

1 Enhancing Temporal Understanding in Video-LLMs through Stacked 2 Temporal Attention in Vision Encoders - Supplementary Document

3 A Prompt Examples

4 This section provides example prompts used for our experiments.

5 Box 1 illustrates the prompt structure used in our few-shot experiments on the SSv2-T10 dataset. In
6 this setup, the model is presented with one or more in-context examples, each consisting of a video
7 and its corresponding action label. After reviewing these examples, the model is tasked with inferring
8 the action occurring in a new query video. This format encourages the model to generalize from the
9 provided samples and demonstrate its capability to recognize similar actions.

Box 1: Example prompt with in-context examples for SSv2-T10 dataset

Instruction: Look at the provided examples and answer the last question.

Example 1 - <video> The action happening in this video is: Moving [something] from left to right.

Example 2 - <video> The action happening in this video is: Moving [something] from right to left.

Final Prompt - <video> Now considering the previous examples, what action is happening in this video?

10

11 Box 2 presents the prompt format used for the SSv2-VSM dataset, which is designed to evaluate the
12 model’s ability to perform similarity matching. The model is provided with multiple labeled video
13 examples and a final query video, and it must determine whether any of the examples depict an action
14 related to the one shown in the query. If a related action exists, the model is expected to return the
15 example number along with the action label; otherwise, it should respond with "No related action."
16 This task format emphasizes fine-grained action discrimination and serves as a valuable test of the
17 model’s capacity for visual-semantic matching.

Box 2: Example prompt for SSv2-VSM dataset

Instruction: Look at the provided examples and identify which example is related to the final video.

Example 1 - <video> The action happening in this video is: Moving [something] away from [something].

Example 2 - <video> The action happening in this video is: Moving [something] closer to [something].

Final Prompt - <video> Now considering the previous examples, is there any action related to this video? If not, respond with 'No related action' and if there is, respond with the example number and action.

18

19 In open-ended generation, the model’s output may not precisely match the expected action name. To
20 address this, we employ another LLM, Qwen2-7B, as a judge to evaluate the responses generated
21 by Video-LLMs. Box 3 provides the evaluation prompt. Given a ground truth label and the model’s
22 predicted answer, the judge determines whether the prediction is correct, returning a binary “Yes” or
23 “No” response.

Box 3: Example Prompt for Evaluation

Instruction: Look at the ground truth and the LLM’s answer. Decide whether the LLM’s answer matches the ground truth.

Ground Truth: Pulling [something] from left to right

LLM Answer: The action is moving something from right to left on the floor

Question: Based on the ground truth, is the LLM answer correct? Answer with a simple "Yes" or "No".

B Ablation Study

To assess the impact of temporal attention block placement and the number of attention heads in the vision encoder of STAVEQ2, we conducted an ablation study on STAVEQ2 2B, fine-tuned on the SSv2-T10 dataset. We evaluated configurations with temporal attention blocks placed either before or after spatial attention within transformer blocks and varied the temporal attention head scaling factor relative to the baseline number of heads. Results are summarized in Table 1.

Table 1: Ablation study results for temporal attention block placement and head scaling in STAVEQ2 2B on SSv2-T10. Accuracy (%) is reported, with the best result in bold.

Model	Atten. Block Order	Head Scale	Acc (%)
Qwen2-VL 2B	–	1.0	73.14
STAVEQ2 2B	Spatial First	1.0	58.34
STAVEQ2 2B	Spatial First	0.5	71.18
STAVEQ2 2B	Temporal First	0.25	73.20
STAVEQ2 2B	Spatial First	0.25	76.04

Positioning temporal attention after spatial attention with a head scaling factor of 0.25 achieves the highest accuracy (76.04%), outperforming the baseline Qwen2-VL 2B (73.14%) by 2.90%. Reducing the number of temporal attention heads (e.g., 0.25 vs. 1.0) enhances performance, likely due to improved regularization and focus on critical temporal features, particularly when there is limited data, aligning with our emphasis on parameter efficiency. Placing temporal attention before spatial attention (73.20% at 0.25 scale) yields slightly lower performance, indicating that processing spatial context first enhances temporal modeling. These results validate the design of our temporal attention blocks, especially for fine-grained temporal reasoning tasks.

Table 2 highlights the impact of applying STAVE enhancements to the InternVideo2-Chat 8B model, fine-tuned on the SSv2-T10 dataset. The baseline InternVideo2-Chat 8B model achieves an accuracy of 84.17%. With the integration of stacked temporal attention, the performance improves significantly, reaching 95.18%—a substantial gain of 11.01%. This result underscores the effectiveness of our method in boosting the temporal understanding capabilities of Video-LLMs.

Table 2: Performance of InternVideo2-Chat 8B, fine-tuned on SSv2-T10. Adding stacked temporal attention (STA) leads to a significant accuracy gain.

Method	Acc (%)
InternVideo2-Chat 8B	84.17
InternVideo2-Chat 8B + STA	95.18 (↑ 11.01%)

44 C SSv2-VSM Dataset Composition

45 As described in the paper, each sample in SSv2-VSM dataset consists of two reference videos (with
 46 different actions) and a third query video. The task is to determine whether the action in the query
 47 video matches the first, the second, or neither video. For the SSv2-VSM dataset, we explored the
 48 optimal ratio of positive to negative samples for fine-tuning the models. Positive samples consist
 49 of reference videos where one matches the query video’s action, while negative samples have no
 50 matching actions with the query video.

51 Table 3 shows that increasing the proportion of positive samples generally enhances performance,
 52 with similarity matching accuracy improving from 25.52% at 50% positive samples to a peak of
 53 71.25% at 80% positive samples. However, accuracy drops to 49.18% at 91% positive samples,
 54 indicating that an excessively high proportion of positive samples may reduce dataset diversity and
 55 hinder generalization.

Table 3: Context-Selection fine-tuning results with varying positive-negative sample ratios. Similarity Matching evaluates whether the model can correctly identify a relevant context sample among distractors when such a sample is present.

Dataset Composition	Accuracy (%)
50% positive	25.52
80% positive	71.25
91% positive	49.18

56 In our experiments reported in Table 3, we initially included textual descriptions of the actions
 57 happening in each reference video. However, removing these textual descriptions increases task
 58 difficulty, as the similarity matching accuracy at 80% positive samples drops from 71.25% to 68.65%
 59 by removing the descriptions. Consequently, to evaluate the models on a more challenging task, we
 60 excluded textual descriptions for reference videos in the main experiments reported in the paper.

61 D Prompting Experiments

62 We conducted additional controlled experiments to evaluate how prompt structure and intra-context
 63 formatting affect model performance in experiments on the SSv2-VSM dataset. As shown in Table 4,
 64 we tested four prompting strategies: No-Instruction, where the task is given no explicit instruction;
 65 Introductive, where a task description appears at the start of the prompt; Introductive-Summative,
 66 which adds a summary instruction after the context samples; and Intra-demonstration, where the
 67 instruction is repeated before each context item. Results indicate that most models are largely
 68 unaffected by instruction style.

Table 4: Prompt structure impact on SSv2-VSM performance. We compare four prompting styles: No-Instruction (no task description), Introductive (task instruction at the beginning), Introductive-Summative (instructions both at the beginning and after context), and Intra-demonstration (instruction repeated before each context sample).

Model	No-Instruction	Introductive	Introductive-Summative	Intra-demonstration
LLaVA-NeXT-Video 7B	20.08%	20.11%	20.11%	20.11%
Qwen2-VL 2B	22.76%	21.29%	21.46%	24.97%
Qwen2-VL 7B	20.14%	20.11%	20.22%	20.28%
Internvideo2-chat 8B	35.94%	37.06%	45.02%	34.94%

69 We also tested intra-context ordering—i.e., whether the sequence of video and text tags within each
 70 context sample affects model understanding. As seen in Table 5, this had minimal impact on most
 71 models.

Table 5: Effect of intra-context tag ordering on SSv2-VSM performance. We compare two context formats: Text-Video, where the text appears before the video tag in each context item, and Video-Text, where the video tag comes first.

Model	Text-Video	Video-Text
LLaVA-NeXT-Video 7B	20.11%	20.11%
Qwen2-VL 2B	24.09%	21.29%
Qwen2-VL 7B	22.32%	20.11%
Internvideo2-chat 8B	15.98%	37.06%

E Attention Visualization

To qualitatively evaluate the impact of stacked temporal attention (STA), we visualized the attention maps generated by InternVideo2-1B before and after applying STA. Figure 1 shows a person poking a lighter so that it falls, labeled with the class *poking [something] so that it falls over*. Before applying STA, the attention maps show that the model places minimal focus on the lighter, despite the fact that the action class is primarily defined by the lighter’s movement and fall. After incorporating STA, the model’s attention shifts significantly toward the lighter, especially as it begins to fall, indicating improved temporal modeling and relevance attribution.

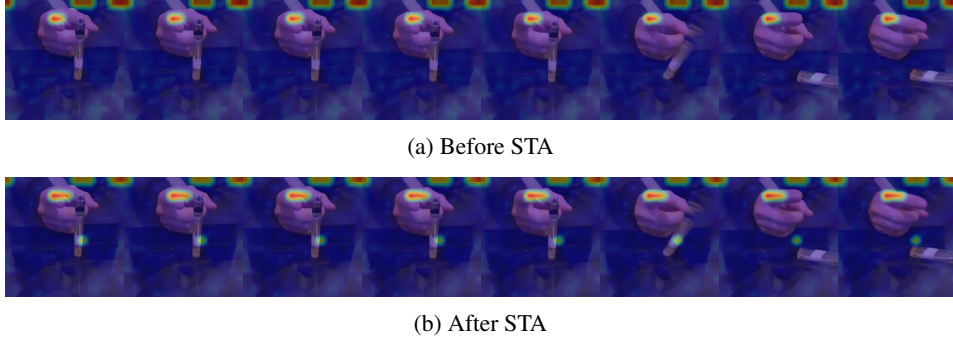


Figure 1: Attention maps for the action *poking [something] so that it falls over*.

These visualizations underscore the effectiveness of stacked temporal attention in helping the model focus on temporally relevant regions of the video. Especially in cases where the action depends on subtle object movements over time, STA enhances the model’s ability to localize and interpret key interactions, thereby improving classification performance.

F STAVEQ2.5

To further validate the effectiveness of stacked temporal attention blocks in enhancing Video-LLMs, we introduce STAVEQ2.5, which extends our temporal enhancement to the Qwen2.5-VL model by incorporating dedicated temporal attention blocks. Qwen2.5-VL employs windowed attention in most layers and full spatial self-attention in only four layers; however, our approach remains unaffected since each patch in the stacked temporal attention block attends to corresponding patches across all frames, bypassing the limitations imposed by windowed attention. This design preserves the temporal attention mechanism of STAVEQ2, ensuring consistent temporal modeling.

We train STAVEQ2.5 using the same two-stage strategy employed for STAVEQ2, leveraging the WebVid-QA dataset. Evaluation on video understanding benchmarks—VITATECS, MVBench, and Video-MME—demonstrates the model’s effectiveness in handling temporally complex scenarios. The results show that our stacked temporal attention approach not only enhances Qwen2-VL but also improves the newer Qwen2.5-VL, confirming its broad applicability across Video-LLM architectures.

Table 6: Accuracy (%) on video understanding benchmarks for our STAVEQ2.5 compared to other models. For VITATECS, aspect-wise results are shown; other benchmarks report overall accuracy. *(Video-MME without/with subtitles). † Results collected from the Video-MME leaderboard. – indicates results not reported in the original paper and unavailable from other sources.

Model	VITATECS						MVBench	*VMME (wo/w)
	Comp.	Dir.	Int.	Loc.	Seq.	Type		
Qwen2-VL 2B	80.8	82.1	69.6	76.1	72.2	85.9	63.2	55.6 / 60.4
STAVEQ2 2B (Ours)	81.3	83.0	70.1	76.9	72.9	86.6	65.1	56.2 / 61.3
ST-LLM 7B	–	–	–	–	–	–	54.9	–
TG-Vid 7B	–	–	–	–	–	–	56.4	–
LLaVA-OneVision 7B	–	–	–	–	–	–	56.7	58.2 / –
Qwen2-VL 7B	88.9	86.6	78.2	80.6	82.8	88.8	67.0	63.3 / 69.0
Qwen2.5-VL 7B	86.1	80.0	73.0	77.3	78.8	88.2	69.6	65.1 / 71.6
STAVEQ2 7B (Ours)	89.8	87.6	78.7	80.9	83.9	88.9	70.1	66.8 / 71.8
STAVEQ2.5 7B (Ours)	88.0	82.1	74.2	77.9	79.7	88.9	70.3	66.2 / 72.5
LLaVA-OneVision 72B	–	–	–	–	–	–	59.4	66.2 / 69.5
VideoLLaMA2 72B	–	–	–	–	–	–	62.0	61.4 / 63.1
LLaVA-Video 72B	–	–	–	–	–	–	–	70.6 / 76.9
Qwen2-VL 72B	89.8	87.8	77.9	85.3	84.8	90.4	73.6	71.2 / 77.8
Qwen2.5-VL 72B	92.1	88.9	81.9	87.1	89.4	91.8	70.4	73.3 / 79.1
STAVEQ2 72B (Ours)	92.8	90.1	82.3	87.9	90.3	92.8	74.5	73.9 / 79.9
STAVEQ2.5 72B (Ours)	93.1	90.9	82.1	88.0	90.8	93.3	72.4	74.2 / 79.8
GPT-4o [†]	–	–	–	–	–	–	–	71.9 / 77.2

97 As shown in Table 6, STAVEQ2.5 consistently outperforms its base model, Qwen2.5-VL, across
98 VITATECS, MVBench, and Video-MME, validating the effectiveness of our stacked temporal
99 attention approach in enhancing video understanding. Notably, STAVEQ2.5 72B surpasses STAVEQ2
100 72B on most benchmarks, demonstrating superior performance of our approach when applied to the
101 newer Qwen2.5-VL architecture. These results underscore the adaptability of our stacked temporal
102 attention mechanism, which drives performance gains across Video-LLMs of varying scales and
103 architectures, excelling in tasks that demand sophisticated temporal understanding.