Factorized Learning for Temporally Grounded Video-Language Models

Supplementary Material

A. More Implementation Details

Here we provide more implementation details in addition to the main paper.

Following existing practice [9, 11], we adopt 1 FPS frame sampling for both training and testing. Frames are resized to 224×224 before being fed into the network.

To determine the salient tokens for grounding and explicit event-level visual semantic capture, we treat frames within the ground-truth interval as salient during training. During inference, if the feature similarity between a frame-level video token and <evi> exceeds 60% of the maximum similarity between the current <evi> token and all frame tokens, the corresponding frame-level video token is considered salient and included. Sec. C.4 demonstrates the performance robustness to threshold variation.

Before similarity calculation, the <evi> token is first projected through a 2-layer MLP. This projection helps to distinguish its two functional roles: serving as a generation token during autoregressive decoding (via a standard LM classification head), and acting as a query token for similarity-based grounding and visual semantic aggregation. This design facilitates the joint learning of these related but functionally distinct tasks.

For the performance comparison among different methods, most reported numbers are directly taken from the original papers, except for Qwen2.5-VL [1] on E.T. Bench [11], which we re-implemented due to the absence of official results. We found that the performance is highly sensitive to the prompt and pixel configurations, which aligns with findings discussed in the context of video temporal grounding on GitHub[1]. We combine the official cookbook from Qwen2.5-VL and the practice from lmms-eval[2], resulting in considerably higher performance compared with directly using the official cookbook which may not be tailored for benchmarking purposes. We report the highest performance of Qwen2.5-VL that we were able to achieve in our paper.

B. Data Annotation Formats

Here, we also provide the annotation formats for model training, which offer an intuitive and clear understanding of D^{2}VLM's generation objective. Based on the input-output format, we categorize different tasks into three main types: (1) Grounding-focused task (e.g., temporal video grounding, action localization, etc.), which can involve single-event grounding and multi-event grounding. (2) Dense captioning-related task, which requires grounding multiple events throughout the entire video while also providing a textual description for each grounded event. (3) temporally grounded video question answering, which involves answering the user's open-ended questions while also providing the temporal position of the answer (evidence).

[1] https://github.com/QwenLM/Qwen2.5-VL/issues/837

[2] https://github.com/EvolvingLMMs-Lab/lmms-eval

Method	Year	TEM (Rec)	GVQ (Rec)
Video-ChatGPT-7B [12]	ACL'24	15.9	0.0
Video-LLaVA-7B [10]	EMNLP'24	7.5	0.1
LLaMA-VID-7B [9]	ECCV'24	7.0	0.9
Video-LLaMA-2-7B [3]	arXiv'24	0.0	0.1
PLLaVA [14]	arXiv'24	4.1	1.2
VTimeLLM-7B [6]	CVPR'24	6.8	1.9
VTG-LLM-7B [5]	AAAI'25	8.9	1.4
TimeChat-7B [13]	CVPR'24	18.0	1.5
LITA-13B [7]	ECCV'24	16.0	2.2
E.T. Chat-3.8B [11]	NeurIPS'24	16.5	3.7
D^{2}VLM-3.8B (Ours)	ICCV'25	**29.2**	**7.1**

Table 1. Performance comparison on TEM (Temporal Event Matching) and GVQ (Grounded Video Question answering).

The examples are shown in Fig. 1. Here, we provide the definitions of some important keys in the annotation files. The "conversations" key is the main component, which consists of two sub-parts: "from human" and "from gpt". The value corresponding to the "from human" part represents the input prompt, which mainly includes the video (represented here as a place-holder <image>, but will be actually replaced by video frames) and the user question (instruction). The second part, "from gpt", represents the desired model response sequence, which typically consists of two stages: the pure evidence grounding stage and the interleaved text-evidence token generation stage. These two stages are separated by the </evi> token, which the model should also generate to indicate the end of the evidence grounding stage and the beginning of the interleaved response. Another important key is "time_gt," which indicates the ground-truth temporal event position. This is used to supervise the similarity calculation between the <evi> token and frame-level tokens, as mentioned in this paper. Here, the ground-truth annotations for the evidence grounding stage and the interleaved response stage are the same, based on the natural assumption that the grounded evidence should be consistent with the answer.

C. More Experimental Results

C.1. Performance on E.T. Bench Complex Dataset

Here we also compare the performance on E.T. Bench Complex dataset [11] that involves two sub-tasks: temporal

Method	MVBench	Video-MME (w/o subs)
Video-LLaVA-7B [10]	43.0	39.9
E.T. Chat-3.8B [11]	36.4	34.5
D^2VLM-3.8B (Ours)	**43.9**	**43.9**

Table 2. Performance comparison on general video-question-answering benchmarks.

Threshold	Grounding	Dense Captioning	
	Avg_{F1}	Avg_{F1}	Avg_{Sim}
0.4	40.9	36.2	20.9
0.5	41.7	37.1	21.3
0.6	42.3	37.5	21.8
0.7	42.1	35.6	21.2
0.8	39.7	31.5	19.9

Table 3. Threshold analysis on E.T. Bench data.

event matching and grounded video question answering. The results are shown in Tab. 1. It can be seen that our approach also outperforms the existing state of the art by large margins, further demonstrating its superiority.

C.2. Extension to General QA Tasks

We test our model on general video question answering benchmarks (MVBench [8] and Video-MME [4]). To enhance basic instruction-following capability, we incorporate automatically constructed multiple-choice questions during the proposed factorized preference optimization process. Due to our proposed factorized preference data synthesis, we can easily generate diverse distractor options based on different causes of failure and combine them with the original correct answer to form multiple-choice questions, without requiring additional external data sources.

As shown in Tab. 2, our method outperforms the grounding-focused counterpart E.T. Chat [11] and achieves results comparable to some general video understanding models (e.g., Video-LLaVA [10]) trained on large-scale generic data, but are usually less effective on grounding. We attribute the performance gap between our model and recent SOTA methods [1, 2] to the absence of large-scale generic pretraining and the relatively smaller model size. Incorporating such data and scaling up the model could further improve our framework. Meanwhile, it is also worth exploring how to train a model that can simultaneously achieve strong general reasoning and accurate temporal grounding.

C.3. Cost of Frame-Wise Similarity Calculation

Since the designed `<evi>` token involves additional frame-wise feature similarity computation for temporal grounding and visual semantic aggregation beyond the standard autoregressive decoder, it is natural to evaluate the associated computational cost. Such a frame-wise similarity calculation process is actually lightweight, taking less than 0.4 ms per token generation on a single 3090 GPU—only 1.4% of the total network forwarding time (29 ms).

C.4. Sensitivity Analysis on Similarity Threshold

As shown in Tab. 3, performance is relatively robust across different threshold values for salient frame identification during inference, and the intuitive choice of 0.5 already yields acceptable results. Overall, an overly high threshold causes information loss, while an overly low one introduces less relevant context. The best performance is achieved at a threshold of 0.6.

D. More about the Factorized Data Synthesis

As mentioned in the main paper, we mainly focus on two main factors: temporal event grounding and textual response, where each factor can be further categorized into multiple sub-factors. For temporal event grounding aspects, sub-factors include temporal localization shift, randomly adding or deleting grounded events (corresponding to the simulation of false positives and missed detection), and merging multiple events into one (corresponding to the simulation of a lack of fine-grained distinction in event boundaries). A full demonstration example can be found at Fig. 2. For textual response aspect, this type of perturbation modifies the semantic correctness of the textual response. It includes sub-types such as distorting key information, which disrupts critical content, and repeating responses, a common failure mode observed in video LLMs. Except for the repeating factor, we prompt an off-the-shelf LLM [15] to generate a distracted response based on the original correct event-level response.

E. Visualization Results

Here we provide qualitative results to better demonstrate the capability of our approach. Based on the input-output format, we categorize different tasks into three main types: (1) Dense captioning related task, which requires grounding multiple events throughout the entire video while also providing a textual description for each grounded event. (2) Grounding-focused task (e.g., temporal video grounding, action localization, etc.), which includes single-event grounding and multi-event grounding. (3) Temporally grounded video question answering, which involves answering the user's open-ended questions while also providing the temporal position of the answer (evidence). We also visualize the prediction result from the recent SOTA method [11] for comparison. Note that in the response from D^2VLM, all temporal information is derived from the generated `<evi>` token through the conversion process illustrated in the main paper. For each input, D^2VLM will first perform pure evidence grounding, followed by interleaved text-evidence generation (here we denote this as its actual response part). We show the converted time-involved text for both stages, where the actual response stage begins after the "Answer:" marker.

Dense captioning task. From Fig. 3, we can observe that: (1) Compared with the recent counterpart, our method can better localize the individual events. (2) Our method also generates more coherent and meaningful textual description, whereas the compared method often fails to do so and repeatedly generates similar content. These results essentially demonstrate the superiority of our approach in both event grounding and textual generation.

Grounding-focused task. The qualitative examples are shown in Fig. 4. It can be observed that: (1) Compared with recent SOTA method [11], our approach can localize the desired temporal event position more accurately (Fig. 4 (a)). (2) Our method also better distinguishes the boundaries between individual events, demonstrating its fine-grained discrimination capability (Fig. 4 (b)).

Temporally grounded video question answering. The example is shown in Fig. 5. It can be observed that: (1) Our method can correctly answer the question, while the compared method fails to follow the instruction given by the user (i.e., only responding to the temporal evidence position without answering the question). (2) Our method can also provide more reliable temporal evidence grounding.

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. 1, 2

[2] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29083–29095, 2025. 2

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

[4] 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. 2

[5] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024. 1

[6] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 1

[7] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 1

[8] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[9] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 1

[10] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 1, 2

[11] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.T. Bench: Towards open-ended event-level video-language understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1, 2, 3

[12] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 1

[13] 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. 1

[14] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1

[15] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 2

```
{
    "video": "activitynet/v_6uhLrPgbpUA.mp4",
    "duration": 46.333,
    “time_gt": {
            " evidence_grounding ": [[6.337, 12.098], [21.172, 24.413], [29.238, 38.095]],
            " interleaved_response ": [[6.337, 12.098], [21.172, 24.413], [29.238, 38.095]]
    }
    "conversations": [
        {
            "from": "human",
            "value": "<image>\nLocate all instances of the action 'bmx' in the video and give me the timestamps for each
                    occurrence."
        },
        {
            "from": "gpt",
            “value”: " <evi> <evi> <evi> </evi> The action happens in <evi>, <evi>, and <evi>."
        }
    ]
}
```


(a) Annotation example of grounding-focused task.


```
{
    "video": "how_to_step/-Yt51vMDbo0.mp4",
    "duration": 128.395,
    “time_gt": {
            "evidence_grounding": [[75, 83], [120, 128], [126, 128]],
            "interleaved_response": [[75, 83], [120, 128], [126, 128]]
    }
    "conversations": [
        {
            "from": "human",
            "value": "<image>\nLocalize a series of action steps in the given video, output the timestamps for each step,
                    and briefly describe the step."
        },
        {
            "from": "gpt",
            "value": "<evi> <evi> <evi> </evi> <evi>, remove the skin and meat pieces easily by running the knife along
         the base or cutting them off. <evi>, cut and dice the mango to use in smoothies, salads, or other
                   dishes. <evi>, enjoy your mangoes!."
        }
    ]
}
```


(b) Annotation example of dense captioning-related task.


```
{
    "video": "ego_timeqa/45ad1f4b-11f0-4ee2-aaf8-f06584d434c9_5_155.mp4",
    "duration": 150,
    “time_gt": {
            " evidence_grounding ": [8.823, 11.311],
            " interleaved_response ": [8.823, 11.311]
    }
    "conversations": [
        {
            "from": "human",
            "value": "<image>\nWatch the video carefully and answer the question: 'What did I do with the chopping
                    sticks?'. Your response should mention the timestamps as a reference. For example: '<answer>. The
                    event happens in <evi>'."
        },
        {
            "from": "gpt",
            “value”: " <evi> </evi> Picked them up. The relevant event happens in <evi>."
        }
    ]
}
```


(c) Annotation example of temporally grounded video question answering task.

Figure 1. Qualitative examples for grounding-focused task, dense captioning-related task, and temporally grounded video question answering task.

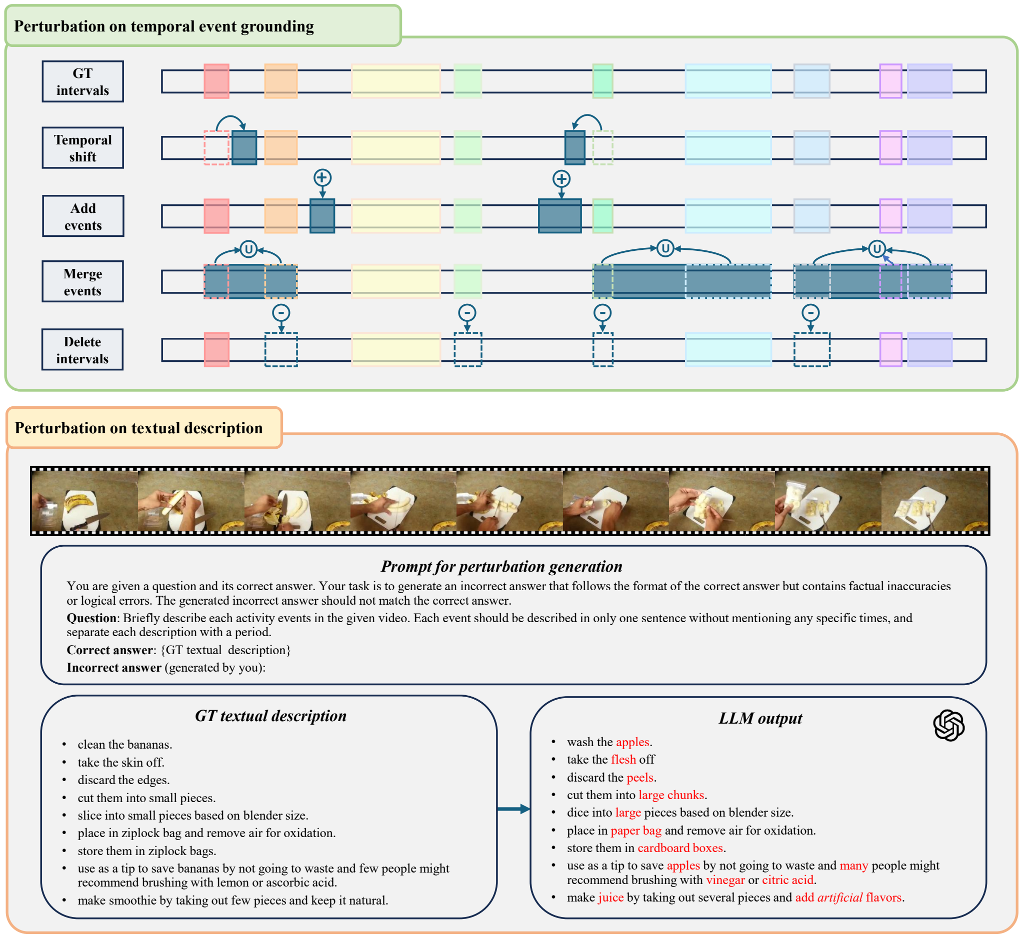


Figure 2. An illustrative example of the data synthesis approach.

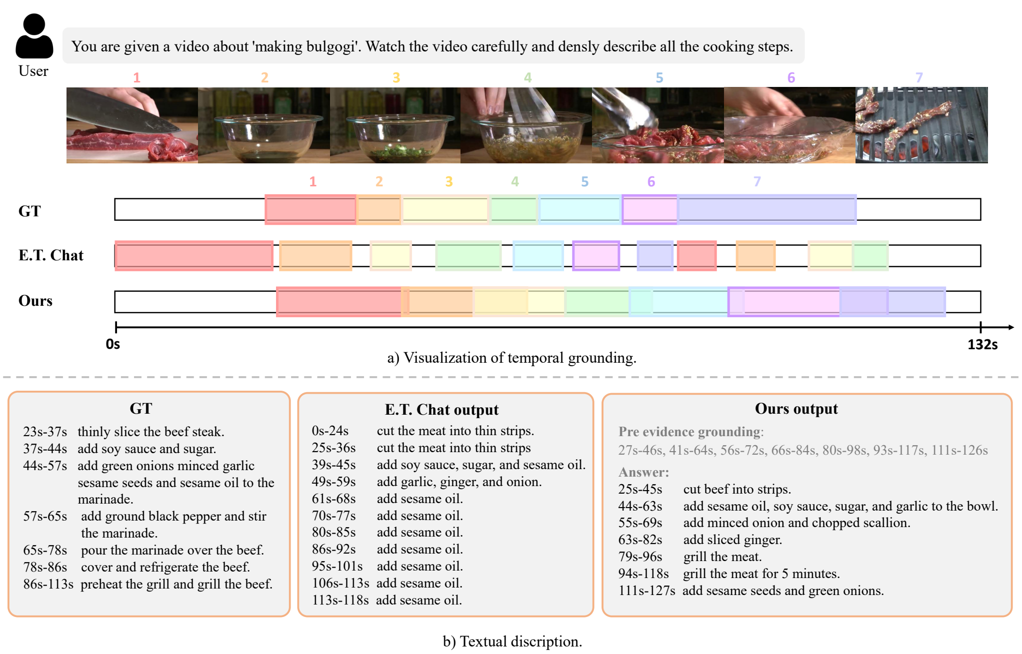
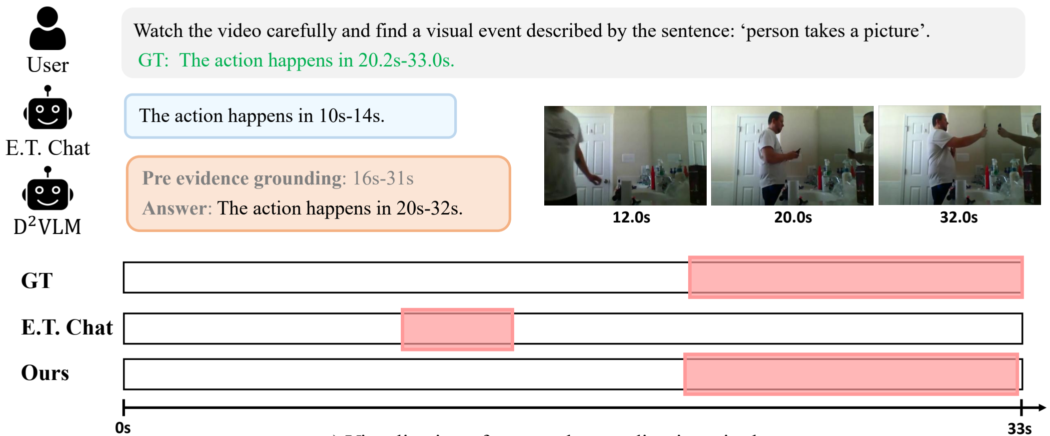
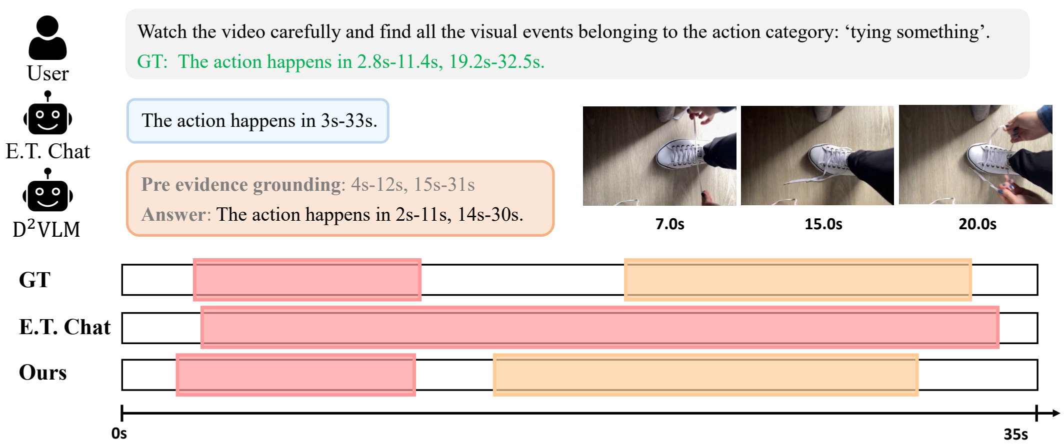


Figure 3. A qualitative example for dense captioning task.



a) Visualization of temporal grounding in a single event.



b) Visualization of temporal grounding in multiple events.

Figure 4. Qualitative examples for grounding-focused task.

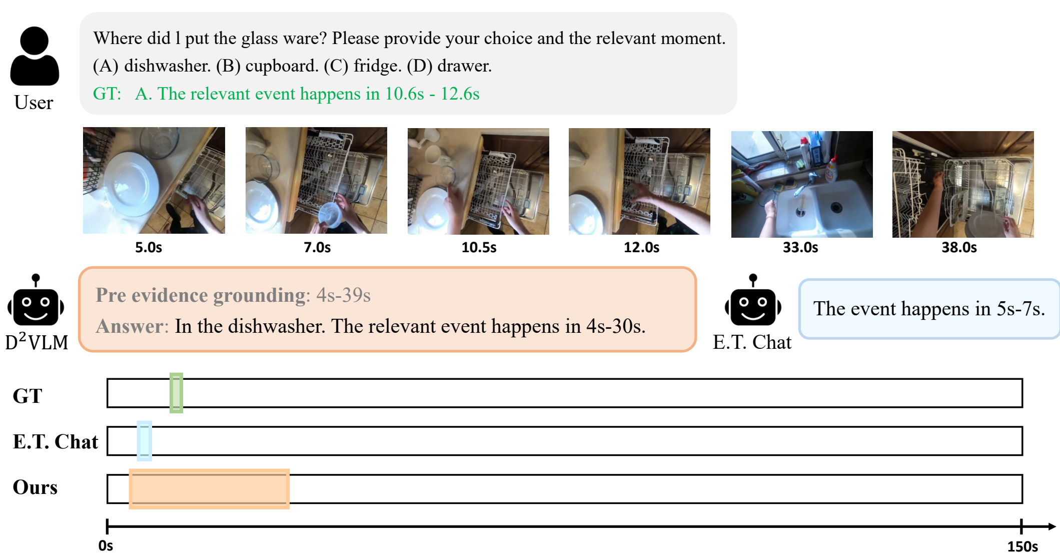


Figure 5. A qualitative example for temporally grounded video question answering task.