GroundVTS: Visual Token Sampling in Multimodal Large Language Models for Video Temporal Grounding

Supplementary Material

A. Influence on General VQA

While GroundVTS significantly improves temporal grounding accuracy, it is essential to evaluate whether its selective token sampling mechanism impacts the model's ability to handle general video understanding tasks. To investigate this, we assess GroundVTS on multiple subtasks of MVBench [9], a benchmark that comprehensively measures various video question-answering (VQA) capabilities, including temporal reasoning, object interaction, and scene-level comprehension.

As shown in Table 1, GroundVTS-Q achieves a slightly higher overall score (+0.8) compared to the base model Qwen2.5-VL-7B, indicating that the introduction of the VTS module does not compromise general VQA capability. The overall performance of GroundVTS-Q is competitive, with significant improvements in several tasks. The most notable gains appear in subtasks that focus on temporal reasoning and fine-grained action reasoning, such as Action Count (AC), Action Localization (AL) and Action Sequence (AS). In these tasks, GroundVTS-Q outperforms the base model by a substantial margin, with an improvement of 8.0 points in AC, 8.0 points in AL and 0.5 points in AS. These results align with the design objective of enhancing temporal sensitivity, confirming that VTS excels at capturing the temporal aspects of video data.

Meanwhile, in tasks that require a more general understanding of the scene or object interactions, such as Scene Transition (ST) and State Change (SC), GroundVTS-Q maintains competitive performance. For example, GroundVTS-Q achieves a slight drop in ST (-1.0) and SC (-2.0) compared to the base model, but still performs well overall. This suggests that the selective filtering mechanism of the VTS module does not undermine the model's ability to grasp global scene awareness or appearance-related information.

In summary, the results suggest that the introduction of VTS enhances the model's performance on tasks requiring fine-grained temporal reasoning, while maintaining strong performance in more general video understanding tasks. This balance between focused temporal sensitivity and general visual understanding demonstrates the versatility and effectiveness of the GroundVTS framework.

B. Out-of-distribution Data Experiment

To further evaluate the robustness and generalization ability of GroundVTS, we conduct out-of-distribution (OOD) experiments on three benchmarks: DiDeMo [1], LongVideoBench [15], and NExT-GQA [16]. None of these datasets are included in the fine-tuning data of either GroundVTS or its base models (Qwen2.5-VL and InternVL3.5). This evaluation examines whether our method can effectively transfer its temporal grounding capability to unseen domains and tasks.

Results on DiDeMo. As shown in Table 2, GroundVTS-Q achieves substantial gains over both the instruction-tuned baseline QwenVL-G and the pretrained Qwen2.5-VL. Specifically, it improves by +13.7 R1@0.3, +10.1 R1@0.5, and +8.0 mIoU, establishing new SOTA performance on two of the three metrics. These results highlight the strong cross-domain adaptability brought by our query-guided VTS mechanism. On InternVL3.5-8B, GroundVTS-I also produces consistent improvements over InternVL-G, yielding gains of +0.9 R1@0.3, +1.3 R1@0.5, and +1.1 mIoU. While the improvements are smaller than those observed with Qwen2.5-VL, they confirm that GroundVTS remains effective even when the underlying base model already possesses strong temporal reasoning capability.

Results on LongVideoBench. LongVideoBench evaluates temporal reasoning on significantly longer videos, with durations ranging from tens of seconds to several minutes. As shown in Table 3, GroundVTS-I achieves the best accuracy on two of the three duration ranges, reaching 65.6% on (8,15]s and 68.0% on (15,60]s, and remains competitive on (180,600]s. These results suggest that the proposed query-guided temporal selection mechanism can effectively scale to long-video scenarios by filtering irrelevant temporal regions and concentrating computation on query-related segments.

Results on NExT-GQA. NExT-GQA evaluates both temporal grounding and question answering accuracy, requiring models to first localize relevant temporal segments before answering questions. As shown in Table 4, GroundVTS-Q achieves the best mIoU (25.8) among all compared methods and remains competitive on other grounding-related metrics. Notably, the strongest baseline, TOGA, is a classical expert model specifically designed for grounded video question answering, whereas GroundVTS is built upon a general-purpose multimodal large language model without task-specific architecture. Despite this difference, GroundVTS-Q still attains comparable performance across most metrics, demonstrating that the proposed query-guided temporal selection mechanism can effectively transfer its temporal localization capability to reasoning-

Table 1. Comparison with base model on MVBench subtasks.

Model	AA	AC	AL	AS	EN	ER	FGA	OI	OS	ST	SC	all
Qwen2.5-VL-7B	77	39	38	69.7	30	50	44.5	66.5	35.5	90.5	50.5	53.7
GroundVTS-Q	75.5	47	46	70.2	28.5	46.5	41	63.5	43	89.5	48.5	54.5

Table 2. Comparison with state-of-the-art methods on DiDeMo test splits.

Model	R1@0.3	R1@0.5	mIoU
Video-LLaMA [17]	20.1	8.2	14.3
Video-ChatGPT [10]	19.8	6.5	13.7
Valley	33.2	13.4	21.8
VideoChat [8]	34.5	14.5	22.4
Momenter	38.2	21.8	26.5
VTimeLLM [5]	45.0	28.8	27.9
TimeChat [13]	42.8	24.4	28.2
HawkEye	44.8	**29.7**	29.5
Qwen2.5-VL-7B	28.7	22.7	22.2
QwenVL-G	32.6	17.7	22.0
GroundVTS-Q	**46.3**$_{(\uparrow 13.7)}$	27.8$_{(\uparrow 10.1)}$	**30.0**$_{(\uparrow 8.0)}$
InternVL3.5-8B	29.5	23.9	23.0
InternVL-G	36.6	21.0	23.1
GroundVTS-I	37.5$_{(\uparrow 0.9)}$	22.3$_{(\uparrow 1.3)}$	24.2$_{(\uparrow 1.1)}$

The baseline results are from reference [3].

Table 3. Comparison with state-of-the-art methods on LongVideoBench test splits (Acc).

Model	(8, 15]s	(15, 60]s	(180, 600]s
VideoTree [14]	61.0	57.5	48.4
VideoMiner [2]	65.1	64.7	**58.6**
GroundVTS-Q	52.9	60.5	44.2
GroundVTS-I	**65.6**	**68.0**	52.4

intensive video QA tasks.

Overall, the OOD evaluation across three diverse benchmarks demonstrates that GroundVTS generalizes well to unseen datasets and tasks, including short video grounding (DiDeMo), long-video reasoning (LongVideoBench), and grounded video question answering (NExT-GQA). These results reinforce the robustness of the proposed framework and its ability to capture transferable fine-grained temporal cues beyond the training distribution.

C. Parameter-free projection

We evaluate a parameter-free relevance estimation method (Table 5). Without additional training, this approach leads to a substantial performance drop due to the mismatch between the sampled-token distribution and the pretrained LLM (w/o training). To alleviate this issue, we further fine-tune the LLM following Stages 2&3 to adapt to the sampled-token distribution (w/ training), which partially recovers the performance. However, it still underperforms the full GroundVTS model with learned relevance projections.

Table 4. Comparison with state-of-the-art methods on NExT-GQA test splits.

Model	mIoU	mIoP	IoU@.5	IoP@.5	Acc@GQA
TOGA [11]	24.4	**40.5**	**21.1**	**40.6**	**24.6**
VidStreaming [12]	19.3	32.2	13.3	31.0	17.8
GroundVTS-Q	**25.8**	37.4	20.4	35.4	23.2
GroundVTS-I	16.7	26.5	11.9	24.3	18.5

D. Dataset Ablation

GroundVTS-Q is trained on the LLaVA-Video-178K and our constructed Grounding-FT datasets. For a fair comparison, the base Qwen2.5VL-7B is trained on the same datasets under three settings, as reported in Table 6: (i) *Qwen-G*, trained only on the Grounding-FT dataset; (ii) *Qwen-(L+G)*, trained on the concatenation of the two datasets; and (iii) *Qwen-(L→G)*, trained following the same Stage 2→Stage 3 curriculum. Note that Stage 1 (VTS warm-up) is not applicable to the base model.

As shown in Table 6, GroundVTS-Q consistently outperforms the Qwen baselines across all evaluation settings. In particular, GroundVTS-Q achieves 50.1 mIoU on Charades-STA, significantly surpassing Qwen-G with 31.7, Qwen-(L+G) with 28.5, and Qwen-(L→G) with 29.8. These results suggest that the proposed grounding-aware training strategy effectively improves performance under matched data and training configurations.

E. Frame Sampling Sensitivity of InternVL3.5

To examine whether the frame density sensitivity is specific to Qwen2.5-VL or reflects a more general phenomenon, we conduct an additional experiment using InternVL3.5. Unlike Qwen2.5-VL, InternVL3.5 adopts a fixed-number frame sampling strategy. Therefore, we vary the number of sampled frames to analyze how visual token density affects VTG performance.

Figure 1 presents the frame sensitivity results of InternVL3.5 on the QVHighlights dataset. Similar to the trend observed with Qwen2.5-VL, the performance again exhibits

Table 5. Parameter-free token sampling vs. GroundVTS.

VTS	Training	Charades-STA				ActivityNet-Captions			
		R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
✓	✓	**71.5**	**57.5**	**34.2**	**50.1**	**51.3**	**33.6**	**21.4**	**36.0**
–	✓	69.6	52.7	29.6	47.5	38.8	25.0	13.7	27.8
–	–	21.2	13.6	6.8	14.5	9.1	5.1	2.6	6.6

Table 6. Dataset ablation on Charades-STA and ActivityNet-Captions test spilt.

Variant	Charades-STA				ActivityNet-Captions			
	R1 @.3	R1 @.5	R1 @.7	mIoU	R1 @.3	R1 @.5	R1 @.7	mIoU
Qwen-G	45.2	32.7	18.7	31.7	40.6	23.9	9.9	26.7
Qwen-(L+G)	41.1	27.7	15.7	28.5	39.1	20.6	7.8	24.9
Qwen-(L→G)	42.5	30.7	16.9	29.8	40.0	22.1	8.6	25.9
GroundVTS-Q	**71.5**	**57.5**	**34.2**	**50.1**	**51.3**	**33.6**	**21.4**	**36.0**

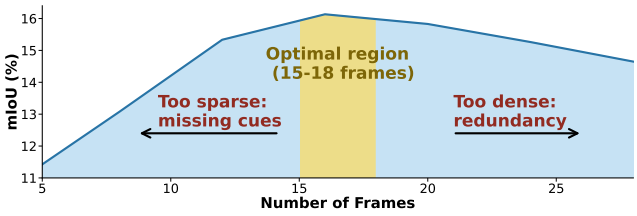


Figure 1. Frame sensitivity of InternVL3.5 on QVHighlights.

a clear non-linear dependency on frame density. Increasing the number of sampled frames initially improves performance by providing richer temporal cues. However, beyond a certain point, further increasing the frame count leads to diminishing returns and eventually performance degradation, suggesting that excessive visual tokens introduce redundancy and interfere with effective temporal reasoning.

These results indicate that the sensitivity to visual token density is not limited to a specific model architecture, but appears to be a general characteristic of multimodal LLM-based VTG systems. This observation further supports our motivation for designing an adaptive token sampling mechanism in GroundVTS.

F. Additional Analysis of Visual Token Density

Table 7 provides the full quantitative results corresponding to the visual token density analysis discussed in the main paper. The results further substantiate the trends previously observed. For the pretrained Qwen2.5VL-7B, performance grows steadily as token density increases, but drops rapidly in sparse conditions. This confirms its heavy dependence on dense temporal evidence: when the effective density falls below 1.0, all metrics decrease sharply (e.g., R1@0.5 drops from 47.1 to 18.8 as density reduces from 2.0 to 1.0). The fine-tuned QwenVL-G shows improved overall accuracy but remains highly sensitive to token density.

In contrast, GroundVTS-Q demonstrates remarkable stability across all density levels. At extremely sparse levels (e.g., density 0.2–0.6), its performance remains comparable to the best performance of its base model, avoiding the sharp degradation observed in QwenVL-G. As the density increases, the improvement of GroundVTS-Q is much more gradual, forming a plateau rather than a steep curve. This consistency appears across all evaluation metrics, including R1@0.3, R1@0.5, R1@0.7, and mIoU, illustrating that GroundVTS effectively mitigates the vulnerability of Vid-LLMs to insufficient visual tokens. These results further reinforces the conclusion that our query-guided sampling mechanism yields reliable grounding accuracy regardless of input density, while base Vid-LLMs suffer substantial degradation when token budgets are reduced.

G. Training Details

This section provides the detailed configurations and parameters used for training GroundVTS across its different stages, as well as the parameter values for the model variants. Table 8 outlines the settings for each stage of training, including the learning rate, optimizer, batch size, and other critical training details. Table 9 lists the total and trainable parameters for both GroundVTS-Q (Qwen2.5VL-based) and GroundVTS-I (InternVL-based) models.

H. Grounding-FT Dataset

Grounding-FT is a curated dataset designed for instruction fine-tuning on Video Temporal Grounding (VTG) tasks. It aggregates the training splits of Charades-STA [4], QVHighlights [7], and ActivityNet-Captions [6], resulting in 70K annotated clips paired with instruction-style queries. The goal is to unify multiple VTG formulations under a consistent question-answering (QA) framework, facilitating language model training with natural conversational inputs rather than fixed task templates.

H.1. Overview and Construction

Grounding-FT covers two main VTG task types:

Table 7. Quantitative analysis of visual token density on Charades-STA test split.

FPS $*$ ρ	Qwen2.5VL-7B				QwenVL-G				GroundVTS-Q			
	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
0.2	22.7	13.3	6.8	16.1	28.0	16.5	8.3	18.7	61.2	43.6	23.3	41.0
0.4	23.5	13.5	6.6	16.4	33.0	20.2	10.2	21.8	67.1	50.8	29.2	46.0
0.6	26.5	14.1	6.8	17.6	36.2	24.0	12.6	24.9	69.5	54.4	32.6	48.2
0.8	30.3	16.4	7.2	19.7	41.4	28.3	15.7	28.4	70.9	56.8	33.6	49.6
1.0	34.2	18.8	8.6	22.1	45.2	32.7	18.7	31.7	71.5	57.5	34.2	50.1
1.2	36.8	21.4	9.1	24.2	49.3	36.6	22.0	35.2	72.3	58.4	34.8	50.7
1.4	42.2	25.0	10.6	28.0	54.8	40.8	24.5	38.9	72.9	58.5	35.3	51.0
1.6	48.6	29.1	12.5	32.2	62.7	44.6	24.8	42.8	72.8	58.3	34.7	50.8
1.8	59.5	37.0	16.6	38.7	72.2	53.0	27.4	48.2	73.0	58.3	34.5	50.9
2.0	68.8	47.1	23.5	45.4	74.4	56.3	30.6	50.0	72.8	58.5	34.7	50.9

Table 8. Training configuration for each stage of the GroundVTS model.

Stage	Trainable Modules	Learning Rate	Optimizer	Batch Size (per GPU)	Grad. Acc. Steps	Epochs	LoRA Config	Dataset
Stage 1: VTS Warm-up	VTS	1e-5		2	4	1	–	LLaVA-Video-178K
Stage 2: Joint LoRA Adaptation	LLM (LoRA) + VTS + Projector	2e-4	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$	2	4	2	rank = 8, $\alpha = 16$, dropout = 0.05	LLaVA-Video-178K
Stage 3: Grounding Fine-tuning	LLM (LoRA) + VTS + Projector	1e-4		2	4	3	rank = 8, $\alpha = 16$, dropout = 0.05	Grounding-FT

Table 9. Parameter statistics for GroundVTS-Q and GroundVTS-I models.

Model	Total Params	Trainable Params			
		VTS	Projector	LoRA	All
GroundVTS-Q	8.32B	29.4M	44.6M	79.0M	153.0M
GroundVTS-I	8.56B	34.6M	33.6M	77.0M	145.2M

(a) Moment Retrieval (MR)—identifying the temporal segment in a video that corresponds to a given natural language query.

(b) Highlight Detection (HD)—output all salient moments relevant to the query in the video together with their corresponding relevance scores.

For MR, we aggregate annotations from the training splits of Charades-STA, QVHighlights, and ActivityNet-Captions. For HD, we use the training split of QVHighlights. All samples are reformulated into an instruction-response style and stored in the `ShareGPT` format, where each instance contains a conversational pair between a `user` (prompt) and an `assistant` (answer), along with the corresponding video path. To enhance linguistic diversity and improve generalization to natural language instructions, we construct a pool of prompt templates and randomly select one for each instance rather than relying on a single fixed phrasing. This variation helps the fine-tuned model better adapt to free-form human queries. Note that timestamp information is not provided in the text prompt, and all models must rely on the positional encodings of visual tokens to infer temporal information.

H.2. Moment Retrieval Task

Each MR training instance contains at least the video name `<video>`, a query phrase {`query`}, and the ground-truth {`start`} and {`end`} timestamps. We construct diverse instruction templates and randomly sample one for each example to enhance linguistic variability. The prompt templates and expected output format are summarized in Table 10. Examples before and after the conversion are as follows:

Example 1 (Charades-STA).

Original annotation:

```
Y6R7T 20.8 30.0##person start playing on their
    phone.
```

Reformatted instance:

```
{
  "messages": [
    {"role":"user",
     "content":"<video>At what point in the
         video did the following events occur:
         person start playing on their phone.
         Output the start and end timestamps."},
```

Table 10. Prompt templates and output format for the MR task.

Type	Content
Prompt Templates	<video>At what point in the video did the following events occur: {query}? Output the start and end timestamps. <video>What is the location of the moment: {query}? <video>Find when the following event happens in the video: {query}. Give me the start and end times. <video>Please indicate the start and end timestamps for the event: {query}. <video>Please predict start and end time of the following moment: {query}. <video>During which time interval does this happen in the video: {query}? <video>Locate the moment in the video where this occurs: {query}. Provide start and end times. <video>For the video, when does this event take place: {query}? Answer with start and end timestamps. <video>I want to know the start and end times of the following event in the video: {query}. <video>Could you tell me from what time to what time this happens: {query}? <video>Can you tell me the time window of this event: {query}? <video>Please find the timestamps that mark the occurrence of this event: {query}. <video>Identify the start and end of the following event in the video: {query}.
Expected Output	from {start}s to {end}s

```
    {"role":"assistant",
     "content":"from 20.8s to 30.0s"}
  ],
  "videos":["Y6R7T.mp4"]
}
```

Example 2 (ActivityNet-Captions).

Original annotation (compact):

```
{"video_id":"v_nwznKOuZM7w",
 "timestamps":[[0,15.51],[11.39,31.65]],
 "sentences":[
   "Two people are seen moving around a kitchen
       quickly performing various tasks and
       sitting down.",
   "They then wax down a ski in the kitchen
       while continuing to move around."
 ]}
```

Reformatted instances:

```
{"messages":[
  {"role":"user",
   "content":"<video>What is the location of the
       moment: Two people are seen moving around
       a kitchen quickly performing various
       tasks and sitting down."},
  {"role":"assistant","content":"from 0s to
      15.51s"}
 ],
 "videos":["v_nwznKOuZM7w.mp4"]},

{"messages":[
  {"role":"user",
   "content":"<video>Can you tell me the time
       window of this event: They then wax down
       a ski in the kitchen while continuing to
       move around."},
  {"role":"assistant","content":"from 11.39s to
      31.65s"}
 ],
 "videos":["v_nwznKOuZM7w.mp4"]}
```

H.3. Highlight Detection Task

Each HD training instance contains at least the video name, a query phrase, relevant temporal segments, and their saliency scores. The prompt templates and output format are summarized in Table 11. Specific examples are as follows:

Example (QVHighlights).

Original annotation (compact):

```
{"qid":10016,
 "query":"Man in baseball cap eats before doing
     his interview.",
 "vid":"j7rJstUseKg_210.0_360.0",
 "relevant_windows":[[96,114]],
 "saliency_scores":[
   [2,3,3],[4,3,2],[2,3,1],[2,3,0],
   [2,3,3],[2,3,2],[2,3,1],[2,3,0],[1,3,3]
 ]}
```

Reformatted instance:

```
{
  "messages": [
    {"role":"user",
     "content":"<video>Please highlight the most
         exciting parts related to: Man in
         baseball cap eats before doing his
         interview."},

    {"role":"assistant",
     "content":"The highlights are: important
         from 96.0s to 98.0s, 98.0s to 100.0s,
         104.0s to 106.0s; less important from
         100.0s to 102.0s, 102.0s to 104.0s,
         106.0s to 108.0s, 108.0s to 110.0s,
         110.0s to 112.0s, 112.0s to 114.0s"}
  ],
```

Table 11. Prompt templates and output format for HD task.

Type	Content
Prompt Templates	<video>Please highlight the most exciting parts related to: {query}. <video>Find the most relevant or important moments for: {query}. <video>Which moments in the video best reflect: {query}? <video>Highlight the key segments that correspond to: {query}. <video>Show the most interesting clips about: {query}. <video>What are the highlight moments for: {query}? <video>Mark the time intervals that are most significant for: {query}.
Expected Output	The highlights are: very important from {start}s to {end}s, ...; important from {start}s to {end}s, ...; less important from {start}s to {end}s, ...

```
  "videos":["j7rJstUseKg_210.0_360.0.mp4"]
}
```

Based on the above methods, Grounding-FT reformulates heterogeneous VTG annotations into unified, instruction-response pairs. The diversity of prompt phrasing and conversational structure better aligns the dataset with large language model fine-tuning paradigms, leading to improved robustness and generalization.

I. Discussion on Early vs. Multi-Stage Token Sampling

In this work, VTS performs query-guided token sampling before multimodal fusion. We adopt this design because it is simple, efficient, and well aligned with the VTG setting: early suppression of query-irrelevant visual content helps reduce noise before constructing the joint representation. This also makes the sampling behavior more interpretable, since token relevance is estimated directly from the text query and visual features prior to deeper cross-modal interactions. At the same time, we acknowledge that later-layer or multi-stage sampling could leverage richer multimodal semantics and potentially improve token selection further. Such designs may offer a different trade-off between efficiency, interpretability, and representational power. We view this as an interesting direction for future work.

J. Additional Qualitative Analysis

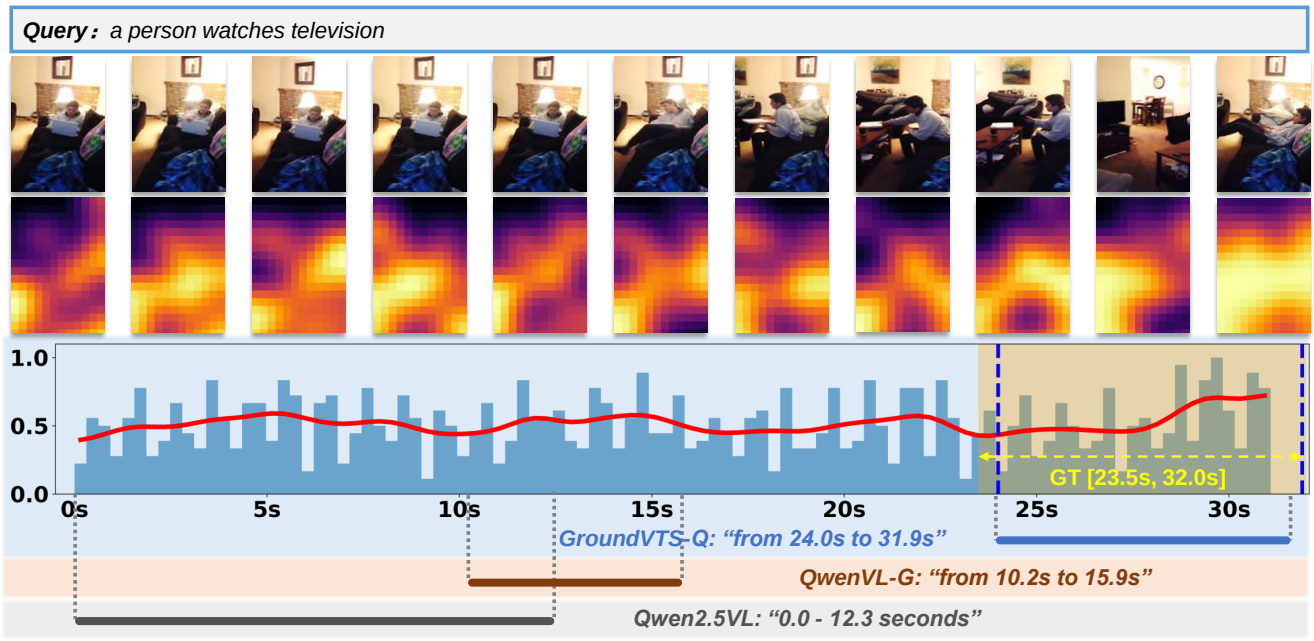
To complement the qualitative study, we provide additional visualization examples for both GroundVTS-Q and GroundVTS-I. All examples follow the same visualization format as Figure 5 in the main paper, where the bottom curve denotes the normalized token density produced by the VTS module, with higher peaks indicating segments the model regards as more relevant to the query.

Across these cases, GroundVTS consistently exhibits highly accurate temporal localization. In Figure 2(a) (GT: 23.5–32.0 s), GroundVTS-Q predicts 24.0–31.9 s, aligning almost perfectly with the ground-truth, whereas QwenVL-G and the pretrained Qwen2.5VL shift the interval far earlier and fail to localize the correct moment. A similar trend appears in Figure 2(b) (GT: 0.0–5.9 s), where GroundVTS-Q outputs 0.0–5.8 s with near-exact precision, while both baselines truncate or deviate from the target boundary.

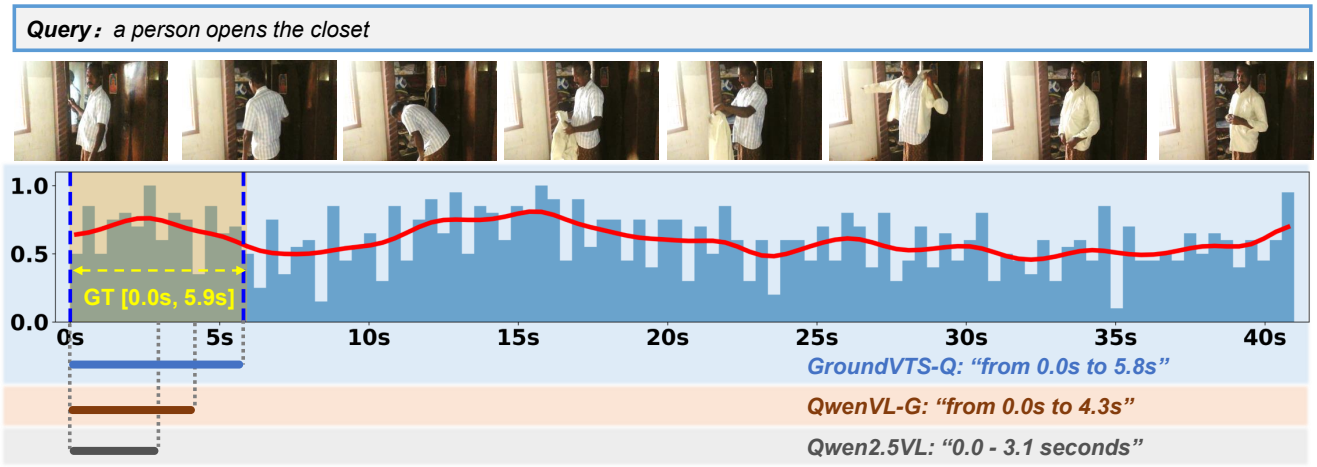
The InternVL-based examples exhibit the same trend. In Figure 3(a) (GT: 22.3–30.9 s), GroundVTS-I produces a tightly aligned prediction of 22.0–31.0 s, whereas InternVL-G shortens and shifts the interval (18.0–26.0 s), and the base InternVL3.5 mislocalizes the event to a distant region. In Figure 3(b) (GT: 0.0–7.0 s), GroundVTS-I again matches the ground-truth boundaries accurately (0.0–6.9 s), while the baseline models either overextend the span or capture only a partial portion of the event.

These visualizations reveal three consistent advantages of GroundVTS. First, its temporal predictions are markedly more accurate and better aligned with annotated spans, regardless of model backbone or event duration. Second, its predicted spans consistently fall within the regions where the sampled token density reaches (local) maxima. In every example, the model's final prediction aligns with the peaks of the VTS density curve, indicating that GroundVTS relies on the most informative temporal segments identified by the sampling module. Third, the token allocation patterns produced by VTS are adaptive across different scenarios. In some cases (as illustrated in Figure 5 of the main paper), the density distribution forms sharp peaks with strong contrasts between attended and suppressed segments, typically corresponding to short or well-isolated grounding moments. In other cases, the differences between peaks and valleys are more moderate; nevertheless, the VTS curve still places a clear relative emphasis on the correct temporal region. These variations indicate that VTS does not rely on a fixed sparsity pattern but adjusts its sampling behavior according to the temporal structure of each video-query pair.

In addition, Figure 2(a) visualizes the spatial distribution of visual tokens selected by VTS. It can be observed that VTS mainly focuses on the middle and lower regions of the frames, which correspond to areas around human activities. When the action of watching TV occurs at the end of the



(a)



(b)

Figure 2. Additional qualitative comparison between GroundVTS-Q and its base models. Example (a) additionally illustrates spatial token retention maps, which correspond to spatial token selections.

video, the VTS module attends to most regions of the frame.

Overall, by concentrating tokens at the most semantically relevant moments while downweighting redundant frames, VTS enables GroundVTS to encode fine-grained temporal cues more effectively, leading to significantly sharper and more accurate temporal boundaries.

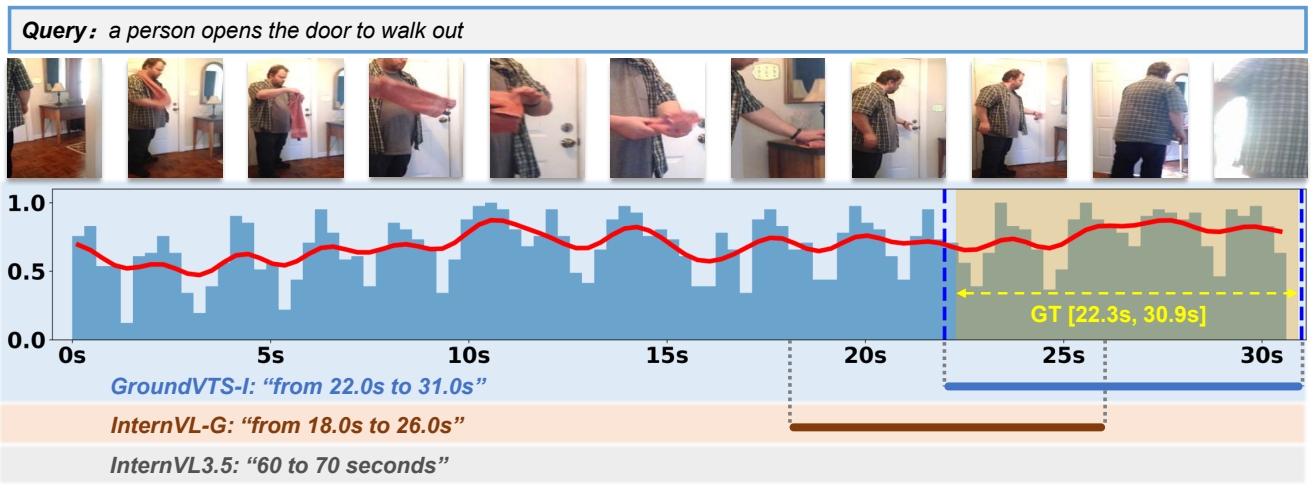
References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 1

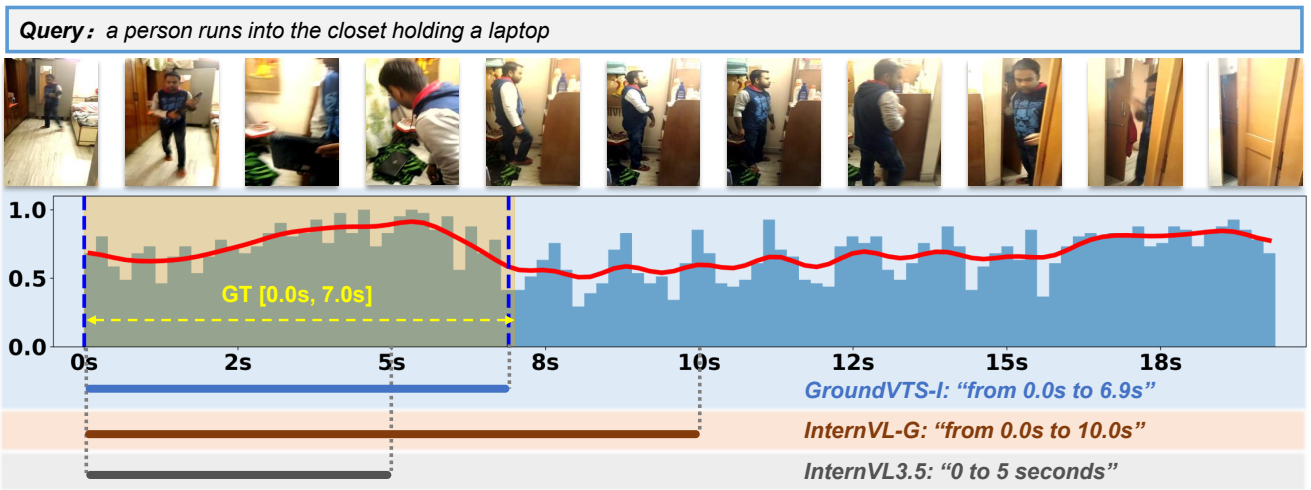
[2] Xinye Cao, Hongcan Guo, Jiawen Qian, Guoshun Nan, Chao Wang, Yuqi Pan, Tianhao Hou, Xiaojuan Wang, and Yutong Gao. Videominer: Iteratively grounding key frames of hour-long videos via tree-based group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23773–23783, 2025. 2

[3] Pengcheng Fang, Yuxia Chen, and Rui Guo. When and what: Diffusion-grounded videollm with entity aware segmentation for long video understanding. *arXiv preprint arXiv:2508.15641*, 2025. 2

[4] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In



(a)



(b)

Figure 3. Additional qualitative comparison between GroundVTS-I and its base models.

Proceedings of the IEEE international conference on computer vision, pages 5267–5275, 2017. 3

[5] 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. 2

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[7] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 3

[8] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[9] 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. 1

[10] 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. 2

[11] Your Name and Coauthor Name. Toga: Temporally grounded open-ended video qa with weak supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[12] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neu-*

ral Information Processing Systems, 37:119336–119360, 2024. 2

[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. 2

[14] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 2

[15] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 1

[16] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 1

[17] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 2