towards generation failure. Due to semantic failures being the most common, we study robustness towards generation failure by injecting deliberate semantic failures to visual generation prompts (e.g., word flips like "red ball"→"yellow ball") at 10%, 20%, 30% rates on VCR-Bench. From Table 5, ACT2SEE remains robust and still outperforms retrieval-only at 30% failure rate. We argue ACT2SEE is robust to realistic generation failure as we observe a 10.2% real-world failure rate the analysis above.

Table 5. Accuracy with varying generation failure rates.

Accuracy (%)	**10% failure**	**20% failure**	**30% failure**	**Retrieval-only**
ACT2SEE	46.88	46.40	45.69	45.56

References

[1] 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. 4, 5

[2] Ali Borji. Qualitative failures of image generation models and their application in detecting deepfakes. *Image and Vision Computing*, 137:104771, 2023. 7

[3] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. 4

[4] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1

[5] 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. 5

[6] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 2

[7] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 4

[8] Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Famemind: Frame-interleaved video reasoning via reinforcement learning. *arXiv e-prints*, pages arXiv–2509, 2025. 5

[9] Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *arXiv preprint arXiv:2506.00318*, 2025. 6

[10] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191, 2025. 4

[11] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025. 5

[12] 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. 3

[13] Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025. 1

[14] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025. 1

[15] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025. 4

[16] Team Qwen3-VL, 2025. 5

[17] MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. Egonormia: Benchmarking physical social norm understanding. *arXiv preprint arXiv:2502.20490*, 2025. 4

[18] 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. 5

[19] Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song, and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. *arXiv preprint arXiv:2509.23652*, 2025. 6

[20] Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. *arXiv preprint arXiv:2507.09876*, 2025. 4, 5