

Thinking With Videos: Multimodal Tool-Augmented Reinforcement Learning for Long Video Reasoning

Supplementary Material

To facilitate a deeper understanding and reproducibility of our work, this supplementary material provides additional details on the implementation of the proposed VITAL framework, including **method** details (Sec. A), **experiment and efficiency** details (Sec. B), and **dataset** construction (Sec. C). We further present extended **ablation studies** (Sec. D) and additional **case analyses** (Sec. E) to comprehensively evaluate the effectiveness and robustness of our approach. Finally, we discuss the **limitations** of our current work and outline potential directions for **future works** (Sec. F). For reference and reproducibility, we also provide the experimental **code** along with detailed documentation.

A. Method Implementation Details

This section describes the methodology details of the proposed VITAL framework in Sec. 3. We first illustrate the implementation details of toolbox g_{tool} and then demonstrate the details of Difficulty-aware Group Relative Policy Optimization (DGRPO).

Table 9. Inputs and outputs format of visual tools

Tool Name	Inputs	Outputs
Clip Captioning	start, end	caption of the clip
Clip QA	start, end, question	answer to the question
Video Clipping	start, end	video clip visual tokens

A.1. Tool Implementation Details

In order to enhance the video reasoning ability of MLLMs by incorporating new evidences related to the user query, we implement three visual tools in the visual toolbox:

1. **Video clip captioning tool:** This tool takes the `start` and `end` timestamps of a video clip as input, and generates a descriptive `caption` for the specified segment.
2. **Video clip QA tool:** This tool receives the `start` and `end` timestamps of a video clip, together with a natural language `question`, as input. It outputs an `answer` to the given question based on the visual content of the specified clip.
3. **Video clipping tool:** This tool takes the `start` and `end` timestamps as input and outputs the visual content (represented as `visual tokens`) corresponding to the selected video segment.

Tab. 9 summarizes the input and output formats of these visual tools.

```
def get_video_clip_caption(start_time: float, end_time: float)
-> str:
    """
    Retrieve the text caption associated with a specified
    video segment.

    Args:
        start_time (float): The start time (in seconds) of
        the video segment, keeping two decimal places.
        end_time (float): The end time (in seconds) of the
        video segment, keeping two decimal places.

    Returns:
        str: The caption string describing the video segment.
    """
    pass

def ask_video_clip_question(start_time: float, end_time:
float, question: str) -> str:
    """
    Answer a question about the content of a specified video
    segment.

    Args:
        start_time (float): The start time (in seconds) of
        the video segment, keeping two decimal places.
        end_time (float): The end time (in seconds) of the
        video segment, keeping two decimal places.
        question (str): The question about the content of the
        video segment.

    Returns:
        str: The answer to the given question.
    """
    pass

def get_video_clip_frame(start_time: float, end_time: float)
-> List[Image]:
    """
    Get frames of a video segment from start_time to end_time.

    Args:
        start_time (float): The start time (in seconds) of
        the video segment, keeping two decimal places.
        end_time (float): The end time (in seconds) of the
        video segment, keeping two decimal places.

    Returns:
        List[Image]: A list of frame images of the segment.
    """
    pass
```

Figure 7. Python-style pseudo code for the three video reasoning tools. Here shows input parameters, output types, and function descriptions.

For all tools, the model parses and extracts tool call parameters following the Hermes-style protocol [1, 65]. Each tool call is formatted as a JSON string, enclosed within the `<tool_call>` and `</tool_call>` tags. After tool execution, the tool response is returned as a JSON string; if the tool outputs video tokens, the response includes a corresponding video token sequence. In cases where tool execution fails—due to incorrect parameter formatting or other

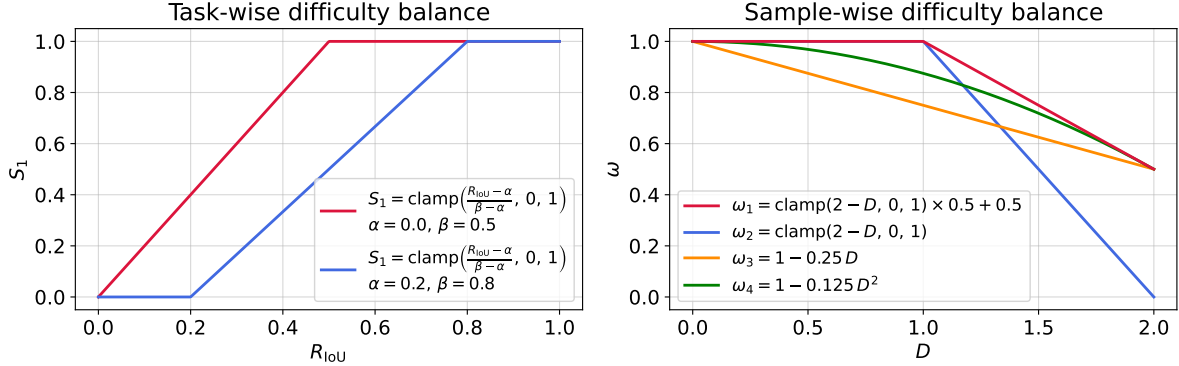


Figure 8. **DGRPO clamped difficulty balance functions.** Left: task-wise difficulty balance functions. Here we adopt two hyper-parameter settings for short video temporal grounding and long video temporal grounding tasks. Right: sample-wise difficulty balance functions. Here we adopt ω_1 for simplicity and effectiveness.

Table 10. **Reward design of DGRPO with tools.** For DGRPO without tools, tool reward is removed and the format reward ranges from 0 to 1.

Task	Sub-task	Accuracy Reward	Max Value		
			Acc	Format	Tool
Temporal Grounding	–	IoU	1	0.5	0.5
Reasoning VQA	MCQ	EM	1	0.5	0.5
	Number	EM	1	0.5	0.5
	Open-ended	Rouge	1	0.5	0.5
	OCR	WER	1	0.5	0.5
	Regression	L1	1	0.5	0.5
Grounded VQA	MCQ	$(IoU + EM) / 2$	1	0.5	0.5
	Open-ended	$(IoU + Rouge) / 2$	1	0.5	0.5

Table 11. **DGRPO hyper-parameters α and β** for each task and data source of MTRV-RL-110k training dataset.

Task	Data source	α	β
Temporal Grounding	Charades-STA	0.2	0.8
	ActivityNet-MR	0.2	0.8
	VidChapters-7M	0.0	0.5
Reasoning VQA	All	–	–
Grounded VQA	ReXTime	0.2	0.8
	NEt-GQA	0.2	0.8

unexpected errors—an error message is returned in the form of a JSON dictionary string.

The system prompt includes tool schemas derived from pseudo code in Fig. 7.

A.2. DGRPO Algorithm Details

We propose Difficulty-aware GRPO to mitigate the task-wise difficulty imbalance and sample-wise difficulty imbalance, as discussed in Sec. 3.3 and Alg. 1. In this subsection, we present the hyper-parameters used in our approach and provide the rationale behind their selection.

Reward design. In DGRPO reinforcement learning, we adopt three rewards: the accuracy reward, the format reward and the tool reward, as shown in Tab. 10.

For the accuracy reward, we adopt following rewards for each sub-task following [13]:

- **IoU:** the Intersection over Union value of predicted time range and ground truth time range.
- **EM:** Exact Match reward, is equal to 1 when the prediction matches ground truth answer exactly.

- **Rouge:** Recall-Oriented Understudy for Gisting Evaluation [41], calculated as an average of Rouge-1, Rouge-2 and Rouge-L between the prediction and the ground truth answer.
- **WER:** Word Error Rate, measuring the edit distance between the prediction and the ground truth answer.
- **L1:** calculated as $\text{clamp}(1 - |\hat{y} - y|/|y|, 0, 1)$, where, \hat{y} is the predicted number and y is the ground truth number.

The format reward is equal to 0.5 only when the model response τ matches this format exactly: `<think>... </think> <tool.call>... </tool.call> <think>... </think> <answer>... </answer>`, or `<think>... </think> <answer>... </answer>`. If the response does not match any of these formats, the format reward is set to 0.

Similarly, the tool reward is equal to 0.5 only when the model successfully performs a tool call, otherwise it is 0. For the second stage DGRPO without tools, the tool reward is removed and the format reward ranges from 0 to 1.

Table 12. **Training configurations.** Here `group_size` is the number of rollouts, and `num_turns` is the number of tool call rounds.

Configuration	Stage-1	Stage-2	Stage-3	Stage-4
method	SFT	DGRPO	SFT	DGRPO
using_tools	False	False	True	True
freeze_visual_encoder			True	
learning_rate	1e-5	1e-6	1e-5	1e-6
kl_loss_coef (β)	0	1e-2	0	1e-2
optimizer		AdamW		
AdamW_betas		(0.9, 0.999)		
weight_decay		1e-2		
warmup_ratio	0.1	0	0.1	0
lr_scheduler		cosine		
group_size	–	8	–	8
batch_size	256	64	256	64
mini_batch_size	256	64	256	64
micro_batch_size_per_device	4	2	4	2
number_of_samples	54k	94k	18k	16k
number_of_epochs	1	1	1	1
max_num_turns	0	0	2	2
max_total_pixels	64×224×224		3×64×224×224	
max_sequence_length	4096	4096	10240	10240
max_response_length	0	1024	0	1024
training throughput (/gpu/h)	1408.0	142.9	1429.3	32.7
sample speed (s)	2.6	25.2	2.5	110.3

Task-wise difficulty balance. DGRPO algorithm uses clamped linear transformations for task-wise difficulty balance. As shown in Tab. 11 and in the left of Fig. 8, we adopt two hyper-parameter settings for short video temporal grounding and long video temporal grounding tasks based on their difficulty distributions. For multiple-choice question VQA tasks, we do nothing about task-wise difficulty balance as their reward metric is originally discrete (0 or 1). For Gounded VQA tasks, we transform the IoU value for balancing before adding it with EM or Rouge value. The choice of hyper-parameters in Tab. 11 is studied in Tab. 17.

Sample-wise difficulty balance. As presented in the right of Fig. 8, we apply sample-wise difficulty balance with another clamped linear transformation ω_1 , which results in a updating weight w based on sample difficulty D . The sample difficulty D is estimated by averaging all rollout rewards \hat{R} of the sample. We adopt the piecewise linear function ω_1 with the motivation to provide *soft penalty to easier samples* regardless of the format reward and the tool reward. (format reward + tool reward ≤ 1 , and they always converge to a constant value after several training steps) We also compare other transformation functions in Sec. D.1.

B. Experimental Details

In this section, we provide more details about the four-stage training procedure, the evaluation settings and met-

rics, and efficiency analysis. For image samples in the training dataset, we set `max_pixels` = 448×448. For video samples in the training dataset and evaluation datasets, we first sample them at FPS = 2 and then bound the number of frames and frame pixels according to the video length. We adopt the number prompt technique [80] during training and evaluation to print absolute timestamps on frames, providing additional temporal information for MLLMs for accurate temporal perception.

B.1. Training Details

The training configurations are listed in Tab. 12. Generally, we split the training procedure into two phases, post-training without tools and post-training with tools, each of which contains a cold-start SFT stage and a RL stage. For the 3rd and the 4th stage training with tools, we only train the model on long video datasets since our motivation is to enhance long video understanding with visual tools. We also tried training with tools from the pretrained MLLM directly, but found this paradigm is inefficient and the model is optimized slowly, since tool-augmented RL takes longer time than text-based RL because it contains multi-round generation, as shown in Tab. 12.

In each training stage, we sample short videos with `max_frames` = 64 and `max_pixels` = 224×224, while sampling long videos (e.g., LongVideo-Reason or Vid- Chapters-7M) with `max_frames` = 256 and `max_pixels` = 112×112 for the initial input video. For densely sampled

Table 13. **Evaluation configurations for each dataset.** Here VTG stands for video temporal grounding, VQA for video question answering, GQA for grounded VQA and RQA for reasoning VQA. In all evaluations, tool_max_frames = 64, and tool_max_pixels = max_pixels.

Task	Dataset	Max frames	Max pixels	Is zero-shot?
Long VTG	Vid-Chapter-7M [83]	1024	224*224	No
	VUE-TR-Vision [64]	1024	224*224	Yes
Long VQA	Video-MME [14]	1024	224*224	Yes
	LongVideo-Reason [8]	1024	224*224	No
	LongVideoBench [78]	1024	224*224	Yes
	LVBench [67]	1024	224*224	Yes
	CGBench [4]	1024	224*224	Yes
VTG	Charades-STA [15]	256	384*384	No
	ActivityNet-MR [28]	256	384*384	No
GQA	NExT-GQA [81]	256	384*384	No
	ReXTime [6]	256	384*384	No
RQA	VSI-Bench [85]	256	384*384	Yes
	Video-MMMU [21]	256	384*384	Yes
	MMVU (mc) [93]	256	384*384	Yes

video clip, *i.e.*, the tool results, we set max_frames = 64 and max_pixels = 224×224 for all videos. For DGRPO rollout generation, we set temperature = 1.0, top_p = 1.0, group_size = 8 following common practices of MLLM RL post-training [13, 38, 71].

B.2. Evaluation Details

We evaluate the VITAL-7B model on fourteen challenging benchmarks. The evaluation details are illustrated in table Tab. 13. In all evaluation experiments, we keep temperature = 0.01 and top_p = 0.001, with max_response_length = 1024 for each round of generation, which guaranties stable and reproducible results. For tool calling, max_num_turns = 2, tool_max_frames = 64, and tool_max_pixels = max_pixels of the original sampling setting. For all benchmarks, we test models with their officially published test split. For some benchmarks, their corresponding train split is included in our proposed training dataset; while for others, we test model with zero-shot setting, as detailed in Tab. 13.

B.3. Efficiency Analysis

Training Efficiency. Tab. 14 highlights the high training efficiency of our model, with SFT stages achieving the highest throughput, while DGRPO post-training involves more complex optimization but still maintains reasonable resource usage. The fourth stage DGRPO-Tool includes two to three times of conversation compared to that of the third stage, so the throughput is still reasonable. All experiments were conducted on H100 GPUs with DeepSpeed.

Inference Efficiency. Tab. 15 demonstrates efficient inference, where the higher latency in the two-call setting is

Table 14. Statistics of training efficiency.

Stage	Data Size (k)	GPU Hours (h)	Throughput (/gpu/h)
SFT	54	36	1500
DGRPO	94	390	241
SFT-Tool	18	12	1500
DGRPO-Tool	16	202	79

Table 15. Statistics of inference latency.

Inference Latency (s)	zero-call	one-call	two-call
Sample Latency	2.04	5.61	8.88
First-round Latency	2.04	2.38	2.42
Average Tool Latency	–	3.23	3.23

due to the increased conversation rounds, without incurring extra computational overhead. Note that the inference time for tools involving MLLM calls (e.g., the clip captioning or QA tool in Tab. 9) is variable and significantly higher than that of the video clipping tool (usually between 30 to 60 seconds per call). All experiments were conducted on 8 H100 GPUs with vLLM deployment. During inference, the GPU memory consumption is within 80GB.

C. Dataset Construction Details

C.1. Dataset Statistics

We present comprehensive statistics of the MTRV dataset in Figs. 9 and 10. The dataset comprises four subsets: MTRV-CoT (54k), MTRV-RL (94), MTRV-CoT-Tool (18k), and MTRV-RL-Tool (16k). Fig. 9 illustrates the distributions of video duration and token length across all subsets. No-


```

SYSTEM_PROMPT:
You are an advanced video analysis assistant.
All your thinking processes should be a concise, logical, visual-based chain-of-thought, less than 300
tokens. Think naturally as a human would, using inner dialogue like "let me think," "hmm," "wait," etc.
All your formal output should be a brief sentence, less than 50 tokens.
-----
PROMPT:
This is a video with a duration of {duration} seconds.
When answering the user's question, please carefully analyze and reason step by step, making sure to
incorporate and reference the visual content of the video as part of your thought process. Think as a
human would: use natural internal dialogue such as "let me think," "hmm," "wait," "oh, I see," "let's
break it down," etc., and include self-reflection or verification as you reason through the problem.
Your thought process should be output within <think>...</think> tags, and your final answer within
<answer>...</answer> tags. For example: <think> ... </think><answer> ... </answer>.
Important:
- Your reasoning and answer must be based on the visual information from the video.
- Do not mention or refer to any "ground truth", "reference answer" or similar terms in your output.
- You must arrive at your final answer solely through step-by-step visual reasoning, and your final answer
must match the provided Ground Truth Answer exactly.
User Question: {question}
Ground Truth Answer: {answer}

```

Figure 11. Prompts for text-based CoT generation.

Here L denotes the duration of the input video, $\text{rand}()$ is a random float number between 0 and 1. λ is a randomness parameter, which is set to 0.2 by default. The clamp function ensures that the resulting values remain within the valid range $[0, L]$. For other tasks like video question answering, we do not use the tool parameter suggestion and remove the corresponding sentences in the prompts in Fig. 12. We turn on the thinking mode of reasoning MLLM in round 1 and 3.

C.4. Data Post-Process

After generating chain-of-thoughts automatically, we perform a rule-based post-processing to control the data quality. A sample is excluded from the training dataset if it meets any of the following criteria:

1. The chain-of-thought is incomplete or does not reach a final answer.
2. The generated answer does not match the ground truth.
3. The sample contains irrelevant or off-topic content, *e.g.*, direct description of “ground truth” or “suggestion”.

After data post-processing, we obtain the final **MTVR-CoT-72k** dataset, which consists of high-quality and well-formatted samples suitable for cold-start supervised fine-tuning. The dataset consists of two subsets, **MTVR-CoT (54k)** for training stage-1 and **MTVR-CoT-Tool (18k)** for training stage-3, as illustrated in Fig. 5 in the main paper.

D. More Ablation Studies

D.1. Ablation Study on DGRPO Transformations

We conducted ablation studies on the selection of transformation functions for Sample-wise Difficulty Balance of

DGRPO. As presented in Sec. A.2 and Fig. 8, we design four transformation functions:

$$\omega_1(D) = \text{clamp}(2 - D, 0, 1) \times 0.5 + 0.5 \quad (8)$$

$$\omega_2(D) = \text{clamp}(2 - D, 0, 1) \quad (9)$$

$$\omega_3(D) = 1 - 0.25D \quad (10)$$

$$\omega_4(D) = 1 - 0.125D^2 \quad (11)$$

Tab. 16 shows the experimental results of using these functions to train for the first two stages. Comparing $\omega_1(D)$ and $\omega_2(D)$, we observe the importance of soft penalty. $\omega_2(D)$ style hard penalty will totally remove the contribution of easy samples, which is harmful for training. $\omega_1(D)$ also outperforms continuous functions $\omega_3(D)$ and $\omega_4(D)$. We attribute it to the phenomenon that the format reward + tool reward is always equal to 1 after tens of training steps. After that, the reward will vary from 1.0 to 2.0 in most situations. Therefore, $\omega_1(D)$ maximizes the range of sample-wise difficulty coefficient, resulting in better performance.

Table 16. **Ablation study on DGRPO transformations.** Each experiment contains training for the first two stages, *i.e.*, SFT and DGRPO without tools.

	Function	LVR Acc	VidCh mIoU	MMMU Acc	Cha mIoU	Avg
Ⓒ	$\omega_1(D)$	70.2	28.8	52.1	57.1	52.1
	$\omega_2(D)$	58.3	22.6	39.9	44.2	41.3
	$\omega_3(D)$	65.0	26.7	47.3	51.0	47.5
	$\omega_4(D)$	68.9	27.9	49.5	54.6	50.2

SYSTEM_PROMPT:
<p>You are an advanced video analysis assistant. All your thinking processes should be a concise, logical, visual-based chain-of-thought, less than 300 tokens. Think naturally as a human would, using inner dialogue like "let me think," "hmm," "wait," etc. All your formal output should be a brief sentence, less than 50 tokens.</p>
----- PROMPT_ROUND_1:
<p>This is a video with a duration of {duration} seconds. Step 1: Think. Analyze the user's question based on the entire visual content of the video step-by-step. You may start by analyzing each key event in the video, including self-reflection or verification in your reasoning. --- User Question: {question} Suggested Range: {suggestion} --- Your just need to think and analyze. Your thinking should be ended around the Suggested Range, but you MUST NOT mention "suggest" or similar words in thinking or output. Finally, you will think about calling the get_video_clip_frame tool with the Suggested Range as parameters, but you MUST NOT really call the tool. You MUST NOT directly answer the user question, since your answer may be incorrect and you need tool results for further analysis.</p>
----- PROMPT_ROUND_2:
<p>Step 2: Act. Based on the reasoning process you just formulated, call the appropriate tool to analyze the most relevant video segment. Your output must be ONLY the tool call. --- User Question: {question} Suggested Range: {suggestion} --- Your output for this step must be exactly one tool call. The tool call parameters must be exactly the Suggested Range.</p>
----- PROMPT_ROUND_3:
<p>You have received the tool output for the video clip from {start_time:.2f} to {end_time:.2f} seconds. Step 3: Answer. Integrate your initial reasoning with the new visual evidence from the tool. Think step-by-step again and provide the final answer. --- User Question: {question} Ground Truth Answer: {answer} --- Your final answer must exactly match the provided Ground Truth Answer, but you MUST NOT mention or refer to any "ground truth", "reference answer", or similar terms in thinking or output.</p>

Figure 12. Prompts of each round for multimodal CoT generation.

D.2. Sensitivity Analysis of Hyper-parameters

We conduct a sensitivity analysis on the DGRPO hyper-parameters α and β to determine the optimal settings for the task-wise difficulty balance function. As detailed in Tab. 17, we perform two distinct sets of experiments: one training exclusively on short video temporal grounding data (rows 1-4), and another solely on long video data (rows 5-8). The results reveal a clear unimodal performance trend for each data type. For short videos, performance peaks at $(\alpha, \beta) = (0.2, 0.8)$, maximizing scores on relevant benchmarks like MMMU and Charades. For long videos, the op-

timal setting is $(\alpha, \beta) = (0.0, 0.5)$, which excels on LVR and VidCh. For long video temporal grounding, the IoU distribution is heavily skewed towards zero, creating a long tail of extremely difficult samples. Setting $\alpha = 0$ ensures that even minimal IoU gains from these challenging samples provide a valid learning signal, preventing their rewards from being prematurely clamped to zero. The final row confirms that combining these independently optimized settings yields the best overall performance, demonstrating the effectiveness of our DGRPO configuration.

Table 17. **Ablation study on DGRPO hyper-parameters.** Each experiment contains training for the first two stages, *i.e.*, SFT and DGRPO without tools.

Short video α, β	Long video α, β	LVR Acc	VidCh mIoU	MMMU Acc	Cha mIoU	Avg
0.3, 0.7	w/o	65.5	23.1	51.5	56.2	49.1
0.2, 0.8	w/o	67.1	24.2	52.0	57.3	50.0
0.1, 0.9	w/o	66.2	23.5	51.6	56.5	49.5
0.0, 1.0	w/o	64.0	22.0	51.0	55.4	48.1
w/o	0.0, 1.0	69.5	27.5	45.1	48.8	47.7
w/o	0.0, 0.7	69.8	28.0	46.5	50.1	48.6
w/o	0.0, 0.5	70.0	28.5	47.2	51.3	49.3
w/o	0.0, 0.3	69.6	27.8	46.1	49.5	48.3
0.2, 0.8	0.0, 0.5	70.2	28.8	52.1	57.1	52.1

D.3. Ablation Study on Cold Start

Tab. 18 presents an ablation study on cold start SFT. The first two rows serve as baselines without tool usage, where adding cold start SFT shows few performance improvements. Introducing additional training with tools (row 3 and row 4) leads to notable improvements. Comparison between row 3 and row 4 demonstrates that cold start SFT (row 4) is beneficial for tool-augmented reinforcement learning, while directly applying tool-augmented DGRPO after two-stage training (row 3) impedes the model from learning to reason with tool calls.

Table 18. **Ablation study on cold start and video resolution.** The first two rows present results without using tools. * in row-3/4 indicates further training with tools based on row-2 (Exp. ⑥). † denotes experiments trained with double the max_total_pixels. #Tools is the average number of successful tool calls.

Train Stage	LVR Acc	VidCh mIoU	MMMU Acc	Cha mIoU	Avg	#Tools
DGRPO	70.1	28.5	50.9	57.5	51.8	0
⑥ SFT+DGRPO	70.2	28.8	52.1	57.1	52.1	0
+DGRPO*	70.5	30.1	51.3	57.3	52.3	0
⑦ +SFT*+DGRPO*	79.3	35.0	54.2	59.9	57.1	0.95
SFT+DGRPO†	73.7	30.9	51.3	58.9	53.7	0

D.4. Fair Ablation of Tool Calling

We conduct an ablation study on whether to use tools in Tab. 1 in the main paper (Exp. ⑥ vs Exp. ⑦). However, this may be due to the different number of video tokens during the final training stage of them, as presented in Tab. 12 (Stage-2 and Stage-4).

Therefore, for a fairer comparison, we conduct an additional two-stage experiment, shown in row 5 of Tab. 18, which matches the experimental setting of row 2 except for the increased video resolution. In the experiment shown in row 5, we set max_total_pixels to $2 \times 64 \times 224 \times 224$. This

setting ensures that the video token budget is the same as that of Exp. ⑦, as the average number of tool calls in Exp. ⑦ is less than 1. As shown in the last two rows of Tab. 18, VITAL-7B with tools (row 4) still significantly outperforms the improved baseline (row 5) by a large margin.

D.5. Ablation Study on Data Size

We provide evaluation results at different steps during the last training stage, DGRPO with tools. As shown in Tab. 19, although some metric fluctuations occur during training, the overall performance on both long and short video understanding benchmarks consistently improves as the training data size increases.

Table 19. **Ablation study on different data sizes.** Data size denotes the data already used during stage-4 DGRPO training.

	Data Size	LVR Acc	VidCh mIoU	MMMU Acc	Cha mIoU	Avg
	0	68.7	18.9	45.2	52.3	46.3
	4k	73.8	27.0	50.6	55.8	51.8
	8k	76.1	31.5	51.8	58.3	54.4
	12k	75.8	29.5	52.7	58.6	54.1
⑦	16k	79.3	35.0	54.2	59.9	57.1

E. More Case Analyses

E.1. Data Quality Analysis

High-quality training data is the foundation of robust model performance. In this subsection, we examine representative samples from our MTRV dataset to illustrate the diversity and complexity of the video-question pairs. As shown in Figs. 13 to 16, the MTRV training set features precise temporal annotations, detailed event descriptions, and a wide range of real-world scenarios. These examples demonstrate the dataset’s ability to capture fine-grained actions and support complex reasoning, ensuring comprehensive coverage for both short and long video understanding tasks. Such high-quality, diverse samples provide a strong foundation for training models capable of robust and generalizable video-language reasoning, *e.g.*, text-based CoT or tool-augmented multimodal CoT.

E.2. Reasoning Quality Analysis

Understanding the model’s reasoning process is crucial for evaluating its effectiveness in complex video understanding tasks. We present several case studies that compare multimodal CoT reasoning chains generated by VITAL-7B with text-based CoT generated by VITAL-7B (w/o). As shown in Figs. 17 and 18, the results demonstrate that tool-augmented multimodal CoT enables VITAL-7B to perform more accurate and fine-grained reasoning compared to text-only CoT.

By leveraging both visual and textual modalities, as well as external tools for temporal grounding and evidence extraction, the model can systematically integrate multimodal cues, leading to more precise event localization and deeper narrative understanding. These case studies highlight the substantial advantages of multimodal, tool-augmented reasoning chains for challenging long video question answering and temporal reasoning tasks.

E.3. Failure Case Analysis

Despite the overall strong performance of VITAL-7B, certain challenging scenarios expose the limitations of our current approach. As shown in Fig. 19, VITAL-7B fails to accurately localize the event boundaries for the “pregnancy announcement” despite leveraging tool-augmented reasoning. The primary reason for this error is the model’s *insufficient OCR capability*: it struggles to reliably extract and interpret key textual cues from video frames, such as the explicit announcement and supporting details. This limitation leads to imprecise temporal grounding, as the model cannot effectively anchor such kind of events to its correct time segment. To address this, future work should focus on integrating more robust training data or module design regarding OCR ability, enabling the model to better capture and utilize fine-grained textual information for event localization in long videos.

F. Limitations and Future Work

Limitations. While VITAL achieves strong results in long video reasoning, our current framework only provides tools for temporal grounding and question answering, limiting its ability to address other tasks such as spatial grounding. Furthermore, our approach mainly focuses on visual features and ignores audio information, which may restrict the model’s overall understanding. Expanding the toolbox and incorporating multimodal features like audio could further enhance the model’s comprehension of videos.

Future Work. In the future, we plan to expand *the variety of tools* within our framework and enable multimodal chain-of-thought reasoning across more modalities, such as integrating both visual and audio cues. This will allow our method to support a broader spectrum of video understanding tasks, including spatio-temporal video grounding, semantic segmentation, and analysis of videos in complex scenarios. Furthermore, we aim to develop more *adaptive tool selection* strategies to enhance the robustness and versatility of multimodal reasoning in open-ended video understanding. In addition, we will continue to *enrich our training datasets* to cover a wider range of task types, particularly those requiring strong OCR capabilities. By curating and annotating more diverse video samples—including

scenarios with dense textual information and complex event structures—we hope to further improve the model’s ability to generalize and accurately localize events across various video domains.

Question: Please find the visual event described by a sentence in the video, determining its starting and ending times. The sentence: **person talking on the phone.**

Ground Truth: From 4.00 to 16.80 seconds.

Data Source: MTVR-CoT (54k), Charades-STA

Task Type: Temporal Grounding



Thinking without tools

<think>Let me carefully analyze the sequence of images to determine when the person starts and stops talking on the phone.

First, I need to identify when the person begins talking on the phone. In the early frames (0.00s to about 4.00s), the person is holding a phone and a glass, but not yet talking on the phone. At 4.97s, the phone is brought up to the ear, and by 5.47s, the person is clearly holding the phone to their ear, which is a typical gesture for talking on the phone.

Now, I need to find when this action ends. The person continues to hold the phone to their ear while moving into the kitchen, up until around 16.80s. At 16.89s, the phone is no longer at the ear, and the person is now using both hands for other tasks, indicating the phone conversation has ended.

So, the event "person talking on the phone" starts at 4.00s (when the phone is first brought to the ear) and ends at 16.80s (just before the phone is lowered).

Let me double-check:

- At 4.00s, the phone is being raised.
- At 4.97s and after, the phone is at the ear.
- At 16.80s, the phone is still at the ear.
- At 16.89s, the phone is no longer at the ear.

This confirms the event duration. </think>

<answer>The event 'person talking on the phone' happens in the 4.00 - 16.80 seconds.</answer>

Figure 13. This case study highlights the high data quality of the **MTVR-CoT (54k) dataset** for temporal grounding tasks. The precise frame-by-frame annotations enable accurate identification of the event boundaries—here, the action "person talking on the phone" is reliably localized between 4.00s and 16.80s. Such detailed and consistent text-based CoT reasoning label supports robust event understanding and temporal modeling in complex video scenarios.

Question: Based on the content of the video, answer the following question: **How does she style her hair to give it more volume?** In the <answer> </answer> tag, first specify the exact time period in seconds of the video segment that support your answer, then, provide your final answer with a short sentence.

Ground Truth: From 72.58 to 157.25, she blow dries her hair with a white blower, sections her hair, brushes her hair with a roller brush while blow drying it at the same time, and rolls the brush downwards and upwards.

Data Source: MTVR-RL (94k), ReXTime (ActivityNet)

Task Type: Grounded VQA (Open-ended)



Figure 14. This case study demonstrates the high-quality annotations in the **MTVR-RL (94k) dataset** for grounded video question answering. The dataset provides precise temporal grounding for open-ended questions, as shown by the accurate localization of the hair styling process between 72.58s and 157.25s. The annotations are collected from original data source.

Question: Please find the visual event described by a sentence in the video, determining its starting and ending times. The sentence: **First Views of Fairmont Chateau Lake**

Ground Truth: From 91.0 to 163.0 seconds

Data Source: MTVR-RL-Tool (16k), Vid-Chapters-7M

Task Type: Long Video Temporal Grounding



Figure 15. This case study highlights the precise temporal localization enabled by the **MTVR-RL-Tool (16k) dataset** for long video understanding. By providing accurate annotations for complex events, such as identifying the exact segment (91.0s to 163.0s) for "First Views of Fairmont Chateau Lake," the dataset supports fine-grained reasoning and comprehensive analysis in complex, extended, multi-scene, long video content. The annotations are collected from original data source.

Question: What is the primary goal of the person in the video, as inferred from the interaction dynamics, textual cues, and video composition?

- A. To demonstrate proper handling techniques for various reptiles.
- B. To advocate for adopting reptiles as pets by showcasing their docile nature.
- C. To promote a YouTube channel focused on reptile care through direct calls to action.
- D. To document a personal hobby of collecting exotic lizards and geckos.

Ground Truth: C

Data Source: MTRV-CoT-Tool (18k), LongVideo-Reason

Task Type: Long Video Reasoning VQA



Thinking with tools

<think>Alright, here's my take on this video's purpose, distilled as if it were my own internal monologue. Okay, so I need to figure out the main objective of this video based on its content. The person's got a crested gecko, and they're holding it, so maybe it's about reptile handling or showing how docile they are. But the shirt, "I'D RATHER BE AT HOME WITH MY LIZARDS," suggests this goes deeper. Let's break it down.

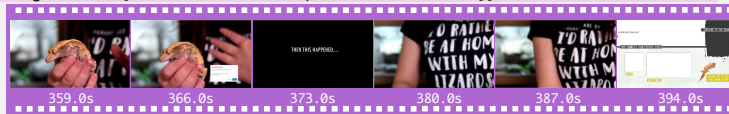
The video jumps around, showing other reptiles: a tortoise, bearded dragons, snakes. Seems like a collection, so maybe it's about their hobby. But then there are those text overlays, asking questions. "Will you get a bearded dragon?" "Will you get a snake?" That's a classic YouTube format. And the answers are shown with the actual reptiles, keeping viewers hooked.

The "Subscribe" button appearing and getting clicked, then the end screen with the subscribe prompts, links to other videos, and social media? That's a dead giveaway. **This isn't just a video about reptiles; it's designed to grow a YouTube channel. It's the core strategy of the content. The reptile content is the vehicle.**

Sure, they are showing off the reptiles. They are likely documenting their personal hobby, and maybe even showcasing their nature. They are definitely trying to engage their audience. But, when you get down to it, the main push is about growing that channel.</think>

<tool_call>{"name": "get_video_clip_frame", "arguments": {"start_time": 359.00, "end_time": 394.00}}</tool_call>

Get_video_clip(start=359, end=394)



<think>Okay, here's how I'm thinking about this, based on the tool's output and my expertise.

Alright, so I need to pick the best answer to this multiple-choice question, and it's all about the video's primary purpose. My initial read was that this video is *specifically* designed to promote a YouTube channel. My reasoning was based on a few things, the Q&A format, which can be a way YouTubers interact with viewers. But the strongest clues were the calls to action I expected, and the elements I thought I would find in a YouTube end screen.

Now, let's see if the tool backs up my initial reasoning. Looking at the clip from the tool (between 359 and 394 seconds), my hypothesis has been *completely* validated. I mean, it's practically a checklist of what I expected to see.

- At the very start of the clip, at 359.55s, we get the classic "Subscribe" button animation - with a click!
- At 360.64s we see "Subscribed".

• Then, **the tool identified a YouTube notification prompt at 362.83s.**

• Finally, at 393.45s, the end screen appears. This is the really telling part. I'm seeing a "SUBSCRIBE TO THIS CHANNEL" bar, a comments section, more video recommendations, and calls to "FOLLOW THE GECKOS" with links.

So, this is undeniable proof. The video is using very deliberate, very YouTube-specific elements.

Here's how I'm synthesizing all of this to pick the correct answer:

• The Q&A format isn't *just* about answering questions. It's a proven engagement tactic for YouTubers.

• Yeah, the video *does* feature reptiles, but it's not an instructional video. It's not designed to teach specific handling techniques - it's more of a personal, and very casual look at keeping reptiles. That would mean rule out Option A.

• The video *does* document someone's hobby, so I could see how option D seems possible, but again it's not the main purpose.

• And yes, the video *may* encourage reptile adoption indirectly, but that's not the video's central thrust, so rule out Option B.

• The *really* crucial part here is the explicit use of those on-screen subscribe buttons, notification prompts, and the dedicated end screen. Those aren't incidental features; they are the core of the video's structure and purpose. They're there to grow a YouTube channel and drive viewer engagement.

Therefore, **the only logical conclusion is C. To promote a YouTube channel focused on reptile care through direct calls to action.** The evidence is absolutely clear.</think>

<answer>C</answer>

Figure 16. This case study illustrates the depth and reliability of the **MTRV-CoT-Tool (18k) dataset** for long video reasoning tasks. By combining multimodal cues—including video structure, on-screen text, and explicit calls to action—the dataset enables nuanced analysis of the video's primary goal. The detailed annotation and step-by-step reasoning process allow for accurate identification of the intent: promoting a YouTube channel focused on reptile care through direct viewer engagement. This demonstrates the dataset's capacity to support complex, high-level, tool-augmented reasoning for long videos.

Question: Please find the visual event described by a sentence in the video, determining its starting and ending times. The sentence: **Regional management approaches should be adopted, as demonstrated by the herring project.**

Ground Truth: From 1682.00 to 1813.00 seconds.

Thinking without tools: From 1300.77 to 1379.73 seconds. IoU = 0.0% **Fail!**

Thinking with tools: From 1699.00 to 1808.00 seconds. IoU = 83.2% **Success!**



`Get_video_clip(start=900, end=1990)`

Thinking without tools

<think>
The user wants to find the start and end times for the event where "Regional management approaches should be adopted, as demonstrated by the herring project" is discussed.

1. I will scan through the provided video frames to locate any mention of **the herring project or regional management approaches**.
2. At frame 1206.72s, a slide titled "The Bristol Channel Herring Project" appears. This marks the beginning of the discussion on this specific project.
3. At frame 1210.73s, the slide changes to "Next phase: Identify and map spawning". This confirms that the herring project is still the focus.
4. At frame 1239.93s, a new slide titled "Regional Approaches to Fisheries Management" appears. This slide outlines four steps for adopting regional management approaches. This is a strong indicator that the discussion about adopting regional management approaches has begun.
5. At frame 1250.73s, the slide "Regional Approaches to Fisheries Management" is fully visible and the speaker is presenting it.
6. At frame 1279.94s, the slide "Fisheries Research and Management Plans" appears, which is part of the regional management approach outlined in the previous slide.
7. At frame 1309.15s, the slide "What needs to be done" appears, summarizing the need for improvements in local governance and co-management.
8. At frame 1320.16s, the speaker is finishing his point on the need for co-management.
9. At frame 1321.16s, a new slide titled "How might we achieve 'leaving the environment in a better state'" appears. This clearly marks the end of the discussion on the herring project and the beginning of a new section. Based on this analysis, the event starts when the herring project is introduced at 1206.72s and ends right after the speaker finishes discussing the need for co-management at 1321.16s. Therefore, the event spans from approximately 1206.72 seconds to 1321.16 seconds.</think>

<answer>The event happens in the **1300.77 - 1379.73 seconds**</answer>

Thinking with tools

<think>Okay, let's pinpoint those timestamps. I'm looking for the exact moment this specific text appears on screen. First, I'll scan through the video, focusing on the slides. Ah, here we go! At around 1261.26s, the slide titled "Regional Approaches to Fisheries Management" pops up. That's our starting point. Now, I need to find when the specific phrase "Regional management approaches should be adopted, as demonstrated by the herring project" actually *appears* on the screen. Let's see... at 1261.26s, the slide is there, but not the full text. It's building up. Okay, at 1262.72s, the complete text is fully visible. Perfect. Now, I need to determine when it disappears. Scanning forward, it's still there at 1264.18s. But at 1265.64s, it's gone. So, the text is fully displayed between approximately 1262.72s and 1264.18s. To be precise, I'll use the start time of 1262.72s and end time of 1264.18s. Let's check it.</think>

<tool_call> {"name": "get_video_clip_frame", "arguments": {"start_time": 900.00, "end_time": 1990.00}}</tool_call>



<think>Okay, let's break this down. I'm tasked with pinpointing the start and end times for the appearance of the phrase "Regional management approaches should be adopted, as demonstrated by the herring project."

First, I did a quick scan of the entire video. I saw the phrase "Regional management approaches should be adopted..." pop up on a slide, specifically under the "Regional Approaches to Fisheries Management" section. My initial guess was somewhere between 1262.72 and 1264.18 seconds, based on a quick look at the full video. Next, I zeroed in on the relevant section using the tool. I found the phrase "Regional management approaches should be adopted..." on a slide, and **it appeared around 1703.99 seconds. The slide also mentioned the Bristol Channel Herring Project. This is promising!** Now, I need to refine my search. I'm looking for the *first* time the phrase appears. I went back and reviewed the frames leading up to 1703.99 seconds. At 1699.00 seconds, the previous slide was still present. Then, at 1703.99 seconds, the target slide appeared. So, the start time is definitely around 1703.99 seconds. I also need to determine the *last* time the phrase is visible. I went through the subsequent slides. The phrase remained on screen until about 1803.15 seconds. Then, at 1808.31 seconds, the next slide appeared. Therefore, the end time is approximately 1803.15 seconds. Finally, I want to make sure I have the right context. The phrase is part of a larger discussion about regional approaches. The slide also mentions the Bristol Channel Herring Project. This confirms that the phrase is indeed the key focus. So, the phrase "Regional management approaches should be adopted, as demonstrated by the herring project" first **appears on screen at approximately 1703.99 seconds and disappears around 1803.15 seconds.** I'm confident in these timestamps.</think>

<answer>The event happens in the **1699.00 - 1808.00 seconds**.</answer>

Figure 17. This example demonstrates the effectiveness of the VITAL-7B model's multimodal chain-of-thought (CoT) reasoning for **video temporal grounding tasks**, which leverages tool-augmented analysis to integrate visual and textual information. By systematically using external tools to extract and interpret relevant video segments, the model accurately locates where "regional management approaches should be adopted, as demonstrated by the herring project" is discussed. This tool-augmented multimodal CoT enables robust, step-by-step reasoning and significantly enhances temporal localization in complex, real-world video tasks.

Question: Based on the narrative progression and environmental details, what is the most likely connection between the framed photo of another man on the desk (0:00:20-0:00:30) and the later appearance of the robot in the office (0:05:30-0:05:40)?

- A. The man in the photo is a rival executive who sent the robot to sabotage the protagonist.
- B. The photo depicts a younger version of the protagonist, hinting at a time-travel plot involving the robot.
- C. The photo shows the inventor of the robot, who died before completing the project the protagonist is reacting to.
- D. The robot is an AI replica of the man in the photo, created to fulfill a corporate strategy revealed in the documents.

Ground Truth: C
Thinking without tools: D *Fail!*
Thinking with tools: C *Success!*



Get_video_clip(start=20, end=30)

Thinking without tools

<think>The reasoning begins by analyzing the framed photo of another man on the desk. The man in the photo is dressed in a suit and tie, suggesting he holds a position of authority or importance. This detail establishes him as a key figure in the narrative. Next, the robot appears in the office, shown interacting with a computer and displaying a "CODE" screen. This scene indicates the robot is designed to execute specific tasks, likely tied to corporate objectives. The robot's presence in the office suggests it is intended to assist or automate certain functions, implying a connection to the man in the photo. Further analysis of the video reveals a document titled "CODE KEN-2" dated 1987, which outlines a plan to "build a robot to do all the work." This document directly links the robot to a corporate strategy aimed at automating tasks. The robot's appearance in the office aligns with this plan, reinforcing its role as a tool for corporate efficiency. The man in the photo could be the inventor of the robot, who died before completing the project. However, the robot's advanced features and its ability to interact with the computer suggest it is more than just a physical creation. **It is likely an AI replica of the man in the photo,** designed to fulfill his corporate vision and execute the plan outlined in the document. In conclusion, the most logical explanation is that the robot is an AI replica of the man in the photo, created to fulfill a corporate strategy revealed in the documents.</think>
<answer>D</answer>

Thinking with tools


<think>Alright, let's dissect this. I'm looking at a video, and the task is to figure out how a picture of a guy on a desk relates to a robot in the scene. My initial thought is, "Okay, let's see... what's the connection here?" First, I note the picture. It's a guy, bald, in a suit. He's smiling, and he's got a serious look on his face. Then, the video cuts to a robot, a large, blue, mechanical thing. It's in a lab, and it's talking. The timestamps are key here; the picture appears around 20 seconds, and the robot shows up much later, around 330 seconds. The robot is clearly a character in the story, and the picture is just a piece of background detail. Now, I have to consider the options. Option A, a rival executive? No, the robot isn't attacking anyone. Option B, a time-travel plot? Possible, but not the strongest link. Option C, the inventor? Maybe, but the robot is talking, so it's probably not the inventor. Option D, an AI replica? That's interesting. The robot is talking, and it's acting like a person. It's also in a lab, which suggests it might be a creation. The documents at the end seem to confirm this. So, I'm leaning towards option D. The robot is an AI, and it's designed to act like the man in the picture. The documents at the end will probably explain how this works. The picture is just a visual cue, a way to introduce the character. The robot is the AI version of the man. This makes the most sense given the context.</think>
<tool_call> {"name": "get_video_clip_frame", "arguments": {"start_time": 20.00, "end_time": 30.00}}</tool_call>

<think>Let's analyze the narrative and environmental cues between 0:00:20 and 0:00:30. In these frames, a framed photo of another man is clearly visible on the desk beside the protagonist. The setting is a professional office, and the photo is placed prominently, suggesting the man in the photo holds significance to the protagonist or the company. The protagonist is holding documents, indicating business or project matters are central to the plot. Later, a robot appears in the office (0:05:30-0:05:40). Given the office environment and the visible documents, it's reasonable to infer the robot is linked to a corporate or technological project. The placement of the photo—clearly visible and not just background decor—implies a narrative connection. Option A (rival executive) doesn't fit: there's no evidence of sabotage or rivalry. Option B (time-travel) is unlikely; the photo shows a different man, not a younger version of the protagonist. Option D (AI replica) would require more explicit hints that the robot is modeled after the man in the photo, but nothing in the scene suggests this. **Option C stands out: the photo is likely of the robot's inventor, someone important to the project but no longer present.** The protagonist's reaction to the robot and the documents further support the idea that the inventor died before the project's completion, and the robot's appearance is a direct continuation of that legacy. The narrative progression—from the photo, to the protagonist's focus on documents, to the robot's arrival—suggests the photo serves as a memorial and narrative anchor for the robot's origin. **Therefore, the most logical answer is C:** The photo shows the inventor of the robot, who died before completing the project the protagonist is reacting to.</think>
<answer>C</answer>

Figure 18. This example highlights the VITAL-7B model's ability to perform tool-augmented multimodal chain-of-thought inference for reasoning VQA tasks. By combining visual scene analysis with textual evidence and leveraging external tools for precise temporal grounding, the model systematically interprets narrative context to infer the relationship between the framed photo and the robot. This approach enables explainable, context-aware decision-making in complex, story-driven video understanding tasks.

Question: Please find the visual event described by a sentence in the video, determining its starting and ending times. The sentence: **The scrapbook holds an announcement about a pregnancy.**

Ground Truth: From 68.00 to 74.00 seconds.

Thinking without tools: From 67.02 to 117.04 seconds.

Thinking with tools: From 65.02 to 118.04 seconds.

IoU = 12.0%

Win!

IoU = 11.3%

Loss!



Get_video_clip(start=56.02, end=122.04)

Thinking without tools

```
<think>
The user wants to find the start and end times for the event where the scrapbook holds an announcement about a pregnancy.
1. I will carefully scan through the provided video frames to locate any mention or visual representation of a pregnancy announcement.
2. At frame `67.02s`, a photo of a pregnant woman holding an ultrasound image is shown. This confirms that the "special guest" is indeed a pregnancy.
3. At frame `71.02s`, the text "It was very hard keeping it a secret" appears, further emphasizing the pregnancy theme.
4. At frame `81.03s`, a photo of a woman drinking wine is shown, reinforcing this point.
5. At frame `95.03s`, a piece of paper is placed on the calendar, and at frame `96.03s`, the text "Our baby GIRL is due 14th February 2017" is clearly visible.
6. At frame `114.04s`, a photo of hands making a heart shape over a belly is shown, visually confirming the pregnancy.
7. At frame `117.04s`, the screen goes black, signaling the end of this specific section of the scrapbook.
Based on this analysis, the entire segment dedicated to the pregnancy announcement spans from when the pregnancy is first hinted at (around 67.02s) until the end of the reveal sequence (117.04s).</think>
<answer>The event happens in the 67.02 - 117.04</answer>
```

Thinking with tools

```
<think>Okay, here's what I've got. The user wants me to pinpoint the start and end times for the "announcement about a pregnancy" within this video. My approach is straightforward: I'll scan the video, looking for any text or visuals related to pregnancy announcements.
First pass: I see the video opens with a blank page, then a title card, and then a calendar. Then comes the wedding photos, followed by a photo of a pregnant woman holding an ultrasound image. That's definitely a strong contender! The text "Our baby GIRL is due 14th February 2017" pops up over the photo. This is the core of the announcement. After that, we have some heart-shaped hand gestures, then the video ends.
Now, let's refine this. The announcement *starts* when the photo of the pregnant woman appears, around `68.02s`. The text overlay appears at `70.02s`, but the photo itself is already present. The announcement *ends* when the next scene begins, which is `119.04s`.
So, the entire sequence, including the photo and the text, constitutes the announcement. Therefore, the relevant timeframe is approximately from `68.02s` to `119.04s`.</think>
<tool_call>{"name": "get_video_clip_frame", "arguments": {"start_time": 56.02, "end_time": 122.04}}
</tool_call>
```



```
<think>Okay, let's break this down. I'm tasked with pinpointing the start and end times for the "announcement about a pregnancy" within this video. My initial scan of the whole video gave me a rough idea, placing it somewhere between 68 and 119 seconds, based on the appearance of the pregnant woman and the ultrasound image. That was a quick overview.
Now, focusing on the provided clip, specifically from 56.02 to 122.04 seconds, I can get much more precise. At 64.02 seconds, the text "There was one special guest we didn't tell you about..." appears, setting the stage. Then, at 68.02 seconds, the actual announcement hits: "Our baby GIRL is due 14th February 2017". That's the core of it. The clip continues through various celebratory scenes, including confetti and heart-shaped hands, all reinforcing the joyous occasion. Finally, at 119.04 seconds, the screen goes black, signaling the end of the announcement.
Therefore, the announcement *begins* with the setup text at 64.02 seconds and *concludes* right before the black screen at 119.04 seconds. So, the most accurate range is from 64.02 seconds to 119.04 seconds. That encompasses the lead-in, the explicit statement, and the subsequent celebration, perfectly capturing the essence of the announcement.</think>
<answer>The event happens in the 64.00 - 119.00 seconds.</answer>
```

Figure 19. This example illustrates a **failure case** of the VITAL-7B model in tool-augmented multimodal chain-of-thought reasoning for temporal grounding tasks. While the model leverages external video clipping tool to densely sample video frames and generate structured reasoning steps, it still fail to recognize and localize the correct time range of the specified event.