VidText: Towards Comprehensive Evaluation for Video Text Understanding

Supplementary Material

A. Overview of Appendix

B. Limitations

We summarize the limitations of our work as follows:

- **Limited scenario coverage:** Although VidText includes 27 fine-grained video categories, it still lacks representation of long-tail or high-risk domains such as medical emergencies, industrial workflows, or disaster scenarios.
- **Imbalanced language distribution:** The majority of samples are in English and Chinese, with significantly fewer examples in other languages such as German, Korean, and Japanese. This imbalance prevents a thorough evaluation of multilingual OCR and reasoning capabilities.
- **Scarcity of challenging text instances:** VidText contains relatively few examples involving difficult text conditions such as severe occlusion, low resolution, motion blur, unusual fonts, or multi-line arrangements. This limits the benchmark's ability to fully assess model robustness under real-world noise and distortion.

C. Discussion

These dataset and model limitations are mutually reinforcing. Dataset gaps may conceal important weaknesses in current models, while existing models' deficiencies highlight the need for broader and more diverse benchmarks. Future efforts should focus on expanding long-tail scene and language coverage in VidText, while also improving LMM architectures with better multilingual OCR, noise robustness, and cross-modal reasoning abilities. Furthermore, we also summarize three insights as follows:

- **Weak cross-domain transfer:** Most LMMs are pretrained on image-based OCR tasks and struggle to generalize to unseen video scenes, such as sports broadcasts or livestream interfaces, where text appearance and context are highly dynamic.
- **Insufficient multilingual alignment:** Current models show limited ability in detecting, transcribing, and semantically linking non-English texts to the visual context, resulting in degraded performance on multilingual content.
- **Low robustness to visual noise:** Models often fail when confronted with noisy, blurry, or occluded text, particularly in tasks requiring instance-level grounding. This degrades downstream reasoning performance and reflects a need for stronger visual resilience.

D. Use of Large Language Models (LLMs)

Scope. We used LLMs *only for language polishing and light scripting assistance*. LLMs were not used to generate data, labels, model outputs, evaluation results, or figures.

Writing assistance. LLMs were employed to improve grammar, wording, and clarity of prose and to reformat LaTeX. All technical content (methods, formulas, hyperparameters, tables, and numbers) was authored and verified by the authors.

Scripting assistance. LLMs helped draft boilerplate code such as command-line wrappers, dataset loaders, or small utilities (e.g., argument parsing, logging). All scripts were *manually reviewed, edited, and tested* by the authors before inclusion. Final experimental pipelines are specified in our repository and Reproducibility Statement.

No synthetic labels or data. No LLM-generated text was used as ground-truth labels, dataset entries, or to augment training/evaluation data. Test annotations, metrics, and reported results are human-authored or taken from public benchmarks per their licenses.

Privacy and ToS. We did not upload private or license-restricted raw videos to third-party APIs. Any prompts contained no personally identifiable information. Our usage complies with data-source licenses and platform Terms of Service.

Reproducibility. All scripts produced with LLM assistance are fully documented and version-controlled; seeds, hyperparameters, and command entry points are fixed. Therefore, the use of LLMs does not affect experimental validity or reproducibility.

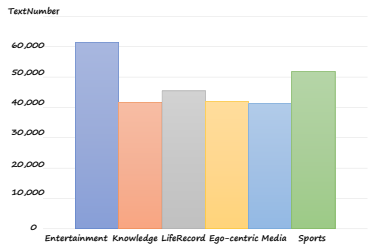

Figure 1. Text quantity distribution across six scene categories.

E. Broader Impact

The VidText benchmark is poised to make a significant contribution to both the OCR and video understanding communities by bridging the gap between low-level text perception [**? ? ?**] and high-level semantic reasoning [**? ?**] in video contexts.

For the OCR community, VidText offers a valuable opportunity to move beyond traditional image-based text detection and recognition [**? ? ? ?**]. By shifting the focus to temporal and contextual dynamics in videos, it promotes the development of algorithms that can track, ground, and interpret visual texts over time.

For the video understanding community, VidText introduces the underexplored yet semantically rich modality of scene text into the landscape of video-language research. By incorporating fine-grained text perception tasks and their paired reasoning counterparts, VidText pushes video-language models to integrate visual texts with multimodal contextual cues, fostering more explainable, interpretable, and grounded video understanding.

F. More details of VIDTEXT

Scene and language distributions. Fig. 1 illustrates the distribution of visual text quantity across six video scene categories. The largest number of text instances appears in **Entertainment** and **Sports**-related content, while **Knowledge** and **Media** are less dense in text content.For completeness, Tab. 1 reports the proportional breakdown of languages. The two largest—**English** and **Chinese**—account for **46.1%** and **32.3%** of the corpus, respectively, while **Japanese (8.2%)**, **Korean (7.0%)**, and **German (6.4%)** together make up the remaining **21.6%**. This skew toward high-resource languages suggests that models may generalize better on English/Chinese content, whereas performance on lower-resource languages could be constrained by data scarcity.

Video duration distribution. VIDTEXT exhibits a wide range of video durations, with an average length of 108.2 seconds. As shown in Fig. 2, this highlights the multi-duration characteristic of VIDTEXT, ensuring the temporal diversity needed to support both short-form and long-form video understanding tasks.

Table 1. Distribution of scene super-categories and languages in VIDTEXT.

Scene super-category	# Videos	Proportion
Media	192	20.4%
Knowledge	164	17.5%
Life-record	180	19.2%
Entertainment	123	13.1%
Ego-centric	101	10.8%
Sport	179	19.0%
Languages		
English	433	46.1%
Chinese	303	32.3%
Japanese	77	8.2%
Korean	66	7.0%
German	60	6.4%

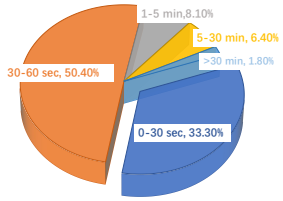

Figure 2. Video duration distribution in VIDTEXT.

Semantic content word cloud. To visualize the semantic richness and diversity of video–text interactions, we construct a word cloud using all questions and answers in VIDTEXT. As shown in Fig. 3, high-frequency words such as *text*, *video*, *content*, and *EXIT* reflect a strong alignment between text and semantic reasoning. The co-existence of spatial keywords (e.g., *LEFT*, *RIGHT*), functional terms (e.g., *score*, *speed*), and contextual references (e.g., *player*, *talent*) highlights the multi-granular reasoning needs of the dataset.

G. More ablation studies

G.1. Further impact analysis of key design choices

Impact of Video Duration To investigate the influence of video duration on various tasks, we grouped videos

Figure 3. Word cloud of all questions and answers in VIDTEXT.

Table 2. Impact of video duration and segment cropping.

Impact of Video Duration				Impact of Segment Cropping on TCR			
Duration Range	**Text Loc. ↑**	**OCR ↑**	**Reasoning ↑**	**Model**	**Full Video**	**Segment**	**Δ**
0–30 s	29.81	27.07	45.45	VideoLLaMA3	43.33	56.67	(↑13.34)
30–60 s	29.01	24.25	54.55	Qwen2.5-VL	26.67	33.33	(↑6.66)
1–5 min	14.45	30.21	42.86				
5–30 min	8.28	18.31	46.57				
≥30 min	0.00	0.00	33.33				

into five duration intervals and evaluated four representative tasks: Text Localization, Temporal Causal Reasoning, Local OCR, and Local Reasoning. The results are shown in Tab. 2.

Observation. Perception-heavy tasks (e.g., Text Localization and OCR) suffer a significant performance drop as video length increases. This indicates the challenge of long-range visual-text dependency modeling. Reasoning tasks also degrade but show more fluctuation. In particular, Temporal Causal Reasoning performs unexpectedly poorly on 30–60 s videos.

We hypothesize this may not only be due to duration, but also due to difficulty in locating the relevant temporal segment where the causal text appears. To verify this, we conduct an additional ablation study: for each sample, we crop the ground-truth segment and feed only that segment into the model. The results in Tab. 2 demonstrate that models perform significantly better when provided with only the relevant segment, confirming that **text localization** is a key bottleneck.

This ablation emphasizes that improving temporal grounding and retrieval capabilities is essential for long-form causal reasoning tasks.

Language diversity ablation. We analyze Holistic/Local OCR across languages (Tab. 3) for Qwen2.5-VL-7B and VideoLLaMA3-7B. We observe **notable performance drops** on lower-resource languages (Korean, Japanese, German) versus high-resource ones (English, Chinese), suggesting bias from training data distributions.

Alignment with training distribution (controlled fine-tuning). Since most model developers do not disclose language ratios, we conduct a controlled study on LLaVA-One-Vision by incrementally adding multilingual OCR data during instruction tuning: Phase 1 (IC13+IC15; English), Phase 2 (+ReCTS+TextOCR; add Chinese), Phase 3 (+MLT17+MLT19; add Korean/Japanese/German). We keep the original video instruction data to preserve video ability. Tab. 4 (Holistic OCR) shows consistent gains with multilingual supervision, especially on lower-resource languages.

G.2. Model-intrinsic factors

We ablate three influential factors (Tab. **??**): **(i) Input resolution**: Increasing resolution improves video-text performance, especially for models that partition inputs into sub-patches (e.g., InternVL2.5), which better preserve text details at higher resolutions. **(ii) OCR ability**: Performance aligns with standard OCR benchmarks (e.g., OCRBench), indicating that fundamental OCR competence is predictive for video text understanding. **(iii) LLM backbone**: Some backbones (e.g., Qwen2.5) show stronger multilingual performance than LLaMA-based variants, suggesting language modeling strength matters alongside perception.

G.3. Chain-of-Thought (CoT) Strategy: Effectiveness and Scope

Why it works. CoT boosts performance mainly due to two factors: (i) it decouples complex video–text reasoning into two sub-problems, *visual text spotting* and *video content understanding*, allowing the model to focus on one aspect at a time; (ii) it mitigates the temporal aggregation bottleneck in current Video LLMs by segmenting long videos and enabling localized reasoning within clips. For global tasks like Holistic Reasoning, CoT processes long videos sequentially, clip by clip; for local tasks (e.g., Local Reasoning, Temporal Causal Reasoning), it transforms the search space from vision to text, using a retrieval-like strategy to locate relevant cues.

Ablations. We design two ablated CoT variants to understand its impact(Tab. 5: (1) *Partial-Task CoT*, where prompts only include either video summarization (VS) or visual text spotting (VTS); and (2) *Partial-Time CoT*, where the model reasons over the full video instead of clip-wise segments. Results show that the full CoT strategy outperforms both ablations, demonstrating that joint modeling of segmented temporal reasoning and dual-modality decomposition is essential for complex video text understanding.

Scope and limitations. Despite its effectiveness, CoT introduces extra computation (video segmentation and intermediate generations), increasing inference latency. Moreover, its performance depends on the base model's perception and generation capacity: stronger models yield bet-

Table 3. OCR performance (%) across languages. Comparison between Qwen2.5-VL and VideoLLaMA3 on local and holistic OCR tasks.

Language	Qwen2.5-VL↑		VideoLLaMA3↑	
	Local OCR	Holistic OCR	Local OCR	Holistic OCR
English	49.3	45.2	45.5	32.1
Chinese	26.3	25.4	24.5	22.6
Korean	17.6	16.4	18.9	13.5
Japanese	22.8	23.1	23.1	18.3
German	14.5	12.4	14.8	11.2

Table 4. Holistic OCR vs. multilingual supervision during instruction tuning (LLaVA-One-Vision).

Language	Phase 1	Phase 2	Phase 3
English	34.2	36.7	39.2
Chinese	22.3	26.4	29.1
Korean	12.1	13.5	16.0
Japanese	10.0	11.2	14.2
German	8.5	8.2	14.3

Table 5. Ablation of CoT strategies on Qwen2.5-VL-7B. HR, LR, TR, SR denote Holistic, Local, Temporal, and Spatial Reasoning respectively.

Method	HR ↑	LR ↑	TR ↑	SR ↑
Baseline	36.0	26.5	35.4	35.2
Partial-Time CoT	36.2	26.3	35.4	35.3
Partial-Task CoT (VS only)	37.3	27.1	35.9	37.2
Partial-Task CoT (VTS only)	38.5	27.5	36.8	38.4
Full CoT (ours)	**40.5**	**28.7**	**37.2**	**40.9**

ter CoT-based reasoning.

H. Collecting Details of VIDTEXT

This section outlines the procedures for sourcing, filtering, and analyzing the video content in VIDTEXT.

Sources. To ensure a broad coverage of video scenarios and textual styles, VIDTEXT integrates data from six public datasets:

- **BOVText** [**?**] — Multi-scene videos suitable for holistic OCR tasks.
- **RoadText-1K** [**?**] — Dense road-text detection in driving scenarios.
- **DSText** [**?**] — Subtitles from indoor instructional videos.
- **M4-ViteVQA** [**?**] — Clip- and instance-level multimodal QA videos.
- **Video-MME/MLVU** [**? ?**] — Long-form videos with strong temporal reasoning demands.

YouTube supplementation. To supplement long-form data, we collect additional videos from YouTube, focusing on the following categories:

- **Sports highlights:** NBA, FIFA World Cup, and related competitions.
- **Gaming commentary:** live streams and post-game analysis.
- **TV shows and variety entertainment.**

Retrieval and filtering criteria. Candidate videos were retrieved using targeted keyword queries such as *"match subtitles"*, *"game commentary"*, and *"captioned recap"*. We applied the following filtering rules:

- **Minimum duration:** $\geq$3 minutes for YouTube, $>$30 minutes for Video-MME.
- **Scene-text richness:** We use the latest detector **Gomatching** [**?**] to calculate the proportion of frames containing text.
- **Density thresholds:** Videos must meet a minimum ratio of text-bearing frames: **20%** for YouTube videos and **10%** for Video-MME.

Metadata statistics. We also collect metadata such as video length, resolution, and frame rate to ensure coverage diversity across temporal and visual characteristics.

I. Details of Annotation

I.1. Instance annotation

Each video underwent a two-stage text annotation process. In the first stage, annotators drew tight bounding boxes around visible text lines and assigned each to a category: *ClearText* or *Illegible*. A tracking tool automatically propagated bounding boxes across frames using consistent Track IDs. More details are shown in Fig. 4.

Instance Annotation Guidelines

Step 1: Text Detection (Bounding Box)
1.Draw a bounding box around 3-5 visible text instance in each video.
2.Annotate entire text lines, not individual words or characters.
3.If the same text appears across multiple frames, assign it the same Track ID using the tracking tool provided.
Step 2: Text Classification
Each bounding box must be assigned one of the following text categories:
Clear Text:clearly visible text.
Illegible: text that is unreadable due to blur, occlusion, or low resolution.
Step 3: Text Transcription
All Text instances require transcription.
For tracked text across multiple frames, you only need to transcribe it once—the tool will propagate it across the track automatically
Special Handling: Blur or Occlusion
If a text instance becomes blurred or occluded:
If the blur/occlusion lasts 3 frames or fewer, continue the original track.
If it lasts more than 3 frames, end the current track and create a new one labeled as Illegible.
If a text transitions from unreadable to readable (or vice versa), create a new track with the updated label

Figure 4. Instance-level annotation guidelines.

I.2. Clip-level annotation

Videos shorter than 1 minute were split into 5-second clips; longer ones into 20-second clips. For each clip, annotators recorded all visible, legible text and its temporal span. Repeated instances within a clip were marked only once. Illegible or heavily blurred texts were ignored. More details are shown in Fig. 5.

I.3. Video-level text collection

A separate annotation team reviewed the OCR predictions from our model. Annotators removed hallucinated content and added missing instances. Chinese was annotated by full lines; other languages (e.g., English, German) were annotated by words. Each unique string was listed once in the final inventory. More details are shown in Fig. 5.

I.4. Holistic reasoning

Annotators watched the full video and consulted the video-level text inventory to write one multi-label question per video (see Fig. 6). Each question included seven options describing high-level semantics such as scene, role, topic, or sponsor.

I.5. Local reasoning

For every clip (as defined in D.2), annotators created one four-option multiple-choice question requiring reasoning between localized text and visual context (e.g., subtitle or character behavior). The question must require multimodal reasoning and not be solvable using text or image alone. More details are shown in Fig. 7.

I.6. Temporal causal reasoning

Given a reference text (e.g., scoreboard or subtitle), annotators identified the timestamp of its appearance, observed the following 3–30 seconds, and formulated a causal reasoning question. The answer was a single factual sentence describing the resulting action. Each QA pair was anchored to the cue's timestamp. More details are shown in Fig. 8.

I.7. Spatial reasoning

As shown in Fig. 9, at a given timestamp, annotators located a reference text or entity and constructed a question requiring reasoning over its spatial relation to nearby visual elements (e.g., direction, proximity, interaction).

Quality control. All annotations underwent double review. Each item was cross-validated by a second annotator, and disagreements were resolved by expert adjudication. On a random sample of 200 items, we achieved an average inter-annotator agreement of **0.81** (Cohen's κ), indicating high reliability.

Clip & Video-Level Annotation Guidelines

Clip Level:

Video Segmentation

1.Divide the video into consecutive temporal segments： If the video is shorter than 1 minute:divide it into clips of **5 seconds** each.Else; divide it into clips of **20 seconds** each.
2.Each segment should have a clear start_time and end_time.

Text Identification

1.For each clip, annotate all readable text instances that appear within the clip's time span.
2.Ignore illegible, blurred, or heavily occluded text.
3.If the same text appears multiple times within the clip, annotate it only once.

Video Level:

Global Text Collection

1.Watch through the full video and record all clearly visible and legible text content.
2.You will be provided with a preliminary list of detected texts (from an automatic text detection model). In this case, carefully review and correct the list by adding missing texts and removing false positives to ensure accuracy.
3.Each unique text instance should be annotated only once (no need to mark repetitions).

Language-Based Annotation Rules

1.For Chinese text: annotate by complete text lines (e.g., subtitle or sign line).
2.For Non-Chinese languages (e.g., English, German): annotate by individual words.
3.For mixed-language cases, follow the dominant language rule and note exceptions when needed.

Figure 5. Clip- and video-level annotation guidelines.

Annotation Guidelines for Holistic Reasoning

goal:: Given the overall textual and visual content throughout the video—including information across multiple time segments—annotate a global question that requires semantic reasoning across time and space.

Input

You will be given the full video and its OCR transcription.
Your goal is to:Observe the entire video, noting important text and visual elements across different timepoints.Identify high-level topics, roles, actions, or patterns that emerge over time.
Create a multi-option question that tests understanding of the video's overall narrative or semantic structure, including content distributed across time.
Select 3 correct options from a set of 7 plausible answers.

CoT Expectation:

You should simulate how a model would connect multiple distributed cues,
such as:"The subtitle shows the name + stage text shows show name + outfit = talent show"
"Multiple timepoints include branding (e.g., sponsor, stage banner) → context clue"
"Introduction + mid-performance + audience shot = global understanding of scene"

GOOD EXAMPLE:

Question： Based on the video text and description, which of the following statements accurately describe the scene and content of the video?
"A": "The young performer is identified as a 12-year-old talent from a rural background.",
"B": "The show being referenced is \"**中国达人秀**\" (China's Got Talent).",
"F": "The show features a challenge round sponsored by \"**海飞丝**\"

Figure 6. Holistic reasoning annotation guidelines.

Annotation Guidelines for Local Reasoning

goal:: Within a specific time segment of the video, reason over the text and visual context to answer a multimodal question grounded in localized semantics.

Input

You are given a **specific video segment** along with:
•Detected OCR text within the segment
•The corresponding video frames
Your task is to:
Understand the meaning and context of the visible text in the clip.Interpret surrounding visual content (e.g., characters, objects, layout)
Construct a multiple-choice question that tests the model's semantic understanding and reasoning ability
Provide 4 candidate options and select the correct answer

CoT Expectation:

Ask: what does the text cause / reflect / imply?
Simulate the model making the connection:
"If the subtitle says 'stay still', and the character hides behind a wall → he's afraid / threatened"

GOOD EXAMPLE:

Q: "In the clip, the text 'Final Round' is shown. What does it suggest about the competition?"
A: "The winner will be decided in this match."
Q: "When the subtitle says 'Don't move', what is the person doing?"
A: "They are hiding quietly behind the shelf."

Figure 7. Local reasoning annotation guidelines.

Annotation Guidelines for Temporal Causal Reasoning

goal: Track a specific text instance in the video, analyze the sequence of related events, and annotate a question–answer pair that reflects their causal relationships.

Input

1.Locate the reference text
•Find the timestamp where the given text appears clearly (e.g., scoreboard, sign, subtitle).
•Pause at that moment and record the text content and timestamp.
2.Observe what happens next
•Watch the following 3–30 seconds of the video.
•Identify any actions, changes, or reactions that may be caused by or related to the text.
3.Write the QA pair
Question: Frame a question that highlights the relationship (e.g., "what happened after…" / "how did the player respond to…").
Answer: Describe the actual action concisely and factually.

CoT Expectation:

you should consider the temporal progression: what happened after the text appeared, and why it might be related.
Example: a low score triggers a coach's timeout; a red light prompts braking.

GOOD EXAMPLE:

"question": "At a score of 105:83, what move did James make to score?",
"answer": "He stole the ball and dunked it."
"question": "At the beginning of the game when the score was 0:0, how did the Warriors player score while being defended by Player 1?",
"answer": "By scoring a three-point shot",

Figure 8. Temporal causal reasoning annotation guidelines.

Annotation Guidelines for Spatial Reasoning

goal: At a specific timestamp, infer the spatial relationship between a text instance (or person) and surrounding visual elements—such as direction, relative position, or interaction.

Input

1.Locate the reference text

•Find the timestamp where the given text appears clearly (e.g., scoreboard, sign, subtitle).

•Pause at that moment and record the text content and timestamp.

2.Observe what happens next

•Watch the following 3–30 seconds of the video.

•Analyze the scene: what object or person is near, behind, or interacting?

3.Write the QA pair

Compose a multiple-choice or open-form reasoning question and answer

CoT Expectation:

you should consider Reason about spatial layout: who is positioned where, and what action is implied.

Use directional and functional cues: “behind”, “to the right”, “blocking”, “following”.

GOOD EXAMPLE:

"question": "When the score was 31:18 and 2:09 remained in the game, where was Player 8 located when attempting the three-point shot?",

"answer": "Bottom-middle of the image, right 45-degree three-point position"

"question": "With the score 0:0, who is the player defending Timber-wolves' Player 5 (white jersey)?",

“answer": "Player 31"

Figure 9. Spatial reasoning annotation guidelines.

J. Details of Experimental Settings

J.1. Model configuration

We outline the primary baselines evaluated on VIDTEXT. To ensure fair comparison across both open- and closed-source models, we explicitly standardize frame sampling and spatial resolution for each baseline, as summarized in Tab. 6.

For proprietary models such as GPT-4o, Gemini 1.5 (Pro and Flash), and GPT-4-Turbo, we follow their official or API-supported settings. `GPT-4o` supports up to ∼500 image inputs, for which we adopt a uniform sampling rate of 0.5 fps with an input resolution of 512×512 to accommodate most of our videos. `GPT-4-Turbo` is restricted to 16 frames, uniformly sampled across the video, and resized to the same resolution.

For open-source models, we align each configuration with their original public implementations. `VideoChat-Flash`, `Qwen2-VL (7B)`, and all `Qwen2.5-VL` variants (3B/7B/72B) operate under a 1 fps sampling strategy, with a maximum of 768 frames per video. Models supporting extended temporal contexts—such as `VideoLLaMA 3`, `InternVL 2.5`, and `LLaVA-OV`—are provided with 64 uniformly sampled frames, resized to 336×336. `ShareGPT4Video` also uses 64 frames, but with a reduced spatial resolution of 224×224. `LongVU` and `LongVA` are evaluated with sparse and extended frame settings. `LongVU` uses 1 fps sampling, while `LongVA` accepts up to 128 uniformly distributed frames. `MiniCPM-V2.6` applies a fixed 64-frame sliding window, following its official implementation.

J.2. Human performance study

To assess the upper-bound of performance on VIDTEXT, we conducted a controlled human evaluation across all tasks in our benchmark. Three annotators with experience in video analysis and text recognition were recruited to answer a representative subset of questions spanning all eight task types. Each participant was given access to the full video content and instructed to answer using their best judgment, without time constraints. The average human accuracy across all tasks reaches **89.5%**, substantially outperforming all evaluated models. In particular, humans demonstrated near-perfect scores in holistic and local OCR, reasoning, and spatial understanding tasks, highlighting the gap between human-level comprehension and the capabilities of current multimodal large models. These results serve as a reference ceiling for future model development and underline the complexity and nuance of the video-text understanding challenges posed by VIDTEXT. More details are shown in Tab. **??**.

Table 6. Frame–sampling and input-resolution settings for baselines.

Model	Size	Sampling	Resolution
Proprietary MLLMs			
GPT-4-Turbo	–	16 frames	512^2
Gemini 1.5 Flash	–	1 fps	512^2
GPT-4o	–	0.5 fps	512^2
Gemini 1.5 Pro	–	1 fps	512^2
Open-source MLLMs			
LongVU	3 B	1 fps	448^2
Qwen2.5-VL	3 B	1 fps	448^2
Video-XL-Pro	7 B	1 fps	448^2
LongVA	7 B	128 frames	–
MiniCPM-V2.6	7 B	64 frames	448^2
VideoChat-Flash	7 B	1 fps	448^2
Qwen2-VL	7 B	1 fps	448^2
Qwen2.5-VL	7 B	1 fps	448^2
VideoLLaMA 3	7 B	64 frames	336^2
ShareGPT4Video	8 B	64 frames	224^2
Oryx-1.5	32 B	64 frames	336^2
LLaVA-OV	72 B	64 frames	336^2
Qwen2.5-VL	72 B	1 fps	448^2
InternVL 2.5	78 B	64 frames	336^2

J.3. Experiment environment

All experiments are conducted on a server equipped with 4×NVIDIA A100 GPUs (80GB each). Model inference and evaluation are implemented in PyTorch with mixed-precision support.

K. Model prompts

Fig. 10 shows the prompt template used to obtain detailed frame-level captions from the Aria model. The prompt includes instructions to describe the scene, detect visible text, summarize actions, and relate them spatially and semantically. Tab. 7 lists the standardized prompt templates used for each task in VIDTEXT.

L. More visualization results

We present additional visualizations of our VIDTEXT annotation examples in Fig. 11, Fig. 12, and Fig. 13.

M. Ethics and Responsible Dataset Use

M.1. Consent and Compensation for Human Annotators

Our human evaluation involved **three graduate research assistants**. Prior to participation, all annotators were provided with an **Information Sheet** that clearly explained: (i) the **research purpose**; (ii) the nature of the **data to be viewed**; (iii) estimated **workload and duration**; (iv) **voluntary participation** with the right to withdraw at any time; and (v) assurance that **no personal information** would be collected and that results would be reported in **aggregate form**. All annotators **gave informed consent** by signing the document and were compensated at **$25/hour**, following our institution's standard rate for annotation tasks. No demographic or identity-related information was collected, and no audio/video recordings were made. All annotation files are stored on **encrypted drives** accessible only to the author team.

Based on international guidelines (**US 45 CFR 46**, **EU GDPR**, and our institution's ethics policy), this activity is classified as **Not Human-Subjects Research (Not-HSR)** or **Exempt Category #4 (publicly available / anonymized data)**. We completed an **internal ethics self-assessment**, and will include the **signed consent forms** and self-review documentation in the supplementary materials.

M.2. Privacy and Copyright Compliance for YouTube Videos

Regarding the **76 YouTube videos** (sports and esports content) used in our dataset: (i) all videos are sourced from **publicly accessible broadcasts** where faces are generally indistinct; (ii) for any identifiable individuals, we apply **automatic face blurring** and **crop out channel identifiers**; (iii) derived data are **downsampled in resolution** and **watermarks removed**; (iv) the dataset is released strictly for **non-commercial academic research** under the **CC-BY-NC-SA 4.0** license.

We adopt a **takedown policy**: a dedicated **contact email** will be provided on the dataset homepage and GitHub; upon request from content creators or copyright holders, we will **remove the relevant video within 48 hours**; in cases of full takedown, we will retain only **sparse sampled frames** or **annotations/metadata**, following the practice of datasets like **MovieNet** [**?**]. A detailed **Usage and Takedown Policy** is included in the supplementary materials to ensure **privacy protection**, **responsible use**, and **copyright compliance**.

Aria Caption Generation Prompt

You are given images sampled from a video. Please imagine yourself in the scene and describe in detail what you see from your viewpoint. Your description should focus on the following aspects:
1. What is the overall scene or environment?
2. What visible objects or people are present?
3. Are there any texts (e.g., signs, labels, instructions)? If yes, what do they say?
4. What activities or actions are happening in the scene?
5. Are there any meaningful relationships between the scene texts and the objects, people, or actions around them?
Please write the description in a natural and informative way, as if explaining what you are currently seeing. Avoid mentioning “image” or “frame”, and do not speculate beyond what is visible.
Output format:
- Scene description: [...]
- Visible texts: [...]
- Human and object activities: [...]
- Spatial or semantic relationships (if any): [...]

Figure 10. Prompt template used for Aria to generate frame-level captions.

Table 7. Prompt templates used for VIDTEXT tasks.

Task	Prompt template
Holistic OCR	”Recognize all visual texts in the video. If the text is not in English, do not provide an English translation. Do not include any descriptions, narrative, or context. Output only the extracted text lines, each on a new line.”
Holistic Reasoning	”Watch the video carefully and select the correct three answers. Question: {question} Options: {options} Please output your answer in the format: `Correct Answers: A, B, C`”
Local OCR	”Watch the video and answer the following question based on its content. Question: {question} Please output only the texts that appear in the specified time interval as a JSON array of strings, with each element representing one piece of text. Do not include any additional description or translation.”
Local Reasoning	”Watch the video and answer the following multiple-choice question based on its content. Question: {question} Options: Option A: {text} Option B: {text} ... Please select the correct option.”
Text Localization	”Watch the video and answer the following question based on its content. Please provide the time interval (in seconds, precise to 0.1s) during which the text appears in the video. Output your answer in JSON format with keys ’start’ and ’end’. For example: {`"start": 0.0, "end": 30.0`}. Do not include any extra commentary.”
Temporal Causal Reasoning	”Watch the video and answer the following multiple-choice question based on its content. Question: {question} Options: Option A: {text} Option B: {text} ... Please select the correct option.”
Text Tracking	*(Same prompt as Spatial Reasoning)*
Spatial Reasoning	”Watch the video and answer the following multiple-choice question based on its content. Question: {question} Options: Option A: {text} Option B: {text} ... Please select the correct option.”

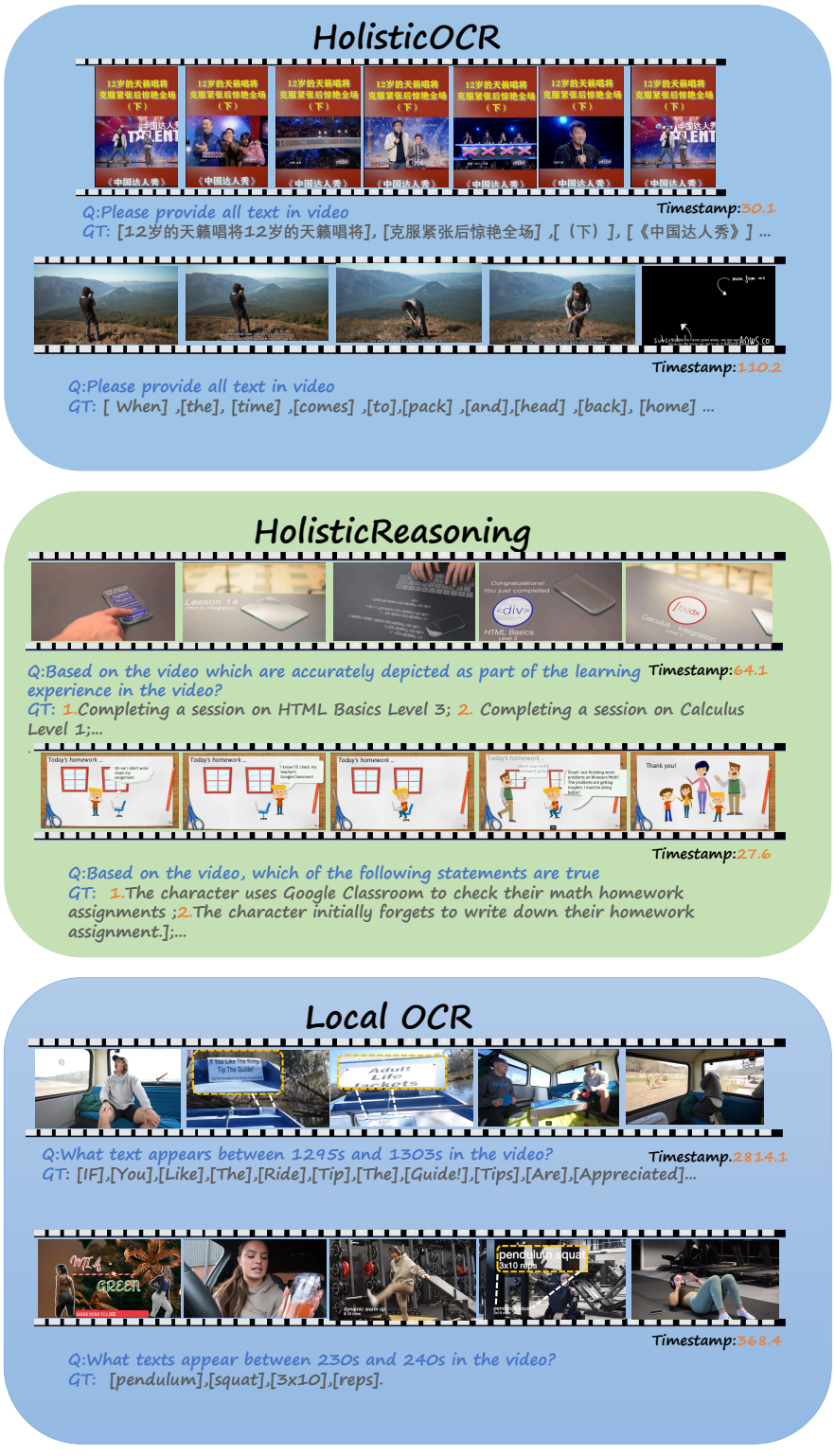

Figure 11. (Top) more examples of HolisticOCR. (Middle) more examples of HolisticReasoning. (Bottom) more examples of LocalOCR.

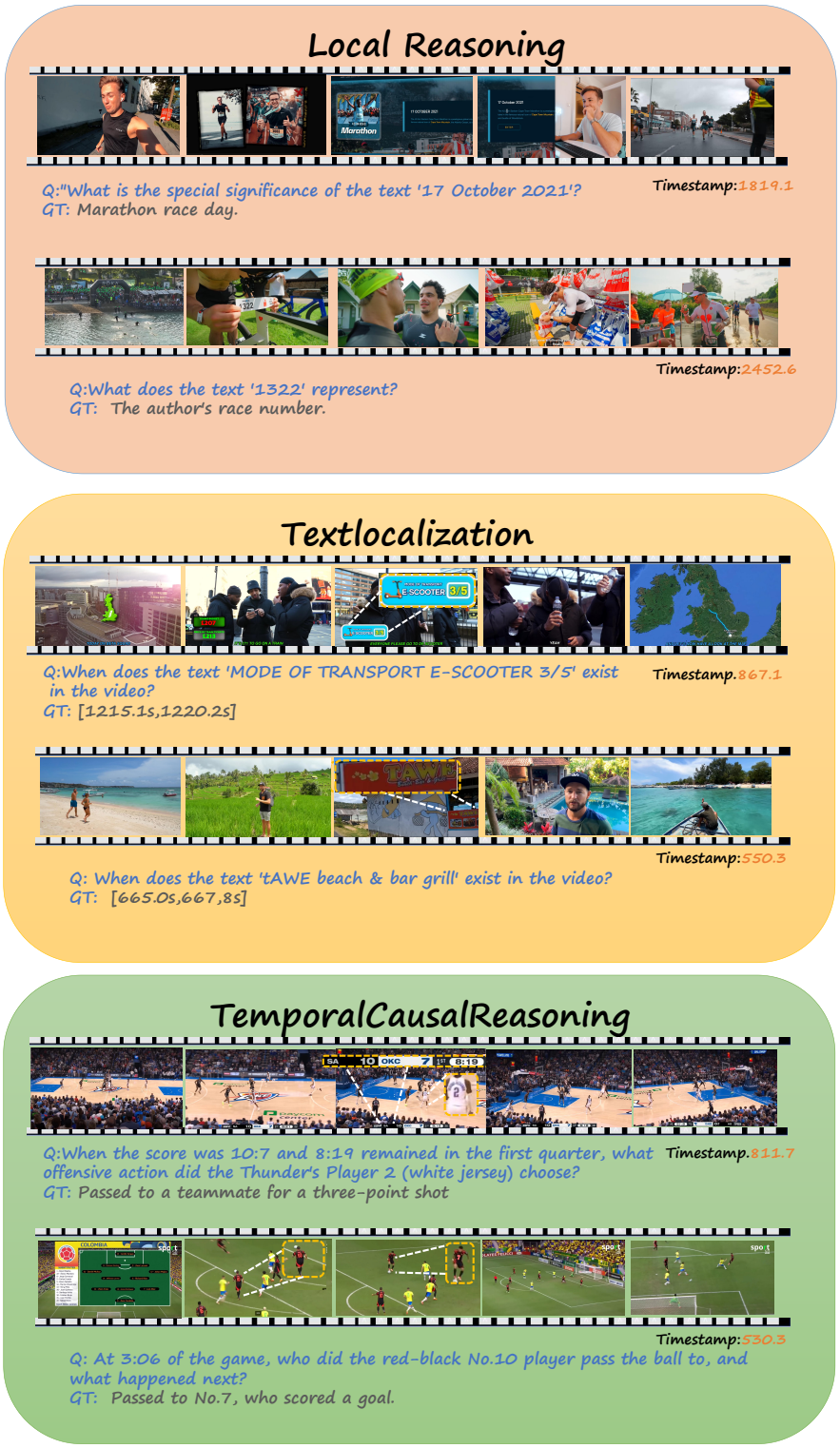

Figure 12. (Top) more examples of LocalReasoning. (Middle) more examples of TextLocalization. (Bottom) more examples of Temporal-CausalReasoning.

Figure 13. (Top) more examples of TextTracking. (Bottom) more examples of SpatialReasoning.