ViTexQA: A Multi-Frame Temporal Perception Dataset for Video Text Question Answering

Zhentao Guo[1*], Chen Duan[1*], Tongkun Guan[2*], Zining Wang[1], Kai Zhou[1(✉)], and Pengfei Yan[1]

[1] Meituan, China {guozhentao, duanchen, wangzining, zhoukai, yanpengfei}@meituan.com
[2] Shanghai Jiao Tong University, China gtk0615@sjtu.edu.cn
https://github.com/ZhentaoGuo/ViTexQA

Abstract. Despite remarkable progress in multimodal understanding, current MLLMs still exhibit limitations in video text understanding, particularly when semantics emerge through the integration of temporally distributed textual cues across multiple frames. This perception challenge fundamentally differs from static image text understanding, yet existing datasets fail to capture: the vast majority of questions remain answerable from single frames, inadequately reflecting real-world video text comprehension demands. To address this, we present ViTexQA, a large-scale video-text QA dataset, and FrameThinker for robust multi-frame temporal reasoning. We build ViTexQA via a quality-controlled Chain-of-Thought (CoT) annotation pipeline boosted with temporal constraints; all its QA pairs demand cross-frame text fusion to solve, enforcing true temporal reliance. FrameThinker adopts two-stage training for explicit temporal modeling: CoT-Guided Supervised Fine-Tuning (SFT) generates frame-aware reasoning chains, followed by Temporally-grounded Reinforcement Learning (RL) optimized with multi-frame coherence rewards. Evaluations show our method outperforms SOTA baselines on ViTexQA, lifting ROUGE-L by 6.3%.

Keywords: Multi-model Large Language Models · Video Text Datasets · Multi-frame Temporal Perception

1 Introduction

The rapid development of Multimodal Large Language Models (MLLMs) [2, 5, 15, 17, 23, 48, 53, 55] has enabled unified approaches to diverse vision-language tasks, from static image understanding [11, 13, 14] to dynamic video comprehension [22, 25, 28, 30, 41, 47, 49]. Among these capabilities, the ability to understand text within visual content has proven essential for real-world applications. While substantial progress has been made in image-based scene text understanding through benchmarks like DocVQA [33], TextVQA [43], and ST-VQA [3], the challenge of comprehending text in videos remains fundamentally different.

*Equal contribution. ✉Corresponding author.

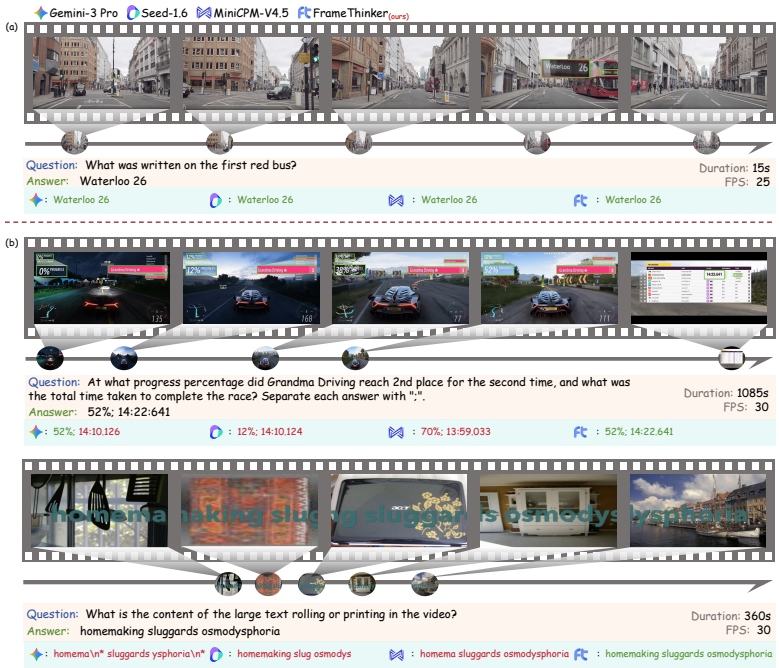

Fig. 1: Single-frame and multi-frame perception examples. **(a)** Single-frame perception, where the answer is derived from textual content within a single frame (such as reading "Waterloo 26" from the red bus.). All compared models answer correctly on such single-frame perception questions. **(b)** Multi-frame perception, which requires the model to integrate textual, spatial, and temporal cues across multiple frames (such as finding the answer to the question based on the progress of the competition, ranking changes, and time spent.). In contrast, while other models frequently produce incorrect or incomplete answers, our **FrameThinker** consistently arrives at the correct answer.

Videos introduce a critical temporal dimension that distinguishes them from static images. Textual information in videos often carries semantics that emerge only through temporal integration rather than appearing complete in isolated moments. Consider sports broadcasts where understanding the game progression requires tracking score changes across frames, news programs where complete stories unfold through sequential caption updates, or instructional videos where procedural knowledge accumulates through temporally distributed textual cues [7, 20]. In these scenarios, comprehension requires synthesizing information dispersed across multiple timestamps to construct complete semantic interpretations. As shown in Fig. 1, while single-frame questions can be answered by examining one snapshot, multi-frame perception demands integrating evidence distributed throughout the video timeline to derive meaningful answers.

Despite the importance of temporal perception over textual content, existing video text QA datasets exhibit a critical limitation. Our statistical analysis reveals that only approximately **13%** of questions in M4-ViteVQA [58] and **18%** in EgoTextVQA [59] genuinely require multi-frame perception. The overwhelm-

ing majority can be answered from single frames, failing to capture the temporal complexity inherent in real-world video understanding. This gap reflects a broader issue in current benchmark design: datasets inadvertently allow models to bypass temporal perception by relying on static, frame-level perception. Consequently, MLLMs trained and evaluated on these benchmarks develop limited capabilities for cross-frame integration, hindering their effectiveness in scenarios demanding true temporal comprehension.

To address this fundamental gap, we introduce ViTexQA, a large-scale dataset explicitly designed to require multi-frame temporal perception over textual content in videos. Unlike existing benchmarks where temporal perception remains optional, ViTexQA ensures that every question is deliberately unanswerable from any single frame alone. Each question necessitates integrating textual cues distributed across multiple timestamps, forcing models to engage in genuine temporal perception rather than frame-level pattern matching. This design philosophy reflects authentic video understanding demands where critical information emerges only through temporal synthesis.

Specifically, ViTexQA is constructed through two complementary pipelines that together address both evaluation and training needs for multi-frame temporal perception. The Video Dataset Construction pipeline ensures genuine multi-frame temporal dependencies through rigorous data collection and quality controlled human annotation, producing 5,147 high-quality videos with 6,864 QA pairs where each question is verified to require cross-frame integration. The Multi-frame perception CoT Construction pipeline further enriches these QA pairs with temporally-grounded CoT annotations that explicitly preserve spatiotemporal textual information across video frames. Together, these pipelines yield a comprehensive dataset spanning approximately 363 hours across diverse domains including sports, news, tutorials, and everyday scenarios. Critically, we adopt a human-centric approach throughout: all annotations are created exclusively by human annotators without automated assistance from GPT or similar systems, ensuring alignment with natural human perception patterns and genuine challenge for current MLLMs [12, 22, 25, 30, 41, 44].

Building upon ViTexQA, we further propose FrameThinker, a systematic method to enhance multi-frame temporal perception capabilities in MLLMs. FrameThinker addresses the temporal perception challenge through a two-stage training paradigm. First, CoT-Guided SFT trains models to generate temporally-grounded perception chains that explicitly enumerate textual evidence across frames with spatial and temporal grounding. Second, Temporally-grounded RL refines these capabilities through composite reward mechanisms that encourage temporal coherence, multi-frame dependency, and accurate evidence aggregation. This approach enables models to develop explicit temporal perception strategies rather than relying on implicit frame-level correlations.

Finally, we conduct extensive experiments evaluating state-of-the-art MLLMs on ViTexQA. Results reveal substantial performance gaps compared to human accuracy, highlighting fundamental limitations in current models' temporal perception abilities. Furthermore, FrameThinker achieves significant improvements

over strong baselines, demonstrating the effectiveness of explicit temporal perception training. These findings underscore both the challenge posed by genuine multi-frame temporal perception and the potential for systematic capability enhancement through appropriate training methodologies.

Our contributions are summarized as follows:

- We present ViTexQA, the first large-scale dataset tailored for multi-frame temporal awareness, constructed through dual complementary pipelines producing both rich question-answer pairs and corresponding temporally-grounded CoT annotations via rigorous human-centric annotation.
- We propose FrameThinker, a training methodology that systematically enhances multi-frame temporal perception through CoT-guided SFT and Temporally-grounded RL.
- Comprehensive benchmarking on ViTexQA reveals critical gaps in current MLLMs' temporal perception capabilities. In contrast, extensive experiments demonstrate that FrameThinker effectively mitigates these limitations, significantly outperforming strong baselines and validating the effectiveness of explicit temporal perception training for video text understanding.

2 Related Work

Image-based text VQA datasets OCR-related datasets are among the most important benchmarks for evaluating MLLMs in visual text understanding. In the OCR domain, a variety of datasets have been proposed to evaluate different aspects of visual text understanding. Natural scene text VQA datasets, such as TextVQA [43], ST-VQA [3], EST-VQA [46], and OCR-VQA [34], focus on question answering over images containing text in natural scenes, where the model must integrate scene understanding with optical character recognition. Document and chart understanding tasks address more structured and information-dense data types. For example, InfoVQA [32] requires models to jointly reason over textual content, graphical elements, and data visualizations; ChartQA [31] is designed for question answering over charts, demanding numerical perception and chart-specific knowledge; DocVQA [33] focuses on document-level text understanding and retrieval-based question answering. Web-based reading comprehension datasets, such as WebSRC [4], target the understanding of web page layouts and interactive elements. In addition, several comprehensive OCR benchmarks have been introduced, such as OCRBench [29], OCRBench v2 [10], and SEED-Bench-2-Plus [21], which integrate multiple OCR-related tasks to provide a holistic evaluation of MLLMs' text understanding capabilities [1, 18].

Video-based text VQA datasets While image-based Text VQA has been extensively studied, real-world scenarios often involve dynamic visual content where textual information appears across multiple frames. In videos, text serves as a crucial semantic carrier, complementing object and action cues with information that is otherwise visually inaccessible [7, 20, 35, 38, 58]. Benchmarking efforts in video text understanding are diverse and comprehensive. For example, MME-VideoOCR [42], VidText [52], FG Bench [9], and EgoTextVQA [59] provide

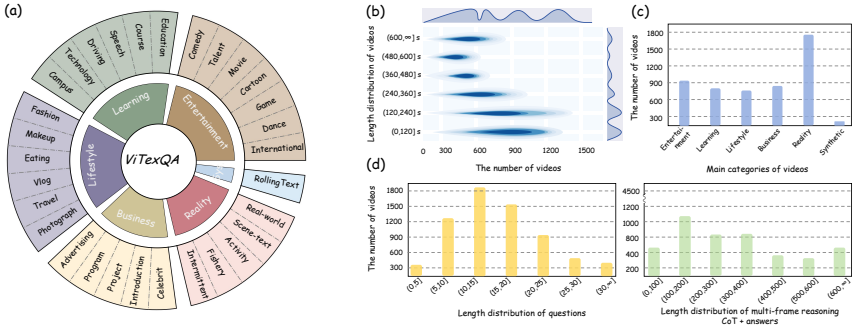

Fig. 2: Statistical overview of our ViTexQA dataset. **(a)** Video genre taxonomy with six major categories and their subcategories. **(b)** Distribution of video durations. **(c)** Video count distribution across six major categories. **(d)** Length distributions of questions **(left)** and CoT annotations with answers **(right)** measured in word count.

comprehensive evaluation suites that assess model performance from multiple perspectives, including text recognition, visual text question answering, and text grounding [36,38,47]. These benchmarks offer valuable insights into the strengths and weaknesses of existing approaches. In contrast, datasets specifically designed for video text question answering remain scarce. RoadTextVQA [45] focuses on driver assistance scenarios, where questions are answered based on textual information such as road signs present in driving videos. NewsVideoQA [19] introduces text-based question answering in news videos, requiring systems to analyze embedded textual content. M4-ViteVQA [58] comprises 7,620 video clips across nine scenes. However, only **13%** to **18%** of the questions in these datasets require the integration of information from multiple frames, which limits their capacity to evaluate temporal perception abilities. Recognizing the importance of multi-frame context for text-rich video question answering, we introduce ViTexQA, a dataset purpose-built for assessing MLLMs' temporal textual perception.

3 ViTexQA Dataset

The ViTexQA dataset is constructed through two complementary pipelines:
(1) Video Dataset Construction pipeline that ensures genuine multi-frame temporal perception requirements through rigorous data collection and quality-controlled annotation, producing 5,147 high-quality videos and 6,864 QA pairs with verified temporal dependencies;
(2) Multi-frame perception CoT Construction pipeline that further yields temporally-grounded Chain-of-Thought (CoT) annotations by preserving spatiotemporal text information across video frames to enrich the perception chain of provided QA pairs.

Together, these pipelines produce a comprehensive dataset (ViTexQA) with diverse video sources, enforced multi-frame dependencies, and temporally-grounded perception CoT annotations, as illustrated in Fig. 2. Compared to existing datasets or benchmarks, ViTexQA uniquely focuses on multi-frame temporal

Table 1: Datasets comparison for video text understanding.

Dataset	#Scenarios	#Videos	#QA	Designed for multi-frame perception?
Train Datasets				
M4-ViteVQA [58]	9	5,444	18,200	No
RoadTextVQA [45]	1	2,557	8,393	No
NewsVideoQA [19]	1	2,407	6,994	No
ViTexQA	30	4,472	6,124	Yes
Evaluation Benchmarks				
OCR Benchmark [10]	20+	25	1,477	No
MME-VideoOCR [42]	44	1,464	2,000	No
M4-ViteVQA [58]	9	680	2,103	No
RoadTextVQA [45]	1	329	1,052	No
EgoTextVQA [59]	2	1,507	7,064	No
NewsVideoQA [19]	1	346	964	No
VidText [52]	27	939	2,857	No
ViTexQA	30	675	740	Yes

perception, requiring models to integrate text information across multiple video frames rather than relying on single-frame understanding, a critical capability where current MLLMs exhibit significant weaknesses. As shown in Table 1, while prior benchmarks offer limited or no multi-frame perception requirements, ViTexQA ensures multi-frame dependency through rigorous verification protocols, providing both evaluation data and rich CoT training annotations (6,864 QA pairs from 5,147 videos) to address this fundamental gap in video text understanding [1, 55].

3.1 Video Dataset Construction

To construct a high-quality benchmark for genuine multi-frame temporal text perception, we establish a rigorous data curation and annotation framework encompassing diverse video sources, quality-controlled annotation protocols, and strategic synthetic augmentation.

(1) Data Collection and Curation. Videos are sourced from: (i) YouTube across 30 diverse domains (games, fashion, technology, etc.), selected for rich temporal text patterns [20, 24]; (ii) challenging cases from existing datasets with original annotations discarded; (iii) To enhance temporal text pattern coverage, we developed a Rolling-Text synthesis pipeline (Fig. 3(b)) that generates 100 synthetic scrolling text videos. All videos were standardized to 720p resolution to ensure text clarity and uniform quality. The dataset is strictly for research purposes and complies with data privacy policies.

(2) Quality-Controlled Annotation Pipeline. As shown in Fig. 3(a), real videos undergo iterative three-round annotation. (Round 1) Eight trained annotators independently create question-answer pairs with mandatory multi-frame perception requirements. Each video receives annotations from two different annotators to capture diverse perception perspectives. Annotators are explicitly instructed to verify that questions are unanswerable from any single frame alone, enforced through frame-by-frame answerability checks. (Round 2) Four senior

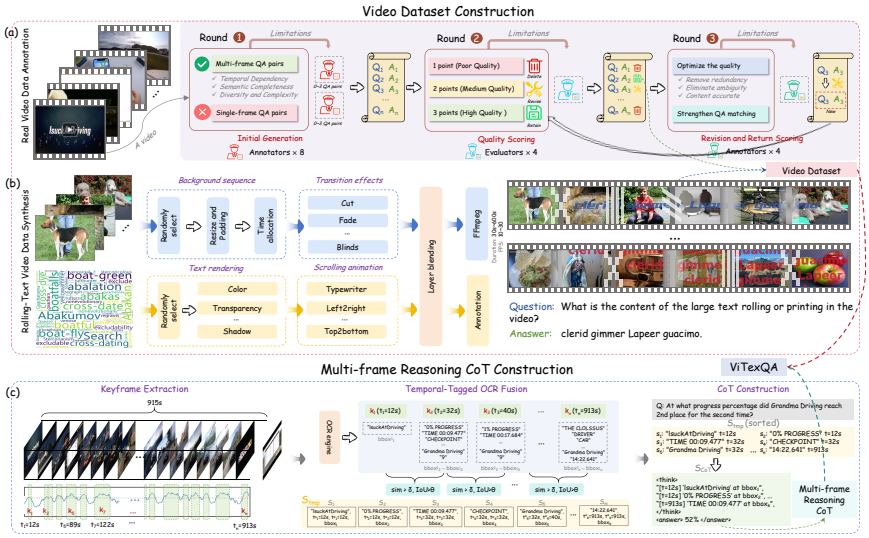

Fig. 3: Video Dataset Construction pipeline of our proposed ViTexQA. (a) Manual three-round annotation pipeline for real videos. (Round 1) Initial Generation. (Round 2) Quality Scoring. (Round 3) Revision and Return Scoring. (b) Synthetic scrolling text video pipeline. By randomly selecting background images and text corpus, setting random rendering parameters and scrolling methods, it generates 30s∼600s random videos with QA annotations. (c) Multi-frame perception CoT construction pipeline. The text information, timestamps, and location annotations of multi-frame are integrated into S_{cot} through three steps.

evaluators assess each annotation on three criteria—temporal dependency necessity (does the question require cross-frame integration?), answer correctness, and question clarity—assigning scores: 1 (reject: single-frame answerable or low quality), 2 (revision needed: minor issues in temporal grounding or clarity), 3 (accept: meets all quality standards). (Round 3) Score-2 annotations undergo targeted revision by four expert annotators, then return to Round 2 for re-evaluation. This iterative process continues until all samples achieve score-3, ensuring dataset-wide quality consistency.

3.2 Multi-frame Perception CoT Construction

Building upon the quality-controlled Video QA pairs described above, we propose a CoT data generation pipeline to enable multi-frame temporal perception. This pipeline preserves spatiotemporal text information across frames, providing detailed perception annotations for each question-answer pair in Fig. 3(c).

Keyframe extraction. Given a video sequence of duration T seconds, we aim to extract a sparse set of keyframes $\mathcal{K} = \{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$ to reduce computational overhead while preserving salient temporal information. We employ a hash-based detection [39] to identify frames with significant visual changes. Specifically, for

consecutive frames $\mathbf{f}_i$ and $\mathbf{f}_{i+1}$, we compute the perceptual hash difference:

$$\Delta H_i = \mathrm{d_H}(\mathrm{H}(\mathbf{f}_i), \mathrm{H}(\mathbf{f}_{i+1})) \tag{1}$$

where $\mathrm{d_H}$ denotes the Hamming distance metric and $\mathrm{H}(\cdot)$ denotes the perceptual hash function. Frames are selected as keyframes when ΔH_i exceeds a predefined threshold τ, ensuring that only frames with substantial visual content changes are retained. This approach effectively reduces redundancy while maintaining temporal coherence across the video.

Temporal-Tagged OCR Fusion. Conventional OCR fusion methods [58] typically perform simple text deduplication, discarding temporal information about "when" and "where" text appears. However, in multi-frame perception scenarios, the temporal order and spatial location of text are crucial for correct inference. We propose Temporal-Tagged OCR Fusion to explicitly preserve spatiotemporal metadata for each detected text instance.

For each keyframe $\mathbf{k}$ (omit the subscript for notation simplicity) of all extracted keyframe $\mathcal{K} = \{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$, we apply an OCR engine [8] to detect text regions. The j-th detected OCR text instance in keyframe $\mathbf{k}$ is defined:

$$\mathbf{o}_j = (\texttt{text}_j, t, \texttt{bbox}_j) \tag{2}$$

where $\texttt{text}_j$ is the recognized text, t is the corresponding timestamp, $\texttt{bbox}_j$ is the bounding box. The complete OCR detection set is $\mathcal{O}_{\mathbf{k}} = \{\mathbf{o}_j \mid \mathbf{k} \in \mathcal{K},\ j = 1, 2, \ldots, n\}$.

Then, we define two text instances $\mathbf{o}_p$ and $\mathbf{o}_q$ from different keyframes as content-equivalent if the following condition is met:

$$\mathrm{sim}(\texttt{text}_p, \texttt{text}_q) > \delta \quad \text{and} \quad \mathrm{IoU}(\texttt{bbox}_p, \texttt{bbox}_q) > \theta \tag{3}$$

where $\mathrm{sim}(\cdot,\cdot)$ measures string similarity and $\mathrm{IoU}(\cdot,\cdot)$ computes bounding box overlap. Based on this equivalence relation, all text instances in $\mathcal{O}_{\mathbf{k}}$ are grouped into M content-equivalent classes $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\}$, where each class $\mathcal{C}_m$ ($m = 1, 2, \ldots, M$) contains text instances that are mutually content-equivalent across multiple keyframes. For each class $\mathcal{C}_m$, we create a time text span:

$$\mathbf{s}_m = (\texttt{text}_m,\ t^{\mathrm{s}}_m,\ t^{\mathrm{e}}_m,\ \texttt{bbox}_m) \tag{4}$$

where $t^{\mathrm{s}}_m = \min_{\mathbf{o}\in\mathcal{C}_m} t(\mathbf{o})$ and $t^{\mathrm{e}}_m = \max_{\mathbf{o}\in\mathcal{C}_m} t(\mathbf{o})$ denote the first and last appearance times. Then, sort the data by time according to t^{s}_m to obtain the time corpus $\mathcal{S}_{\mathrm{tmp}} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M\}$. This representation retains both the temporal sequence and persistence of text, which are critical for multi-frame perception.

CoT Construction. Given a question Q with ground-truth answer A and the temporal corpus $\mathcal{S}_{\mathrm{tmp}}$ obtained from the previous fusion step, we construct perception CoT annotations that explicitly capture the multi-frame temporal perception process. Specifically, for each temporal text span $\mathbf{s}_m = (\texttt{text}_m, t^{\mathrm{s}}_m, t^{\mathrm{e}}_m, \texttt{bbox}_m)$ in $\mathcal{S}_{\mathrm{tmp}}$, we construct a structured perception trace that interleaves temporal information with textual content. The complete CoT sequence is formulated as:

$$\langle \texttt{<think>} \bigoplus_{x=1}^{N} \texttt{``[t=}t^{\mathrm{s}}_x\texttt{s] `text}^{x}_{\mathrm{gt}}\texttt{' at bbox}^{x}_{\mathrm{gt}}\texttt{''} \texttt{</think>} \rangle \tag{5}$$

where $\bigoplus$ denotes temporal concatenation ordered by appearance time t^{s}_x. $\texttt{text}^{x}_{\mathrm{gt}}$ and $\texttt{bbox}^{x}_{\mathrm{gt}}$ represent the recognized text and bounding box of the t^{s}_x timestamp, respectively.

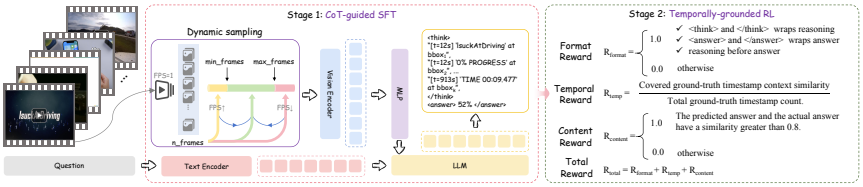

Fig. 4: Overall pipeline of FrameThinker. In stage 1, video frames sampled via dynamic sampling are fed into multi-frame perception CoT-guided SFT. In stage 2, following the standard GRPO approach, r_{fmt}, r_{tmp}, and r_{cnt} are designed to enable the model to be further optimized for multi-frame perception.

To ensure perception quality and temporal coherence, we then connect the temporal evidence to the final answer. The resulting temporally-grounded CoT annotations provide rich supervision signals that teach models not only what to predict, but how to integrate text information across multiple frames through explicit temporal perception. This representation fundamentally differs from existing single-frame text understanding approaches, as it requires models to track information flow across time and reason about temporal dependencies between text instances.

4 FrameThinker

Build upon ViTexQA, we further propose FrameThinker, a multi-modal training method for multi-frame video text understanding. As illustrated in Fig. 4, FrameThinker takes uniformly sampled video frames and temporal-aware CoT annotations as input, and employs a two-stage training paradigm: CoT-guided Supervised Fine-Tuning (SFT) for perception consistency, followed by Temporally-grounded Reinforcement Learning (RL) for task-specific optimization. By grounding predictions in explicit spatiotemporal textual evidence, FrameThinker achieves improved accuracy and interpretability for complex video understanding tasks.

4.1 CoT-guided SFT

In the first stage, we perform SFT on video-text pairs augmented with our generated temporal CoT annotations. The model input consists of two modalities: (1) visual information from uniformly sampled video frames $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\}$, and (2) textual information comprising the question Q and the temporal-aware CoT constructed from $\mathcal{S}_{\text{tmp}}$. Following standard practices in multimodal learning [27], visual features are first projected through a multilayer perceptron (MLP), then concatenated with text token embeddings, and finally fed into the language model backbone for autoregressive generation. Specifically, our training objective is to maximize the likelihood of generating the complete perception-answer sequences:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(Q,\mathcal{F},\text{S}_{\text{cot}})\sim\mathcal{D}} \left[\log \pi_\theta \left(\text{S}_{\text{cot}} \mid Q, \mathcal{F}\right)\right] \tag{6}$$

where $\mathrm{S}_{\mathrm{cot}} = \big\langle$<think> $\bigoplus_{x=1}^{N}$ ''[t=t_x^{s}s] 'text$_{\mathrm{gt}}^{x}$' at bbox$_{\mathrm{gt}}^{x}$'' </think> <answer> S_{gt} </answer>$\big\rangle$. θ denotes model parameters. This formulation encourages the model to explicitly articulate its temporal perception process by: (1) enumerating all detected text with spatiotemporal tags, (2) identifying relevant temporal patterns such as sequential order and co-occurrence, and (3) deriving the answer grounded in multi-frame evidence. Unlike conventional approaches that directly predict answers from visual features, CoT-guided SFT enforces intermediate perception steps that mirror human cognitive processes, thereby improving both accuracy and interpretability.

4.2 Temporally-grounded RL

In the second stage, we apply Temporally-grounded RL via GRPO [26] to further refine the model beyond the limitations of supervised imitation. While CoT-guided SFT provides a strong foundation by teaching the model to mimic human-annotated perception patterns, it is inherently constrained by the quality and coverage of training annotations. RL addresses these limitations by enabling the model to explore alternative perception paths and optimize multi-frame perception through reward-based feedback.

Specifically, we design three complementary reward functions targeting different aspects of output quality, each addressing a specific dimension of the desired model behavior:

Format Reward (r_{fmt}). The output format comprises two essential components: (i) final output format, typically encapsulated within <answer> </answer> tags, and (ii) the intermediate perception process format, typically enclosed within <think> </think> tags. Furthermore, the proper nesting order is enforced such that the perception section strictly precedes the answer section. Formally, the format reward is defined as:

$$r_{\mathrm{fmt}} = \begin{cases} 1, & \text{if format is correct,} \\ 0, & \text{if format is incorrect.} \end{cases} \tag{7}$$

Temporal Reward (r_{tmp}). The temporal grounding reward quantifies how well the model's perception chain references specific temporal evidence from the video, while ensuring that each referenced timestamp is semantically consistent with its surrounding context in the predicted CoT.

Assump the model output $\mathrm{P}_{\mathrm{cot}} = \big\langle$<think> $\bigoplus_{x=1}^{N}$ ''[t=t_x^{s}s] 'text$_{\mathrm{pre}}^{x}$' at bbox$_{\mathrm{pre}}^{x}$'' </think> <answer> $\mathrm{S}_{\mathrm{pre}}$ </answer>$\big\rangle$,where $\mathtt{text}_{\mathrm{pre}}^{x}$ and $\mathtt{bbox}_{\mathrm{pre}}^{x}$ represent the predicted text and predicted bounding box of the t_x^{s} timestamp, respectively. Then the temporal reward is defined as:

$$r_{\mathrm{tmp}} = \frac{1}{|\mathcal{T}_{\mathrm{gt}}|} \sum_{t \in \mathcal{T}_{\mathrm{gt}}} \mathbb{I}[t \in \mathcal{T}_{\mathrm{pre}}] \cdot \mathrm{sim}(\mathtt{text}_{\mathrm{pre}}, \mathtt{text}_{\mathrm{gt}}) \tag{8}$$

where $\mathcal{T}_{\mathrm{pre}}$ denotes the set of predicted timestamps in $\mathrm{P}_{\mathrm{cot}}$, $\mathcal{T}_{\mathrm{gt}}$ denotes the set of ground-truth timestamps in $\mathrm{S}_{\mathrm{cot}}$. $\mathtt{text}_{\mathrm{pre}}$ and $\mathtt{text}_{\mathrm{gt}}$ denote the textual context surrounding timestamp t in the predicted and ground-truth text, respectively.

Content Reward (r_{cnt}). The content reward evaluates the semantic correctness of the final answer:

$$r_{\text{cnt}} = \mathbb{I}[\text{sim}(\text{S}_{\text{pre}}, \text{S}_{\text{gt}}) > \gamma] \tag{9}$$

where S_{pre} is the predicted answer, S_{gt} is the ground-truth answer.

The threshold γ is set empirically (typically $\gamma = 0.8$) to balance strictness and robustness. This binary reward ($r_{\text{cnt}} \in \{0, 1\}$) provides a clear training signal for answer correctness while allowing reasonable flexibility in answer formulation.

Overall Reward and Optimization The overall reward is therefore defined as: $r = r_{\text{fmt}} + r_{\text{tmp}} + r_{\text{cnt}}$.

Consider the standard GRPO approach, it samples a group of generated output sequences $\{z_1, z_2, \ldots, z_G\}$ for each multimodal input query $(Q, \mathcal{F}) \in \mathcal{D}$ from the old policy model $\pi_{\theta_{\text{old}}}$. Then GRPO maximizes the following objective to optimize the model π_θ:

$$\mathcal{J}(\theta) = \mathbb{E}_{\{z_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|Q,\mathcal{F})} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Bigg(\min \Bigg(\frac{\pi_\theta(z_i \mid Q, \mathcal{F})}{\pi_{\theta_{\text{old}}}(z_i \mid Q, \mathcal{F})} \mathcal{A}_i, \\ \text{clip} \left(\frac{\pi_\theta(z_i \mid Q, \mathcal{F})}{\pi_{\theta_{\text{old}}}(z_i \mid Q, \mathcal{F})}, 1-\epsilon, 1+\epsilon \right) \mathcal{A}_i \Bigg) - \beta \, \mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \Bigg) \Bigg] \tag{10}$$

where ϵ and β are the PPO clipping hyper-parameter and the coefficient controlling the Kullback-Leibler (KL) penalty, respectively. π_{ref} is the reference model initialized from the SFT stage. To avoid notation conflict with the ground-truth answer A, we denote the advantage as $\mathcal{A}_i$. Specifically, for a group of $\mathcal{G}$ outputs $\{z_1, \ldots, z_G\}$ sampled from the same input $(Q, \mathcal{F})$, the advantage is calculated:

$$\mathcal{A}_i = \frac{r^{(i)} - \text{mean}(r^{(1)}, r^{(2)}, \ldots, r^{(G)})}{\text{std}(r^{(1)}, r^{(2)}, \ldots, r^{(G)})} \tag{11}$$

where $r^{(i)}$ represents the total reward evaluated on the i-th generated output z_i. Through this RL stage, the model is encouraged to generate not only structurally compliant CoT perception paths, but also temporally grounded and semantically accurate answers that faithfully interpret the complex multi-frame content.

5 Experiments

Implementation details. We use the latest Qwen3-VL [50] as the base model. All experiments were performed on eight NVIDIA A100 GPUs with 80GB of memory each. Additional details about dataset curation, training settings, and visualizations are provided in supplementary materials.

Evaluation metric. We chose ROUGE-L as the evaluation metric because its recall calculation method, based on the longest common subsequence, can effectively tolerate redundant prefixes, format symbols, and other irrelevant content in the model output.

5.1 Quantitative validation of multi-frame dependency

We verify multi-frame dependency in ViTexQA from two perspectives: ***Part I)*** Explicit judgment on query-video pairs. We classify a query-video pair as

multi-frame dependent or not, via annotators and prompted MLLMs. ***Part II)*** Implicit assessment via single/multi-frame performance gap. We evaluate four inputs: **Single-Rand** (a random video frame), **Single-AnsRand** (a random answer frame), **Single-Oracle** (the answer frame with the richest text), and **Multi-frame** (Full video). If models fail under single-frame input but succeed with multi-frame input, we regard query as multi-frame dependent.

Table 2: Quantitative validation of multi-frame dependency in ViTexQA.

Part I: Explicit Judgment				
Method	**The ratio of multi-frame dependent query-video pairs**			
Human	100.0			
Gemini-3 Pro	95.4			
Claude-Sonnet-4.6	97.7			
Part II: Implicit Assessment				
Method	**Single-Rand**	**Single-AnsRand**	**Single-Oracle**	**Multi-frame**
Gemini-3 Pro	0.0	25.2	27.3	71.6
Qwen3-VL-8B	0.0	25.6	26.5	66.7

As shown in Table 2, both parts consistently confirm the multi-frame dependency in ViTexQA: 1) Human annotators and MLLMs identify over 95% of questions as multi-frame dependent. 2) Multi-frame input yields a dramatically higher score, confirming that a single frame is insufficient to answer the question and highlighting the multi-frame nature of ViTexQA.

5.2 Effectiveness of Our ViTexQA Training Data

To validate the quality and generalizability of the ViTexQA training data, we fine-tune three representative models on the ViTexQA training data and evaluate them across four benchmarks. As shown in Table 3, fine-tuning on the ViTexQA training data can significantly improve the three models' performance on the ViTexQA benchmark (by 10.2%, 7.1%, and 10.0%, respectively), demonstrating the effectiveness of our curated training data for video text understanding tasks.

Table 3: Performance comparison before and after SFT on ViTexQA training data across multiple benchmarks. **VTD.**: ViTexQA training data.

Method	**VTD.**	**ViTexQA**	**MME-VideoOCR**	**Video-MME**	**TextVQA**
Qwen3-VL	✗	66.7	68.5	71.4	84.5
	✓	**76.9**(+10.2)	**74.3**(+5.8)	**72.6**(+1.2)	**84.6**(+0.1)
MiniCPM-V4.5	✗	70.1	77.5	68.8	82.4
	✓	**77.2**(+7.1)	**82.3**(+4.8)	**70.6**(+1.6)	**82.8**(+0.4)
FrameThinker	✗	73.5	78.3	74.1	84.2
	✓	**83.5**(+10.0)	**88.1**(+9.8)	**75.3**(+1.2)	**85.3**(+1.1)

* All models use CoT during inference, regardless of whether SFT is performed.

More importantly, models fine-tuned on ViTexQA also achieve improvements on out-of-domain benchmarks. These results confirm that ViTexQA serves not only as a challenging evaluation benchmark but also as a high-quality training resource that can broadly enhance a model's ability to integrate and reason over textual information distributed across video frames. Furthermore, the results from TextVQA show that it also plays a positive role in extracting text information from single images.

Table 4: Performance of existing MLLMs in zero-shot setting on the ViTexQA benchmark. **Avg.**:the average performance of the 30 scenarios; **Dri.**:Driving; **Adv.**:Advertising; **Edu.**:Education; **RT.**:RollingText. The highest score are highlighted, and the second highest is underlined.

Method	Size	Max_Frames	Resolution	Avg.	Dri.	Adv.	Edu.	RT.
Closed-source MLLMs								
GPT-4o [37]	-	128	512*512	51.6	55.2	41.4	51.9	46.6
Gemini-3 Pro [6]	-	128	512*512	71.6	**70.9**	70.3	58.9	59.2
Seed-1.6 [16]	-	128	512*512	61.1	63.4	60.2	54.9	56.8
Open-source MLLMs								
Qwen3-VL [50]	2B	768	448*448	57.2	56.9	56.9	69.1	52.4
LongVU [40]	3B	768	448*448	43.1	45.3	47.5	46.5	36.4
Qwen2.5-VL [2]	3B	768	448*448	55.4	58.1	53.3	63.2	49.4
Qwen3-VL [50]	4B	768	448*448	62.3	62.1	59.0	71.7	57.7
MiniCPM-V2.6 [54]	7B	64	448*448	50.7	55.6	51.7	57.0	38.2
Qwen2.5-VL [2]	7B	768	448*448	63.5	64.4	63.9	73.3	58.5
VideoLLaMA3 [56]	7B	64	336*336	52.4	54.5	62.1	58.5	49.6
LLaVA-Video [57]	7B	16	336*336	36.0	58.3	46.3	41.0	37.6
Keye-VL [51]	8B	768	336*336	39.9	54.9	47.1	49.1	35.3
MiniCPM-V4.5 [54]	8B	768	448*448	<u>70.1</u>	48.6	77.4	51.1	<u>60.1</u>
Qwen3-VL [50]	8B	768	448*448	66.7	63.7	61.7	<u>73.6</u>	53.4
InternVL3 [60]	8B	64	448*448	64.2	52.1	<u>85.5</u>	60.0	52.3
FrameThinker	8B	768	448*448	**73.5**	<u>67.4</u>	**87.2**	**75.9**	**62.3**

5.3 Effectiveness of Our FrameThinker on ViTexQA.

Zero-Shot Comparison with Existing MLLMs. To evaluate the effectiveness of FrameThinker, we first conduct a fair comparison with existing MLLMs under a zero-shot setting, where no ViTexQA training data is utilized by any of the compared methods. The corresponding results are presented in Table 4. FrameThinker, trained on the custom multi-frame perception CoT data from M4-ViteVQA, achieves the best overall performance with an average score of 73.5%, outperforming all competing baselines by a significant margin.

Table 4 shows that the open-source models have a clear technical advantage, significantly outperforming closed-source models. This superior performance is mainly due to the resource limitation of a maximum input of 128 frames, while the open-source model can utilize more frames, which reflects the ViTexQA dataset's emphasis on the core capabilities of model integration and inference of multi-frame content.

At the same time, all models perform worst on Rolling Text, revealing a fundamental weakness in current methods regarding the integration of dynamically changing textual information across frames. This represents the core evaluation challenge that ViTexQA is designed to assess.

Effectiveness of FrameThinker Trained on ViTexQA Data. To further evaluate the effectiveness of FrameThinker when trained on ViTexQA data, we present a comprehensive comparison in Table 5, contrasting the results before and after incorporating ViTexQA training data. As shown in the table, **No.(2)**, **No.(3)**, and **No.(4)** achieve improvements of 2.8%, 7.9%, and 10.0% over **No.(1)**, respectively.

5.4 Ablation studies

The necessity of multi-frame perception. To analyze how video resolution and frame number affect performance on ViTexQA, we perform two ablation studies. Fig. 5(a) fixes the maximum frame count to 768 for resolution evaluation: accuracy rises with resolution, yet gains become negligible beyond 448^2. Fig. 5(b) uses a fixed 448^2 resolution to test frame counts, showing consistent performance improvements with more input frames.

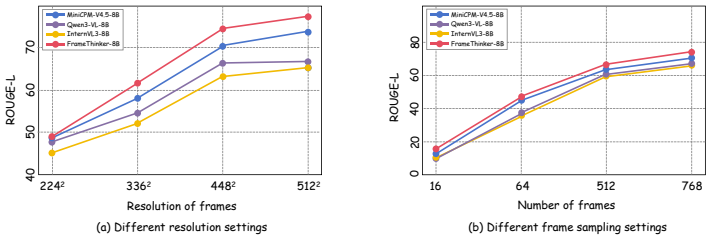

Fig. 5: The model's performance on ViTexQA is compared under different resolutions and frame sampling settings.

Ablation studies on FrameThinker. Table 5 presents a series of ablation experiments, demonstrating that FrameThinker's CoT-guided SFT and Temporally-grounded RL can improve the model's perception ability for multi-frame video text question answering tasks. As shown in the table, **No.(3)**, and **No.(4)** achieve improvements of 5.1% and 7.2%, over **No.(2)**, respectively.

Table 5: Performance of FrameThinker before and after using ViTexQA training data and ablation study on the FrameThinker.

No.	VTD.	CoT-SFT	RL	ViTexQA	MME-VideoOCR	Video-MME	TextVQA
(1)	✗	✗	✗	73.5	78.3	74.1	84.2
(2)	✓	✗	✗	76.3(+2.8)	80.5(+2.2)	74.4(+0.3)	84.2(+0.0)
(3)	✓	✓	✗	81.4(+7.9)	85.3(+7.0)	75.2(+1.1)	84.9(+0.7)
(4)	✓	✓	✓	**83.5**(+10.0)	**88.1**(+9.8)	**75.5**(+1.4)	**85.3**(+1.1)

6 Conclusion

In this paper, we introduce ViTexQA, a large-scale multi-frame temporal perception dataset for video text question answering, addressing the gap in temporal text perception datasets and benchmarks. Evaluations of mainstream MLLMs on ViTexQA reveal weak cross-frame text fusion and unsatisfactory overall performance. To address these challenges, we propose FrameThinker to boost multi-frame temporal perception via explicit key-frame text extraction and optimized perception modules, achieving clear performance gains. We believe that the dataset, multi-frame perception CoT construction strategy, and training framework introduced in this work will serve as valuable resources for the research community, stimulating further progress toward MLLMs that can genuinely comprehend the rich, temporally distributed textual information present in real-world videos.

References

1. Alampara, N., Schilling-Wilhelmi, M., Ríos-García, M., Mandal, I., Khetarpal, P., Grover, H.S., Krishnan, N.A., Jablonka, K.M.: Probing the limitations of multimodal language models for chemistry and materials research. Nature computational science **5**(10), 952–961 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
4. Chen, X., Zhao, Z., Chen, L., Zhang, D., Ji, J., Luo, A., Xiong, Y., Yu, K.: Websrc: A dataset for web-based structural reading comprehension. arXiv preprint arXiv:2101.09465 (2021)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Asaf: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), `https://arxiv.org/abs/2507.06261`
7. De Nadai, M., Damianou, A., Lalmas, M.: Describe what you see with multimodal large language models to enhance video recommendations. In: Proceedings of the Nineteenth ACM Conference on Recommender Systems. pp. 1159–1163 (2025)
8. Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., et al.: Pp-ocr: A practical ultra lightweight ocr system. arXiv preprint arXiv:2009.09941 (2020)
9. Fei, Y., Gao, Y., Xian, X., Zhang, X., Wu, T., Chen, W.: Do current video llms have strong ocr abilities? a preliminary study. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 9860–9876 (2025)
10. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., et al.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. arXiv preprint arXiv:2501.00321 (2024)
11. Guan, T., Lin, C., Shen, W., Yang, X.: Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In: European Conference on Computer Vision. pp. 130–147. Springer (2024)
12. Guan, T., Shen, W., Yang, X.: Ccdplus: Towards accurate character to character distillation for text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
13. Guan, T., Shen, W., Yang, X., Wang, X., Yang, X.: Bridging synthetic and real worlds for pre-training scene text detectors. In: European Conference on Computer Vision. pp. 428–446. Springer (2024)
14. Guan, T., Wang, Z., Fu, P., Guo, Z., Shen, W., Zhou, K., Yue, T., Duan, C., Sun, H., Jiang, Q., et al.: A token-level text image foundation model for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23210–23220 (2025)

15. Guan, T., Yang, Z., Wan, J., Yang, M., Guo, Z., Hu, Z., Luo, R., Chen, R., Jiang, S., Wang, P., et al.: Codepercept: Code-grounded visual stem perception for mllms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33542–33552 (2026)
16. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
17. Guo, Z., Ma, H.: Enhanced point cloud registration for workpieces using triangular constraint sampling consistency in complex industrial environment. The Visual Computer **42**(5), 221 (2026)
18. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5817–5834 (2025)
19. Jahagirdar, S., Mathew, M., Karatzas, D., Jawahar, C.: Watching the news: Towards videoqa models that can read. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4441–4450 (2023)
20. Lee, M.J., Gong, D., Cho, M.: Video summarization with large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18981–18991 (2025)
21. Li, B., Ge, Y., Chen, Y., Ge, Y., Zhang, R., Shan, Y.: Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. arXiv preprint arXiv:2404.16790 (2024)
22. Li, W., Fan, H., Wong, Y., Kankanhalli, M., Yang, Y.: Topa: Extending large language models for video understanding via text-only pre-alignment. Advances in Neural Information Processing Systems **37**, 5697–5738 (2024)
23. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multimodal models. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26763–26773 (2024)
24. Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: Llm-grounded video diffusion models. arXiv preprint arXiv:2309.17444 (2023)
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
26. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
28. Liu, M., Zhang, D., Liu, J., Cui, J., Xie, H., Chen, G., Ye, H., Yang, M.Y., Nex, F., Cheng, H.: Driveva: Video action models are zero-shot drivers. arXiv preprint arXiv:2604.04198 (2026)
29. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
30. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4122–4134 (2025)

31. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
32. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)
33. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
34. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019)
35. Nguyen, P.X., Wang, K., Belongie, S.: Video text detection and recognition: Dataset and benchmark. In: IEEE winter conference on applications of computer vision. pp. 776–783. IEEE (2014)
36. Nguyen, T., Anh, V.N.M., Pham, D.D., Vinh, T.Q., Quynh, N.D.T., Tien, L.A., Le, T.D., Nguyen, B.T.: Horus: multimodal large language models framework for video retrieval at vbs 2025. In: International Conference on Multimedia Modeling. pp. 286–293. Springer (2025)
37. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., and, I.B.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
38. Ozaki, S., Hayashi, K., Oba, M., Sakai, Y., Kamigaito, H., Watanabe, T.: Bqa: Body language question answering dataset for video large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 110–123 (2025)
39. Sadowski, C., Levin, G.: Simhash: Hash-based similarity detection. Tech. rep., Technical report, Google (2007)
40. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H.J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: Longvu: Spatiotemporal adaptive compression for long video-language understanding (2024), https://arxiv.org/abs/2410.17434
41. Shen, Y., Fu, C., Dong, S., Wang, X., Zhang, Y.F., Chen, P., Zhang, M., Cao, H., Li, K., Lin, S., et al.: Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. arXiv preprint arXiv:2502.05177 (2025)
42. Shi, Y., Wang, H., Xie, W., Zhang, H., Zhao, L., Zhang, Y.F., Li, X., Fu, C., Wen, Z., Liu, W., et al.: Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios. arXiv preprint arXiv:2505.21333 (2025)
43. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
44. Team, V., Liu, C., Kuo, C.W., Huang, C., Du, D., Chen, F., Chen, G., Zhang, H., Zhao, H., Zhang, L., et al.: Vidi2: Large multimodal models for video understanding and creation. arXiv preprint arXiv:2511.19529 (2025)
45. Tom, G., Mathew, M., Garcia-Bordils, S., Karatzas, D., Jawahar, C.: Reading between the lanes: Text videoqa on the road. In: International Conference on Document Analysis and Recognition. pp. 137–154. Springer (2023)
46. Wang, X., Liu, Y., Shen, C., Ng, C.C., Luo, C., Jin, L., Chan, C.S., Hengel, A.v.d., Wang, L.: On the general value of evidence, and bilingual scene-text visual question

answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10126–10135 (2020)
47. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European conference on computer vision. pp. 396–416. Springer (2024)
48. Wang, Z., Guan, T., Fu, P., Duan, C., Jiang, Q., Guo, Z., Guo, S., Luo, J., Shen, W., Yang, X.: Marten: Visual question answering with mask generation for multi-modal document understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14460–14471 (2025)
49. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM Multimedia (2017)
50. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., and, C.Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
51. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., Yang, F., Zhou, G.: Kwai keye-vl 1.5 technical report (2025), https://arxiv.org/abs/2509.01563
52. Yang, Z., Shu, Y., Yang, Z., Zhang, Y., Li, Y., Lu, K., Zeng, G., Liu, S., Zhou, Y., Sebe, N.: Vidtext: Towards comprehensive evaluation for video text understanding. arXiv preprint arXiv:2505.22810 (2025)
53. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Chen, C., Li, H., Zhao, W., et al.: Efficient gpt-4v level multimodal large language model for deployment on edge devices. Nature Communications **16**(1), 5509 (2025)
54. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, H., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., Sun, M.: Minicpm-v: A gpt-4v level mllm on your phone (2024), https://arxiv.org/abs/2408.01800
55. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
56. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., Zhao, D.: Videollama 3: Frontier multimodal foundation models for image and video understanding (2025), https://arxiv.org/abs/2501.13106
57. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data (2025), https://arxiv.org/abs/2410.02713
58. Zhao, M., Li, B., Wang, J., Li, W., Zhou, W., Zhang, L., Xuyang, S., Yu, Z., Yu, X., Li, G., et al.: Towards video text visual question answering: Benchmark and baseline. Advances in Neural Information Processing Systems **35**, 35549–35562 (2022)
59. Zhou, S., Xiao, J., Li, Q., Li, Y., Yang, X., Guo, D., Wang, M., Chua, T.S., Yao, A.: Egotextvqa: Towards egocentric scene-text aware video question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3363–3373 (2025)
60. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), https://arxiv.org/abs/2504.10479