VIDEO-SKILL-COT: Skill-based Chain-of-Thoughts for Domain-Adaptive Video Reasoning

Daeun Lee[1,*] **Jaehong Yoon**[1,2,*] **Jaemin Cho**[1] **Mohit Bansal**[1]
[1]UNC Chapel Hill [2]Nanyang Technological University
{daeun,jmincho,mbansal}@cs.unc.edu jaehong.yoon@ntu.edu.sg
https://video-skill-cot.github.io/

Abstract

Recent advances in Chain-of-Thought (CoT) reasoning have improved complex video understanding, but existing methods often struggle to adapt to domain-specific skills (e.g., event detection, spatial relation understanding, emotion understanding) over various video content. To address this, we propose VIDEO-SKILL-COT (a.k.a. VIDEO-SKOT) a framework that automatically constructs and leverages skill-aware CoT supervisions for domain-adaptive video reasoning. First, we construct skill-based CoT annotations: We extract domain-relevant reasoning skills from training questions, cluster them into a shared skill taxonomy, and create detailed multi-step CoT rationale tailored to each video-question pair for training. Second, we introduce a skill-specific expert learning framework. Each expert module specializes in a subset of reasoning skills and is trained with lightweight adapters using the collected CoT supervision. We demonstrate the effectiveness of the proposed approach on three video understanding benchmarks, where VIDEO-SKOT consistently outperforms strong baselines. We also provide in-depth analyses on comparing different CoT annotation pipelines and learned skills over multiple video domains.

1 Introduction

Understanding complex video content requires integrating rich spatiotemporal cues and adapting to diverse domain-specific reasoning needs from cinematic narratives, egocentric recordings, to indoor scenes (Fusier et al., 2007; Huang et al., 2018; Buch et al., 2022; Lin et al., 2023; Chen et al., 2024; Li et al., 2024c). Models should acquire and integrate a wide range of distinct reasoning skills, such as temporal grounding, spatial relationship recognition, and multi-step planning.

Recent work has extended chain-of-thought (CoT) reasoning (Wei et al., 2023; Kojima et al., 2022) to multimodal large language models (MLLMs) for video understanding (Fei et al., 2024; Feng et al., 2025; Li et al., 2025; Liu et al., 2025; Zhi et al., 2025). However, most prior approaches rely on fixed, general-purpose reasoning traces that are insensitive to domain-specific skills. Fig. 1 (left) shows a t-SNE (van der Maaten and Hinton, 2008) plot of embeddings of questions from different video datasets, where questions from the same datasets are strongly clustered as they require shared skills/domains. For example, models pretrained on general corpora such as LLaVA-Video-178K (Zhang et al., 2024) often lack the nuanced narrative understanding needed in CinePile (Rawal et al., 2024). This limits their ability to generalize to unseen domains or specialized skills.

To address this, we propose VIDEO-SKILL-COT (aka VIDEO-SKOT), a novel video understanding framework for creating and leveraging skill-aware CoT supervision, helping effective domain adaptation of MLLMs (Sec. 3). As shown in Fig. 1 (Right), VIDEO-SKOT consists of two main components. First, in skill-based CoT annotation (Sec. 3.2), we introduce a method to automatically construct high-quality, skill-conditioned CoT rationales for video QA tasks. Given a training question, we first extract high-level reasoning skill descriptions (e.g., "Determine object location relative to a person's orientation" and "Inferring emotional state from expressions and body language"), then cluster them into a shared skill taxonomy (Fig. 1 Right-(a)). Then, each question is annotated with its top-K relevant skills and used to generate a multi-step CoT annotation conditioned on these skills (Fig. 1 Right-(b)). This enables diverse and domain-relevant reasoning traces without requiring manual annotation.

Once we have prepared the skill-based CoT annotations, in skill-specific expert learning (Sec. 3.3

[*]Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18435–18449

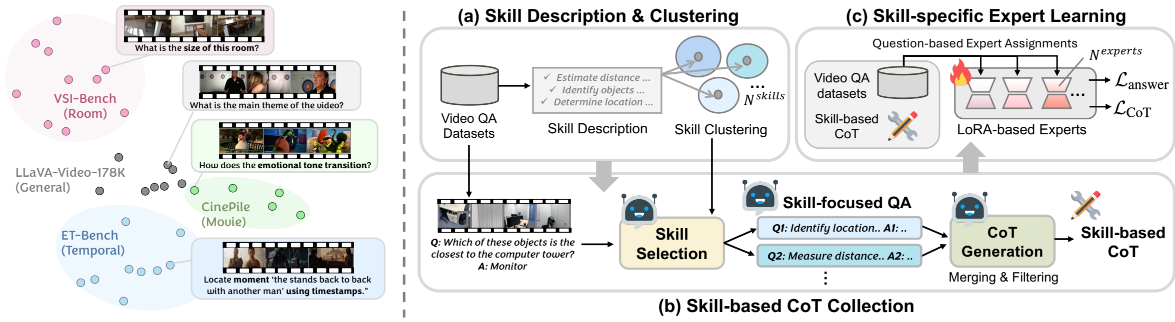

Figure 1: **Left**: Video datasets require different reasoning skills. **Right**: VIDEO-SKOT that automatically constructs and leverages skill-aware CoT supervisions for domain-adaptive video reasoning.

and Fig. 1 Right-(c)), we train skill-specialized expert models with multiple LoRAs (Hu et al., 2022). Each expert specializes in a specific set of reasoning capabilities, determined by a predefined group of related questions. During inference, the model routes each input to the expert aligned with the most relevant question group.

We evaluate VIDEO-SKOT on three video QA datasets with diverse domains (E.T.-Bench (Liu et al., 2024), VSI-Bench (Yang et al., 2024), and CinePile (Rawal et al., 2024)), where VIDEO-SKOT consistently improves over strong baselines, showcasing its strong domain adaptation capabilities. We also present ablation studies on our design choices and visualize the learned domain-specific skills to validate the effectiveness and interpretability of our skill-guided reasoning framework.

2 Related Work

Video Understanding with MLLMs. Prior video understanding models focused on pretraining strategies (Sun et al., 2019; Li et al., 2020; Lei et al., 2021). Recent work incorporates CoT reasoning (Kojima et al., 2022; Wei et al., 2023) from the NLP domain and studies how to collect and learn to generate such CoT reasoning for different video understanding tasks (Fei et al., 2024; Li et al., 2025; Liu et al., 2025; Zhi et al., 2025). Unlike these methods, which often struggle with comprehending videos without explicit skill-specific guidance, our approach introduces a skill-aware reasoning framework incorporating question-adaptive skill selection and skill-guided CoT supervision.

Skill-specific Expert Learning. Modular and expert-based architectures have been widely explored to improve parameter efficiency and mitigate interference in multi-task and multi-domain settings, where each expert learns different knowledge. Mixture-of-experts (MoE) frameworks dynamically route inputs to expert subnetworks (Shazeer et al., 2017), while adapter-based methods introduce lightweight, task-specific modules into pretrained models (Houlsby et al., 2019; Hu et al., 2022). Li et al. (2024b) studies learning skill-specific expert diffusion models for the text-to-image generation task. A concurrent work, Liu et al. (2025) studies a multi-agent system where each agent is implemented as a LoRA (Hu et al., 2022) expert. While Liu et al. (2025) relies on predefined expert roles (planner, grounder, verifier, and answerer), specific architectures, and manually curated role-specific annotations, our expert framework flexibly adapts to any video understanding dataset by automatically discovering and leveraging relevant reasoning skills.

3 VIDEO-SKILL-COT

3.1 Problem Setup

Given a video v and a question q, our objective is to produce both an answer a and a reasoning trace r that offers an interpretable, step-by-step justification. Prior work typically uses a single MLLM f to generate these: $\{r;\, a\}{=}f(q, v)$.

In contrast, VIDEO-SKOT decomposes the reasoning process into two stages: First, given q, we select the most relevant expert $e \in \{1, \ldots, N^{\text{experts}}\}$ based on the set of pre-defined question groups and predicted required skills. Next, a *skill-specific expert MLLM* f^e then generates a *skill-guided reasoning trace* r^s along with the final answer: $\{r^s;\, a\}{=}f^e(q, v)$. We illustrate VIDEO-SKOT in Fig. 1 (right).

This design enables targeted expert learning and adaptation to diverse reasoning skills in a new video domain. In the following, we describe how we automatically construct the skill-based CoT (Sec. 3.2) and how to train MLLMs with the collected skill-based CoT annotations (Sec. 3.3).

3.2 Skill-based CoT Annotation

We first construct skill-based CoT rationale annotations for any Video QA dataset, leveraging skill-aware reasoning to enable domain-adaptive video understanding. We perform the following two steps for each (q, v) in the training set to obtain skill-conditioned reasoning traces.

Step 1: Skill Description & Clustering (Fig. 1 Right-(a)). We define a *skill* as a shared, high-level reasoning capability (e.g., temporal ordering, visual counting, spatial understanding) that recurs across multiple video QA examples within a specific domain. For each question q, we prompt an MLLM to describe what kind of skill is necessary to answer it (e.g., *"Estimate distance between two objects using visual cues"*). Then, we encode all skill descriptions into text embeddings and perform k-means clustering (with k=N^{skills}=10) to form a shared skill taxonomy. Each cluster centroid represents a prototypical skill.

Step 2: Skill-based CoT Collection (Fig. 1 Right-(b)). For each (q, v) pair, we generate a multi-step reasoning trace conditioned on the descriptions of the top 3 assigned skills, a process we refer to as *Skill Selection*. Next, we generate the skill-aware CoT r^s; We prompt an MLLM to produce intermediate sub-questions and corresponding answers, guided by selected skills from the previous stage. These sub-QA pairs are then merged into a coherent CoT paragraph that explicitly reflects the assigned reasoning skills. To ensure the quality of the skill-based CoT rationales, we further verify and filter out reasoning steps that are irrelevant to the correct answer using an LLM evaluator.

After these steps, each training example is now annotated with relevant expert labels e and a verified, skill-grounded CoT trace r^s. These annotations form the basis for downstream training of skill-specific expert models.

3.3 Skill-specific Expert Learning

As illustrated in Fig. 1 Right-(c), we perform modularized fine-tuning to learn task-specific knowledge for skill-based CoT training. Specifically, we first project all questions in training set D^{train} into the text embedding space and perform k-means clustering (with k=N^{experts}=5). Unlike step 2 of Sec. 3.2 where N^{skills} clusters represent the groups of *skill descriptions*, these N^{experts} cluster centroids represent the groups of *questions*. After assigning each training example to its closest N^{experts}, we conduct parameter-efficient training using the corresponding N^{experts} expert LoRA (Hu et al., 2022) modules, ensuring task-specific adaptation while minimizing interference across skills. During test time, we assign each test question by finding the closest question group by finding the closest question embedding centroids.

Training Objective. Following previous work (Hu et al., 2024; Shi et al., 2024), we train an MLLM by minimizing cross-entropy losses for predicting both the answer ($\mathcal{L}_{\text{answer}}$) and CoT tokens ($\mathcal{L}_{\text{CoT}}$), respectively:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{answer}} + \lambda \mathcal{L}_{\text{CoT}} \\ &= \ell(f(q, v), a) + \lambda \ell(f(q, v), r^s), \end{aligned} \tag{1}$$

where we find λ = 0.5 balances the two losses well.

4 Experiments

4.1 Experiment Setups

Implementation Details. To obtain text embeddings (of skill taxonomy in Sec. 3.2 and of questions in Sec. 3.3), we use `all-mpnet-base-v2` SentenceTransformers (Reimers and Gurevych, 2019) implementation. We use LLaVa-Video (7B) (Zhang et al., 2024) as a main backbone model. Additional training details including hyperparameters, the specific MLLMs and LLMs used at each stage, as well as results with the Qwen2.5-VL (7B) backbone are provided in Appendix Secs. A.1, A.2 and B.2.

Datasets and Baselines. We experiment with three different video understanding benchmarks with distinct domains: E.T.Bench (Liu et al., 2024) (temporal understanding), VSI-Bench (Yang et al., 2024) (spatial understanding), and CinePile (Rawal et al., 2024) (movie narrative understanding). For multiple-choice questions, we report the average accuracy. For temporal captioning tasks in E.T.Bench, we use the benchmark's official evaluation script. Baseline MLLMs include mPLUG-Owl (Ye et al., 2024), Video-ChatGPT (Maaz et al., 2023), Video-LLaMA2 (Zhang et al., 2023), LLaVa-OneVision (Li et al., 2024a), and LLaVa-Video (Zhang et al., 2024), GPT4o (Achiam et al., 2023a) and Gemini 1.5 Flash, Pro (Georgiev et al., 2024). Additional details are provided in the Appendix Sec. A.3.

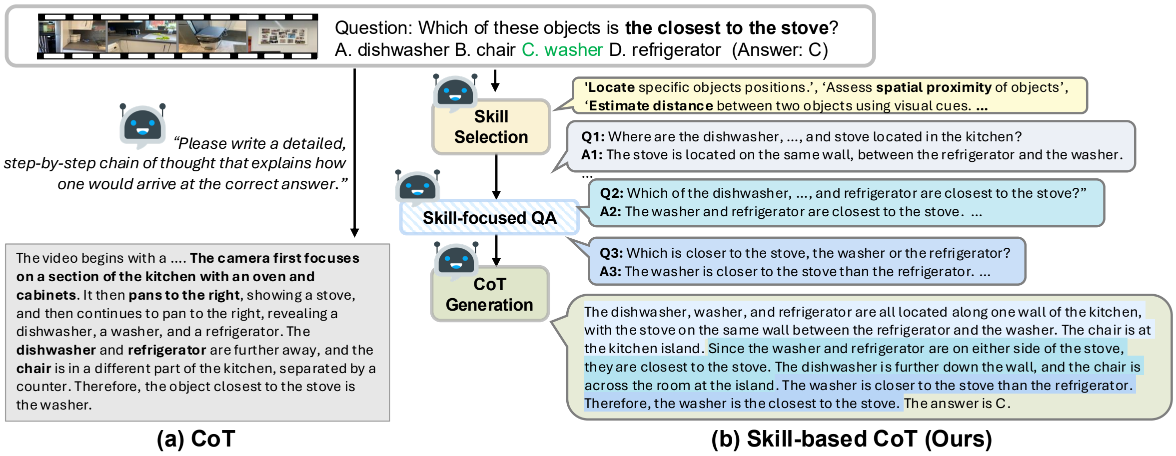

Figure 2: **Comparison of CoT annotations: (a) regular CoT and (b) our skill-based CoT.** Additional examples are provided in Appendix Sec. C.

	Fine-tuned	E.T.Bench (Temporal)	VSI (Spatial)	CinePile (Movie)
Closed-source MLLMs				
GPT4o (Achiam et al., 2023a)		24.69	34.00	56.06
Gemini 1.5 Pro (Georgiev et al., 2024)		26.73	45.40	60.12
Gemini 1.5 Flash (Georgiev et al., 2024)		27.92	42.10	58.75
Open-source MLLMs (7B)				
mPLUG-Owl (Ye et al., 2024)	✗	11.87	-	13.93
Video-ChatGPT (Maaz et al., 2023)	✗	13.02	-	15.08
Video-LLaMA2 (Zhang et al., 2023)	✗	8.30	-	44.57
LLaVA-Video (Zhang et al., 2024)	✗	19.35	35.60	55.83
LLaVA-Video (Zhang et al., 2024)	✓	20.32	47.45	56.29
Ours	✓	**22.21**	**53.15**	**57.88**

Table 1: **Evaluation results on domain-specific video reasoning benchmarks.**

4.2 Quantitative Evaluation

Comparison to Baselines. We compare VIDEO-SKOT to recent MLLM baselines on three video understanding benchmarks (E.T.Bench, VSI-Bench, CinePile) with domains and required skills. Table 1 shows that VIDEO-SKOT consistently outperforms all baselines, achieving improvements of +4.10, +5.70, and +1.59 over the fine-tuned version of LLaVA-Video on E.T.Bench, VSI-Bench, and CinePile, respectively. These results highlight the effectiveness of our modular, expert-driven framework in enabling domain-adaptive CoT video reasoning by leveraging relevant skills.

Ablation Studies. We compare the impact of two key components in our framework: (1) skill-based CoT reasoning and (2) skill-specific expert modules. As shown in Tab. 2, our full model, combining both components (Top row), achieves the highest performance. Removing either the skill-specific expert modules (2nd row), the skill-based CoT (3rd row), or both components (last row) consistently leads to performance degradation, highlighting their complementary roles: skill-CoT enables structured reasoning, while expert modules bring modular specialization. This synergy proves essential for improving video understanding.

Skill-CoT (Sec. 3.2)	Skill-specific Experts (Sec. 3.3)	3-Task Avg.
✔	✔	**44.41**
✔	-	42.91
-	✔	38.53
-	-	41.04

Table 2: **Ablation studies on removing the main components**: Skill-CoT and skill-specific experts.

Criterion	Regular CoT	Skill CoT (Ours)	Δ (Ours - Regular)
Correctness	2.88 ± 1.61	4.97 ± 0.16	+2.09
Relevance	3.16 ± 1.55	4.80 ± 0.59	+1.64
Coherence	2.80 ± 1.42	4.73 ± 0.64	+1.93

Table 3: **Human evaluation results comparing Regular CoT and Skill-CoT.** Scores are reported as mean ± standard deviation.

Human Evaluation. We conduct a human evaluation with five researchers who are familiar with the relevant field, where 15 randomly selected questions were assessed by comparing regular CoT and the proposed Skill-based CoT. Each explanation is rated on a 1–5 Likert scale (5 = best, 1 = worst) across three dimensions: Correctness (factual accuracy), Relevance (task appropriateness), and Coherence (clarity and logical flow). As shown in Tab. 3, Skill-based CoT consistently outperforms regular CoT across all criteria, with substantial gains in correctness, relevance, and coherence, confirming that our method produces explanations that are more accurate, aligned, and easier to follow. These results provide strong evidence that Skill-based CoT produces explanations that are not only more accurate

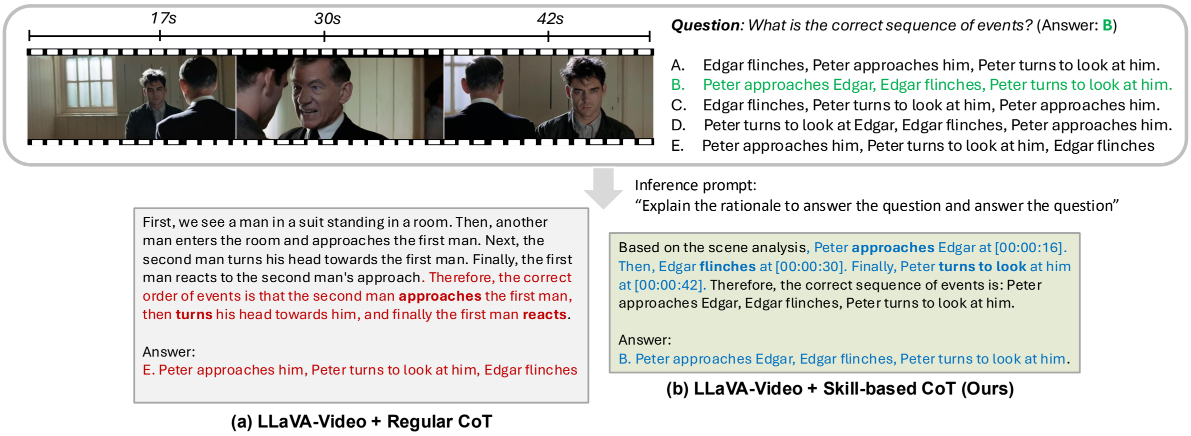

Figure 3: **Inference output comparison: (a) LLaVA-Video trained with regular CoT and (b) LLaVA-Video trained with our skill-based CoT.** VIDEO-SKOT successfully generates temporally grounded and precise rationales that more effectively support accurate answer generation.

but also more relevant and human-readable.

4.3 Qualitative Analysis

Regular CoT vs. Skill-based CoT. Fig. 2 compares the different annotated CoTs from the regular CoT and our skill-based CoT. Given a question about which object is closest to the stove, the regular CoT (left) offers a linear, scene-based narration that lacks structure and includes irrelevant details ("Camera first focuses ... it then pans to the right ..."), making it often harder to extract key spatial information. In contrast, our skill-based CoT starts by identifying relevant skills (e.g., spatial proximity) and breaking the task into focused sub-questions, like comparing the washer and refrigerator.

Inference rationale comparison We compare the inference-time rationales generated by LLaVA-Video trained with (a) regular CoT and (b) the proposed skill-based CoT. During inference, we prompt each model with: *"Explain the rationale to answer the question and answer the question."* As shown in Fig. 3, the model trained with regular CoT produces an incorrect reasoning process, ultimately leading to a wrong answer. In contrast, VIDEO-SKOT successfully generates temporally grounded and precise rationales that more effectively support accurate answer generation.

5 Conclusion

We propose VIDEO-SKOT, a novel video understanding framework for effective domain adaptation of MLLMs. We propose to automatically collect skill-specific CoT annotations from video QA datasets and construct a skill-based reasoning pipeline that combines a lightweight skill assigner with a collection of LoRA-based expert adapters. Empirical results on three diverse benchmarks demonstrate consistent gains of VIDEO-SKOT over strong baselines, highlighting the enhanced quality of our reasoning traces.

Limitations

Our proposed framework demonstrates strong video reasoning capabilities, generating fine-grained, domain-adaptive rationales based on required skills. However, it may still produce occasional inaccuracies or hallucinations (Liu et al., 2023; Wang et al., 2024; Zhou et al., 2024) in its text outputs. Additionally, the overall performance is influenced by the underlying pre-trained backbones, namely, the LLM (Achiam et al., 2023b) and MLLM (Georgiev et al., 2024) used. Nonetheless, we highlight that VIDEO-SKOT can benefit further from future advancements in LLM and MLLM backbones.

Acknowledgments

This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, DARPA Machine Commonsense (MCS) Grant N6600119-2-4031, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, a Bloomberg Data Science PhD Fellowship, and the Accelerate Foundation Models Research program. The views contained in this article are those of the authors and not of the funding agency.

References

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

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" video" in video-language understanding. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024. Sharegpt4video: Improving video understanding and generation with better captions. In Advances in Neural Information Processing Systems (NeurIPS).

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776.

Florent Fusier, Valery Valentin, François Brémond, Monique Thonnat, Mark Borg, David Thirde, and James Ferryman. 2007. Video understanding for complex activity recognition. Machine Vision and Applications, 18:167–188.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. arXiv:1902.00751.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9590–9601.

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. 2018. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learningvia sparse sampling. In CVPR.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

Jialu Li, Jaemin Cho, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. 2024b. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. In Advances in Neural Information Processing Systems (NeurIPS).

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+language omni-representation pre-training. arXiv:2005.00200.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In Proceedings of the International Conference on Computer Vision (ICCV).

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In Proceedings of the International Conference on Learning Representations (ICLR).

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025. Videomind: A chain-of-lora agent for long video reasoning. arXiv preprint arXiv:2503.13444.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Chang Wen Chen, and Ying Shan. 2024. E.t. bench: Towards open-ended event-level video-language understanding. In Neural Information Processing Systems (NeurIPS).

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. CoRR, abs/1701.06538.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. 2024. Unlocking video-llm via agent-of-thoughts distillation. arXiv preprint arXiv:2412.01694.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. arXiv:1904.01766.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research, 9:2579–2605.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In International Conference on Multimedia Modeling. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. arXiv preprint arXiv:2412.14171.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR).

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

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

Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, and 1 others. 2025. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. arXiv preprint arXiv:2504.04471.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In Proceedings of the International Conference on Learning Representations (ICLR).

Appendix

A VIDEO-SKOT Implementation Details **8**
- A.1 Details of skill description & clustering . . . 8
- A.2 Details of skill-based CoT generation 8
- A.3 Details of training . . . 8
- A.4 Prompts . . . 8

B Additional Quantitative Results **8**
- B.1 Per-category performance . . . 8
- B.2 Qwen2.5-VL backbone . . . 8
- B.3 Cross-dataset generalization . . . 8

C Additional Qualitative Results **9**
- C.1 Skill descriptions over different datasets . . . 9
- C.2 Selected skills over different video datasets . . . 9
- C.3 Additional comparison with regular CoT . . . 9

D License **10**

A VIDEO-SKOT Implementation Details

A.1 Details of skill description & clustering

Skill Description. To extract skill descriptions given the training dataset, we prompt GPT-4[1] with its questions and answers. (The prompt is provided in Fig. 8) Each extracted skill is written as a concise skill phrase (6–12 words), preserving the core visual or temporal reasoning concept. Here, we intentionally exclude audio-based cues (e.g., sound, speech, or music) in this process. Specific object names (e.g., "TV", "sofa", "John") are replaced with generic terms, and vague terms (e.g., "reasoning", "analysis") are avoided to enhance clarity. We also provide the exact name of the skills in Tab. 4.

A.2 Details of skill-based CoT generation

For skill-based CoT generation, we utilize Gemini-2.0 Flash with video input. As illustrated in Fig. 9, we first prompt Gemini-2.0 to identify the relevant skills and generate corresponding sub-questions and answers. Then, we construct step-by-step reasoning based on this output. Finally, we use GPT-4 to filter and verify the reasoning by assessing its relevance to the ground-truth answers using Fig. 11 as a prompt.

[1] gpt-4-32k

A.3 Details of training

Training datasets. Instead of using the full video instruction tuning dataset, we randomly sampled 10k and 2.1k examples from ET-Bench and CinePile, respectively. For VSI-Bench, which is intended solely for evaluation and does not provide a training set, we manually split the available data into training and test sets using a 7:3 ratio. We use 3k training dataset for VSI-Bench.

Hyperparameters. For training, we set the learning rate as 1e-5 and the batch size as 1. For LoRA, we use rank 32. We set 1 epoch for ET-Bench training and 3 epochs for the other two datasets. For other parameters, we use the default setup of LLaVA Video. We use 4 A6000 GPUs for training.

A.4 Prompts

In Figs. 8 to 11, we attach prompts for skill-based CoT annotation. We also attach prompt to generate regular CoT in Fig. 12.

B Additional Quantitative Results

B.1 Per-category performance

In Tabs. 5 to 7, we additionally report the per-category performance for each dataset. We also include ablation studies comparing regular CoT vs skill-based CoT, and single-LoRA vs multi-LoRA configurations. VIDEO-SKOT, which combines skill-based CoT with multi-LoRA training, consistently outperforms across all datasets, showing particularly strong gains on reasoning-intensive tasks such as Route Planning in VSI-Bench and temporal understanding tasks in CinePile.

B.2 Qwen2.5-VL backbone

We further evaluate VIDEO-SKOT on VSI-Bench using the Qwen2.5-VL (7B) (Bai et al., 2025) backbone. As shown in Tab. 8, VIDEO-SKOT achieves the highest overall performance (40.76 avg), consistently surpassing both regular-CoT and single-LoRA baselines. These results highlight the robustness and effectiveness of VIDEO-SKOT when applied to a different backbone architecture.

B.3 Cross-dataset generalization

We evaluate cross-domain generalization from ET-Bench (source) to CinePile (target). As shown in Tab. 9, VIDEO-SKOT achieves the best average performance (56.21) among ET-Bench–trained variants, performing competitively with the CinePile

	Skill descriptions
VSI-Bench	"Determine object location relative to a person's orientation.", "Locate specific objects and identify their positions in the scene.", "Identify the closest object to a reference point.", "Recognize objects based on shape, color, and features.", "Identify and count distinct objects based on location and appearance.", "Assess spatial proximity of objects relative to a reference point.", "Identify initial appearances of objects in the video timeline.", "Estimate distance between two objects using visual cues.", "Determine room boundaries using structural elements like walls and floors.", "Identify sequential order of object appearances in a scene."
ET-Bench	"Identify head orientation and gaze direction to infer focus.", "Identify spatial proximity between people in the scene.", "Identifying the timestamp of an action in a scene.", "Detect object using shape, texture, and visual features.", "Track individuals interacting with an object and their actions.", "Identify a person performing an action involving an object.", "Identifying the moment an action occurs in a scene.", "Detect a person using body shape, face, and clothing.", "Identify actions through body movements and postures.", "Identify hand movement and physical interaction with an object."
CinePile	"Identify thematic parallels between actions and overarching narrative themes.", "Inferring emotional tone from facial expressions and actions.", "Tracking emotional shifts through expressions and body language changes.", "Identifying interpersonal conflict through observed actions and interactions.", "Inferring symbolic meaning of an object in a scene.", "Inferring emotional state from expressions and body language.", "Track a person's movements and reactions to scene changes.", "Identifying body language among the scene.", "Identify event sequence to infer action context and significance.", "Identifying the main person or subject in the scene."

Table 4: **Detailed skill descriptions from three datasets.**

fine-tuned model (56.29) and surpassing the zero-shot baseline (55.83). This highlights the effectiveness of skill-guided reasoning for transfer across domains.

C Additional Qualitative Results

C.1 Skill descriptions over different datasets

In Fig. 7, we visualize the skill descriptions for each dataset after performing skill extraction and clustering (Sec. 3.2). To create the visualization, we first obtain text embeddings using Sentence-Transformer and compute N^{skills} cluster centroids. We then apply t-SNE to reduce the dimensionality of the embeddings for visualization purposes. The results highlight that each domain-specific dataset emphasizes different skill sets, though certain skills are shared across datasets. For instance, the skill "*Inferring emotional tone from facial expressions and actions*" from CinePile is distinct from "*Estimating distance between two objects in the video timeline*" from VSI-Bench. However, general skills like "*Identifying objects or people*" appear across multiple datasets. A more detailed list of the extracted skills is provided in Tab. 4.

C.2 Selected skills over different video datasets

In Figs. 4 and 5, we present statistics on the selected top 3 assigned skills for each task in VSI-Bench (presented in Sec. 3.2). As shown in the results, object identification skills are commonly used across tasks. However, each task also requires domain-specific skills. For instance, the Room Size Estimation task necessitates skills such as "*Determining room boundaries using structural elements like walls and floors.*"

C.3 Additional comparison with regular CoT

In Fig. 6, we provide additional comparison with regular CoT and ours.

	Appr. Order	Rel. Dist.	Route Plan	Rel. Dir.	Abs. Dist.	Obj. Count	Obj. Size	Room Size	Avg
LLaVA-Video	31.28	41.57	35.59	46.05	10.78	52.24	47.54	19.77	35.60
LLaVA-Video (fine-tune)	54.16	36.97	41.66	42.26	29.71	56.89	70.16	54.51	47.45
Single-LoRA + Regular-CoT	68.75	36.97	41.66	43.07	31.37	62.62	70.22	65.96	52.58
Single-LoRA + Skill-CoT	64.58	42.01	36.11	44.45	33.11	56.89	74.27	64.35	52.97
Multi-LoRA + Regular-CoT	56.25	34.45	41.66	39.53	12.24	43.00	55.11	23.87	38.27
VIDEO-SKOT (Ours)	68.75	36.13	50.00	47.69	32.13	62.61	70.83	57.09	53.15

Table 5: **Detailed VSI-Bench Results**.

	RAR	EVC	RVQ	TVG	ERM	TAL	EVS	VHD	DVC (F1)	DVC (Sim)	SLC (F1)	SLC (Sim)	TEM	GVQ	Avg
LLaVA-Video	41.6	38.8	56.6	8.2	1.8	14.0	14.8	28.2	20.1	10.0	11.5	8.1	15.7	1.5	19.3
LLaVA-Video (fine-tune)	44.8	34.6	58.2	9.4	2.2	12.0	9.2	29.7	28.6	14.8	10.1	10.2	18.6	2.1	20.3
Single-LoRA + Regular-CoT	43.6	37.8	58.6	10.7	1.9	13.7	7.5	33.8	14	12.2	6.5	10.6	12.4	3.1	19.0
Single-LoRA + Skill-CoT	43.2	40.8	56.8	10.0	1.7	15.2	7.1	30.2	17.3	16.5	10.0	10.6	11.2	1	19.4
Multi-LoRA + Regular-CoT	43.2	40.6	59.8	6.9	1.9	16.5	11.1	31.1	30.3	16.4	6.2	9.5	16.6	1.0	20.7
VIDEO-SKOT (Ours)	49.0	41.2	59.4	15.8	2.4	16.4	8.4	35.5	27.0	15.0	11.9	10.0	16.5	2.4	22.2

Table 6: **Detailed E.T-Bench Results**.

D License

We list the license of the benchmark dataset and models we used. We use these existing artifacts consistently with their intended use.

- LLaVA-Video: Apache License 2.0
- CinePile: cc-by-nc-sa-4.0
- VSI-Bench: Apache License 2.0
- ET-Bench: cc-by-nc-sa-4.0

	CRD	NPA	STA	TH	TEMP	Avg
LLaVA Video	56.78	58.53	60.31	60.52	43.02	55.83
LLaVA Video (fine-tune)	57.74	58.32	60.44	61.05	43.90	56.29
Single-LoRA + Regular CoT	59.18	59.61	61.81	61.05	40.91	56.11
Single-LoRA + Skill-CoT	57.88	56.8	60.77	60.52	40.84	56.36
Multi-LoRA + Regular CoT	58.17	59.17	60.18	62.63	42.44	56.52
VIDEO-SKOT (Ours)	60.00	59.61	61.44	63.15	45.20	57.88

Table 7: **Detailed Cinepile Results.** We ablate the accuracies across the question categories: TEMP - Temporal, CRD - Character and Relationship Dynamics, NPA - Narrative and Plot Analysis, STA - Setting and Technical Analysis, TH - Thematic Exploration.

Model	**Appr. Order**	**Rel. Dist.**	**Route Plan**	**Rel. Dir.**	**Abs. Dist.**	**Obj. Count**	**Obj. Size**	**Room Size**	**Avg**
Qwen2.5-VL	10.41	35.29	36.11	37.80	20.43	23.98	54.22	33.06	31.41
Qwen2.5-VL (fine-tune)	18.75	34.45	36.11	38.65	21.88	31.74	62.55	47.74	36.48
Single-LoRA + Regular-CoT	20.83	35.29	36.11	39.92	21.73	34.56	63.33	48.06	37.48
Single-LoRA + Skill-CoT	27.08	35.61	**41.66**	40.12	21.57	35.33	65.77	**48.58**	39.47
Ours (Multi-LoRA + Skill-CoT)	**31.28**	**37.89**	36.50	**43.67**	**28.12**	**37.48**	**67.31**	43.79	**40.76**

Table 8: **Qwen2.5-VL (7B) Results on VSI-Bench**.

Setting	**Training Data**	**CRD**	**NPA**	**STA**	**TH**	**TEMP**	**Avg**
LLaVA-Video	-	56.78	58.53	60.31	60.52	43.02	55.83
LLaVA-Video (fine-tune)	CinePile	57.74	58.32	60.44	61.05	43.90	56.29
LLaVA-Video (fine-tune)	ET-Bench	53.34	57.02	57.83	60.53	40.55	53.85
Single-LoRA + **Regular-CoT**	ET-Bench	56.23	60.47	58.09	63.68	38.66	55.43
Multi-LoRA + **Skill-CoT (Ours)**	ET-Bench	56.02	61.25	58.88	63.33	41.57	**56.21**

Table 9: **Cross-dataset evaluation on CinePile**. Source: ET-Bench → Target: CinePile. Backbone: LLaVA-Video.

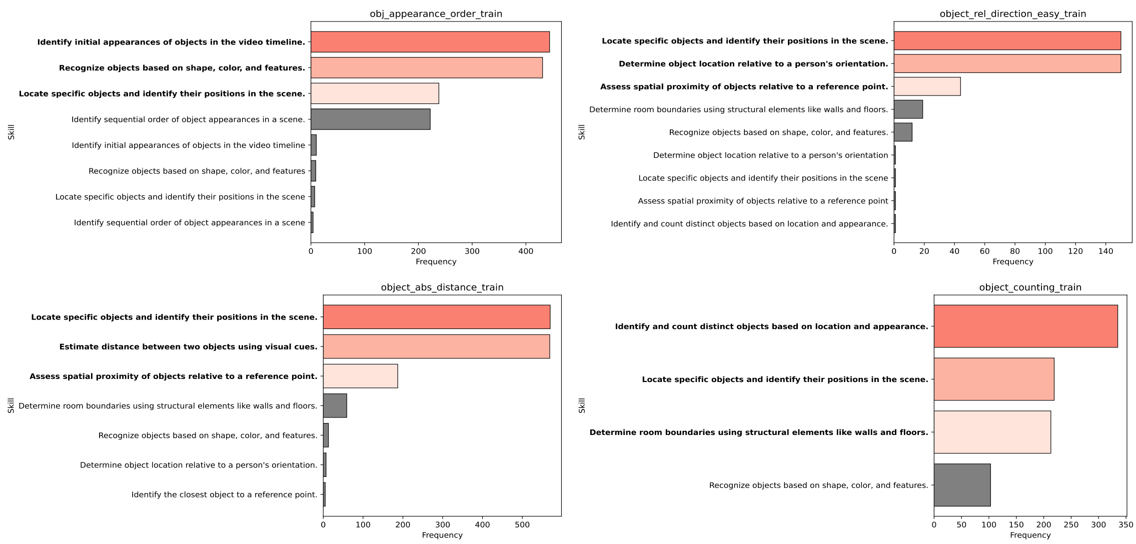

Figure 4: **Skill selection results of VSI-Bench (1)**

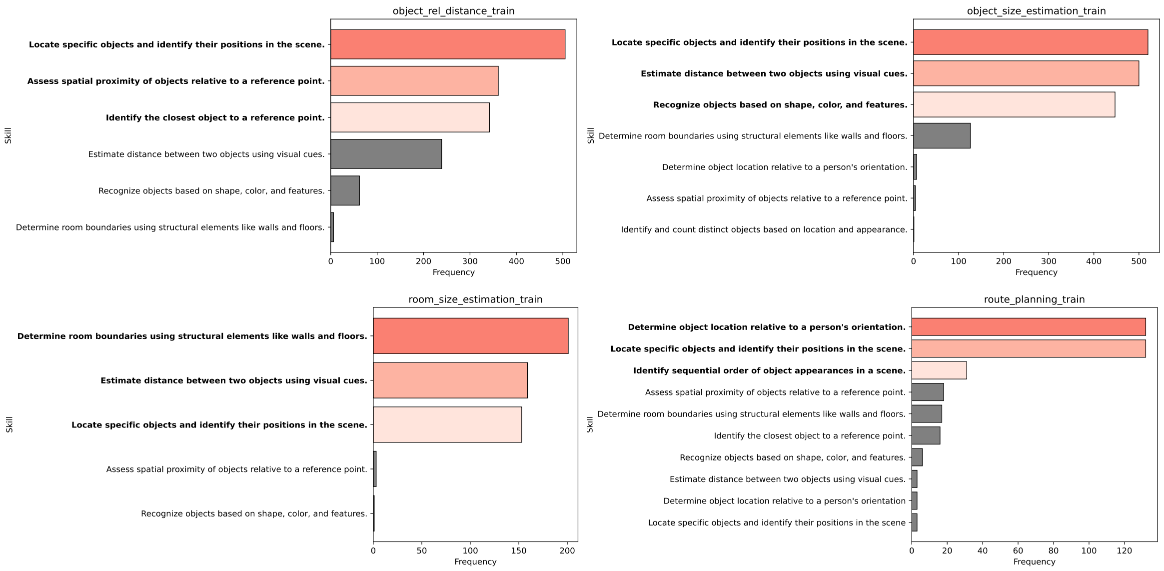

Figure 5: **Skill selection results of VSI-Bench (2)**

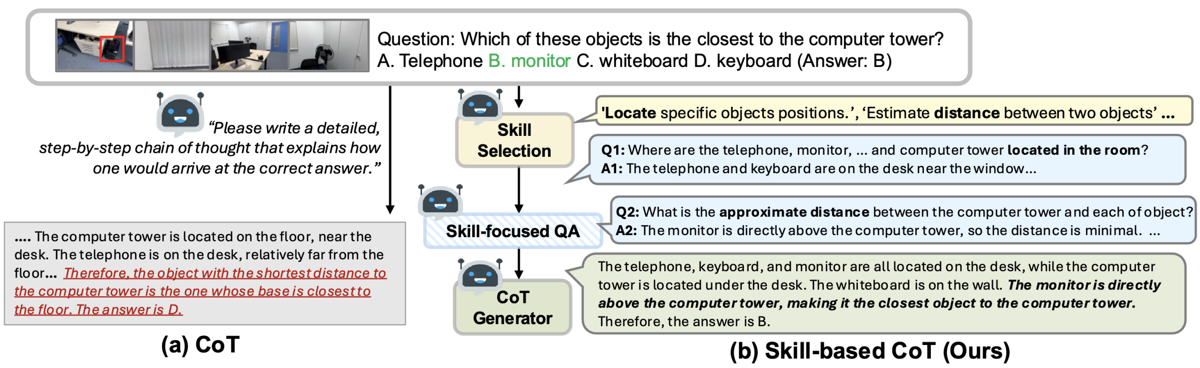

Figure 6: **Additional comparison of CoT annotations: (a) regular CoT and (b) our skill-based CoT.**

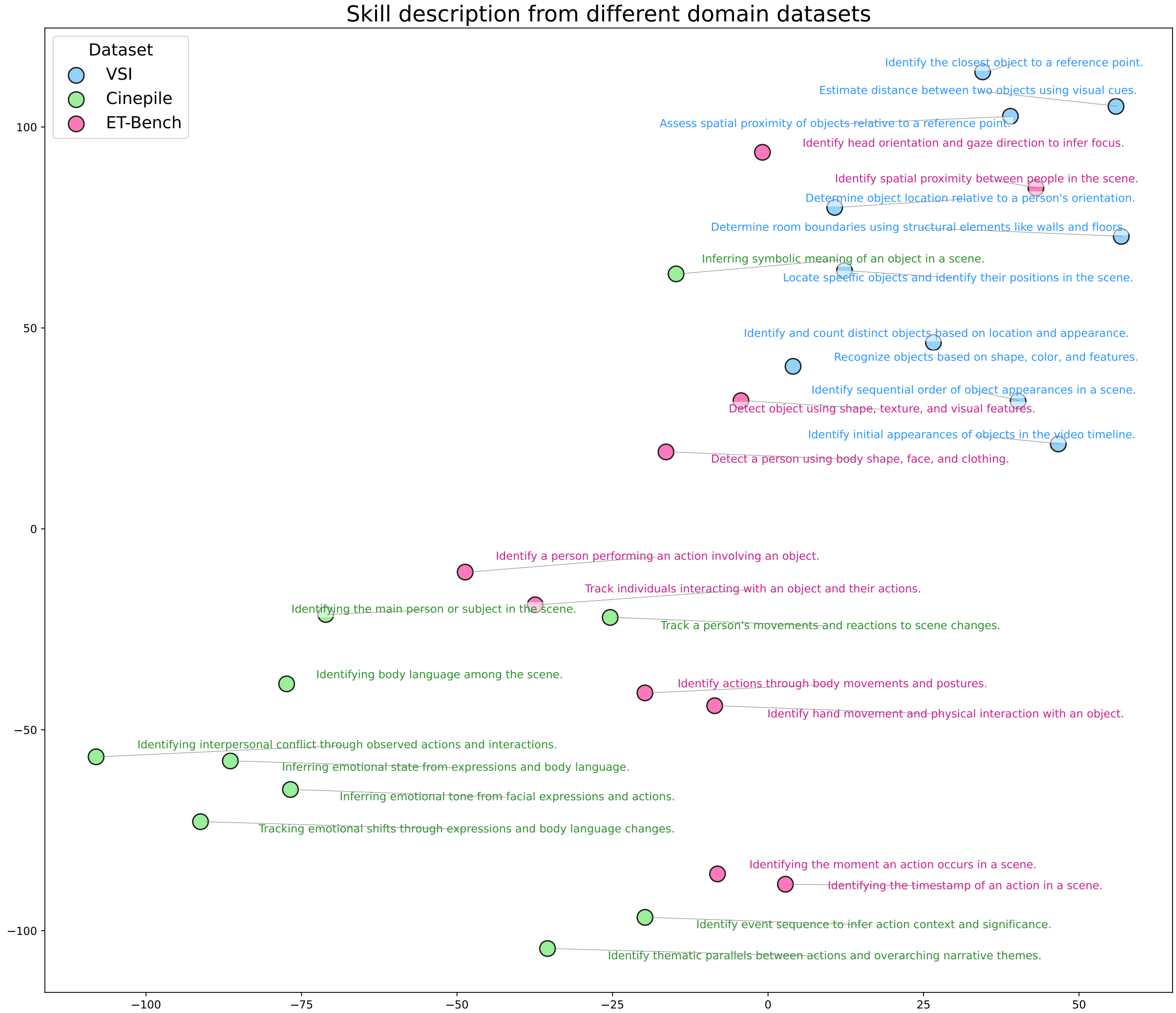

Figure 7: **Skill description from different domain datasets.** We visualize the skill descriptions for each dataset after performing skill extraction and clustering. (Sec. 3.2)

```
Your task is to generate a step-by-step reasoning process to answer the following video-based question correctly, using a chain of
distinct visual or temporal reasoning skills.
Each step must:
- Apply **one specific visual or temporal skill** (e.g., spatial reasoning, object tracking, causal inference)
- Clearly describe what is being inferred or understood at that step
- Build logically on the previous step
- Avoid using audio-based skills (e.g., sound, speech, music)

If fewer than five steps are necessary, return only the needed steps.
Each step should contribute uniquely to the final answer.

Return the output as a JSON list of objects with this format:

[
{{
"skill": "name of the reasoning skill used",
"step": "description of the reasoning step"
}},
...
]

Question: {question}
Answer: {answer}
```

Figure 8: **Prompt for Skill Description**

```
You are a video understanding assistant that performs modular reasoning using visual skills. You are given:
- A video-based question
- The ground-truth answer
- The relevant video content (as input)
- A list of domain-specific visual reasoning skills

Your task is to:
1. Select the skills needed to answer the question.
2. For each selected skill:
a. Generate a focused sub-question that applies the skill.
b. Answer the sub-question using information from the video.

Use the exact format below. Make sure the output is valid and parseable.
---

Question: {question}
Answer: {answer}
Relevant Skills:
{skill_list}

Output Format:
{
"selected_skills": [ "Skill A", "Skill B", ... ],
"skill_reasoning_steps": [
{
"skill": "Skill A",
"sub_question": "...",
"output": "..."
},
{
"skill": "Skill B",
"sub_question": "...",
"output": "..."
}
]
}
```

Figure 9: **Prompt for skill selection and sub-QA generation**

```
You are a reasoning assistant that combines modular skill outputs to solve a complex video understanding task.

Given the original question and a set of skill-based outputs derived from the video,
your task is to use these as evidence to perform multi-step reasoning and produce a final answer.

Use the skill outputs to guide your reasoning. Do not copy them verbatim—use them as evidence in your own words.

---

Question: {question}

Skill Outputs:
{skill_output_text}

Final Answer with Reasoning:
```

Figure 10: **Prompt for skill-based CoT generation**

```
You are given a question, its ground-truth answer, and a reasoning chain.

Your task is:
1. If the reasoning is irrelevant to the answer, return:
{{ "relevance": "No", "revised_reasoning": null }}

2. If the reasoning is relevant, do the following:
- Convert any bullet-point or timestamped list into a natural, coherent paragraph.
- Ensure the final reasoning clearly ends with a statement of the correct answer.

Then respond with:
{{ "relevance": "Yes", "revised_reasoning": "[Revised reasoning with natural paragraph and final answer]" }}

Keep the original reasoning steps as intact as possible.

Respond in JSON format only.

Question: {question}
Answer: {answer}
Reasoning: {cot}
```

Figure 11: **Prompt for CoT filtering**

```
Given the video and the question: {question}
The correct answer is: {answer}

Please write a detailed, step-by-step chain of thought that explains how one would arrive at the correct answer. Your explanation
should be written as a coherent paragraph rather than a list or dictionary.
Focus solely on visual elements and any on-screen text from the video. Do not use or rely on any audio information such as speech,
sound effects, or music.
Do not reveal or repeat the final answer in your reasoning. Focus only on the logical visual steps that would justify the correct
answer without stating it explicitly.
Ensure that each step in your reasoning naturally follows from the previous one, and that the overall explanation clearly supports
why the provided answer is correct (without mentioning the answer itself).
```

Figure 12: **Prompt for regular CoT generation**