VisAssist: A Visually Impaired-Captured Video Question Answering Benchmark for Assistive Systems

Qi Gao[1], Heng Li[1*], Yixin Zhou[1], Meixuan Zhou[1], Jieqiong Chen[2], Xinyu Chai[1]

[1]School of Biomedical Engineering, Shanghai Jiao Tong University, China
[2]Department of Ophthalmology, Shanghai General Hospital, China
{gao1976675161, liheng, yixinzhou, zy140017, xychai}@sjtu.edu.cn
jieqiong.chen@shgh.cn

Abstract

We present VisAssist, the first large-scale video question-answering dataset with 13,413 real-world videos captured by visually impaired users, addressing a critical gap in assistive vision research. Unlike existing benchmarks relying on third-person footage, VisAssist provides authentic first-person perspectives that uniquely capture challenges in blind photography—including unconventional framing, motion artifacts, and frequent information omission. Benchmark evaluations of SOTA multimodal models reveal systematic limitations: severe deficiencies in spatial reasoning when processing dynamic first-person viewpoints, an inability to distinguish missing information from poor capture quality leading to hazardous hallucinations, and fragile text understanding especially for non-Latin scripts under suboptimal conditions. This work establishes a vital real-world benchmark and underscores the need for specialized architectures in visual assistance systems.

Code — https://github.com/gaoCleo/VisAssist
Datasets —
https://huggingface.co/datasets/gaoCleo/VisAssist

Introduction

Computer vision technologies, especially multimodal large language models (MLLMs), can transform assistance for the visually impaired (Tseng et al. 2025). However, their performance is fundamentally limited by a mismatch between training data—typically images from sighted photographers—versus blind users' real-world input, which often has unconventional framing, focus errors, or missing details (Chen, Anjum, and Gurari 2023).

We present the **Visual Assistance Dataset (VisAssist)**, the first large-scale first-person video question-answering (Video QA) dataset for visually impaired users. It consists of 13,413 videos recorded by visually impaired volunteers across diverse indoor and outdoor scenarios, capturing daily situations where visual assistance is critically needed. The annotation framework combines open-ended natural language QA pairs with filming guidance suggestions. When a video's content is insufficient to answer a question, annotators not only provide answers but also offer concrete instructions to improve filming (e.g., Move the camera slightly left). This enables visually impaired users to capture key visual information more efficiently. VisAssist also includes meta-labels for required visual information and video quality, enabling analysis of model limitations—revealing when models succeed or hallucinate with blind-captured videos.

The VisAssist captures unique challenges: (1) high variance in frame utility (from single-frame answers to multi-frame dependencies), (2) extreme quality diversity (e.g., motion blur vs. recoverable low-light frames), and (3) Contextual understanding of text and space, where relationships between elements outweigh raw object detection. This benchmark fills a critical gap by aligning model training with real-world blind photography patterns and failure modes.

Based on this dataset, we evaluate SOTA video QA models under three settings: zero-shot learning, fine-tuning, and input modality analysis. Our tests expose critical weaknesses when these models process videos captured by visually impaired users:

- Incomplete Information Challenges: Current models fail to handle videos with missing key information caused by framing. A key limitation is their inability to distinguish absent information from poorly captured information, affecting answer accuracy.
- Spatiotemporal Understanding Limitations: All models exhibit poor performance on spatial reasoning tasks (e.g., depth estimation, directional queries), revealing fundamental challenges in processing dynamic first-person perspectives - especially with shaky footage typical of visually impaired users' recordings.
- Image Quality Sensitivity: Under common impairments (e.g., low light, glare, or motion blur), model performance declines. Some models generate confident but incorrect responses, posing risks through hazardous hallucinations in vision-assistance applications.

We further analyze the computational efficiency-performance trade-off by comparing four frame sampling strategies: (a) single middle-frame, (b) 0.5fps uniform, (c) 1fps uniform, and (d) adaptive keyframe selection via frame clustering. While higher frame rates generally

*Corresponding author

Question(ZH): 是什么牌子的？要煮多久有写吗？
Question(EN): What brand is it? Is the cooking time written there?
Answer(ZH): 牌子是意大利的Antonio Amato，烹饪方式在您拍摄的包装的第二面，画面模糊无法识别
Answer(EN): The brand is Italian, Antonio Amato. The cooking method is on the second face of the packaging you photographed, but the image is blurry and cannot be identified.
Instruction(ZH): 请调整角度，拍摄这个包装袋的背面，获取它的使用方法信息。
Instruction(EN): Please adjust the angle to shoot the back of the packaging bag to obtain the usage instructions.

Figure 1: Example samples from the constructed dataset.

improve accuracy at greater cost, stronger vision models like Gemini-Pro achieve comparable performance to 0.5fps using only selected keyframes. This reveals a critical dichotomy: vision-capable models thrive with sparse keyframes, even in noisy or low-information-density segments, while weaker models rely on raw frame quantity.

Main Contributions:

- Dataset Creation: We present VisAssist, the first video QA dataset captured by visually impaired users. This fills the gap in real-world assistive video QA benchmarks.
- Comprehensive Benchmarking: We evaluate multiple SOTA multimodal LLMs, exposing their performance bottlenecks.
- Frame Sampling Analysis: We benchmark input strategies for video QA.

Related Work

Video Question Answering Datasets

Video QA datasets (e.g., MSVD-QA, MSRVTT-QA) are vital for evaluating video understanding but often oversimplify visual content (e.g., single-entity recognition). Newer benchmarks target specific capabilities: TGIF-QA (temporal reasoning), Pororo-QA/TVQA (narrative understanding), ActivityNet-QA (action recognition), Social-IQ (social interactions) (Zadeh et al. 2019), and NExT-QA (causal-temporal reasoning) (Xiao et al. 2021). Most use third-person internet videos; Ego4D (first-person) (Grauman et al. 2022) is limited by non-textual annotations. We introduce VisAssist, the first first-person video dataset recorded by visually impaired users for assistive technology research. Its unique challenges—frame repetitions, motion blur, missing cues, and erratic viewpoints—address gaps in existing benchmarks. VisAssist enables tailored model development for visually impaired communities, where third-person datasets fall short.

Video Question Answering Models

VideoQA models fall into two categories: classification-based (Li et al. 2022; Pan et al. 2023) and generative. Early work, limited by computational resources, which select answers from a predefined set. However, their real-world applicability is constrained by their inability to generate novel answers. With improved computational power and LLMs, generative VideoQA has become mainstream. Video-ChatGPT, for example, uses a pretrained visual encoder and frozen language decoder, training only a bridging layer. Subsequent work enhances performance through multimodal fusion: VideoChat combines video descriptions with visual features, while Video-LLaMA integrates audio. For long-form video understanding, temporal modeling is key—Valley employs a Temporal Modeling Module, and TimeChat uses a Time-aware Frame Encoder for better reasoning.

Dataset

We present the VisAssist. Videos in the dataset are all recorded by visually impaired individuals.

Dataset Construction

Video source All videos are captured by visually impaired volunteers, including those who meet the legal blindness criteria and rely on screen readers to operate their phones. Each video must be 448×448 pixels in size and no longer than 15 seconds. Volunteers should verbally state a question or request, covering tasks like object search, navigation, or reading. Scenes must include both indoor and outdoor environments, prioritizing tasks requiring visual support (avoiding those solvable via touch/hearing alone).

Annotation The annotation process consists of three steps. First, each video segment is independently annotated by at least two annotators. Next, the annotation quality is verified using a consistency strategy. Finally, post-processing is performed to filter and compile the results.

Annotation Guidelines: (a) Describe only visual facts in the video. (b) Include potential hazards (e.g., obstacles). (c) For unanswerable questions, specify recording adjustments (e.g., Move camera left). (d) Flag privacy-related information for processing. For interactive annotation: Visually impaired volunteers send recordings for immediate feedback, enabling shooting adjustments. Dialogues are tracked by task ID, with multiple videos forming complete conversation sequences.

Consistency Strategy: An LLM determines whether two annotations conflict. If a conflict is detected, a third annotator arbitrates and modifies the conflicting entries according

to the guidelines.

Post-processing: First, videos are filtered to remove redundant clips and privacy-sensitive content. Next, an LLM translates (Liu et al. 2024) all original Chinese annotations into English. A rapid review is conducted to correct obvious translation errors and ensure quality. Finally, all videos undergo privacy protection processing through pixelation techniques to obscure any visible personal identifiable information. An example from the dataset are shown in Figure 1.

The annotations include these key elements:

- Question (text format)
- Answer (text format)
- Visual information required for answering (multi-label format): The types of visual information involved in answering the question (e.g., color, text, objects).
- Whether the video contains the answer (single-label format): Indicates if the answer is visible or if key information is missing (e.g., blurriness, occlusion).
- Shooting adjustment suggestions (text format): Recommendations for improving the video recording.
- Information type for shooting adjustment suggestions (single-label format): The category of suggested adjustments (e.g., stabilize camera, adjust angle).

Data Statistics

We perform a statistical analysis of video duration and annotation labels. The dataset comprises 13,413 videos totaling 137,554.64 seconds (5,465,939 frames), with average durations of 10.26 seconds and 407.5 frames per video. Most videos last 4-14 seconds, peaking at 14-15 seconds. Frame counts display a bimodal distribution, with primary clusters at 120-360 frames and 840-902 frames; fewer than 0.8% of videos contain under 60 frames (Figure 2d and 2e). Question lengths average 14.1 tokens (Chinese) and 15.8 tokens (English), while answers are longer: 23.5 (Chinese) and 28.7 (English) tokens. Shooting adjustments appear in 17.25% of videos, averaging 13.8 (Chinese) and 18.3 (English) tokens.

Objects and text provide the primary visual information needed, with color and location as secondary cues. Most tasks demand combined visual information (Figure 2a). Regarding answer completeness, 90% of videos contain visible answers (Figure 2b). Primary causes of incompleteness include: target objects being partially or fully out of frame, and insufficient capture of key details. Secondary issues involve improper shooting distance and motion blur (Figure 2c).

Dataset Comparison

Existing Video QA datasets vary significantly in their characteristics depending on video sources and annotations. As shown in Table 1, mainstream datasets primarily consist of third-person perspective videos from movies, social media, and TV programs. EgoVQA and EgoTextVQA, like VisAssist, are first-person perspective datasets but differ significantly. In video content, they consist of videos shot by sighted individuals engaging in certain activities, with each video containing at least one event and most key information being clear. VisAssist, in contrast, features videos captured by visually impaired individuals, where the scene attributes outweigh the event attributes, and key information may be missing. For question types, EgoVQA and EgoTextVQA focus more on reasoning based on events, while VisAssist emphasizes objective visual information.

VisAssist captures the unique characteristics of videos filmed from their perspective and reflects their real-world needs. The key features of the dataset are analyzed below:

Significant variation in effective frames: Some videos require only a single clear frame to answer a question(others being redundant/noisy). In contrast, others demand the integration of information across multiple frames. This challenges models to capture key information efficiently: frame selection is critical, as redundant frames raise computational costs, and missing key frames risks information loss.

Substantial video quality variations: Hardware, environments, and personnel differences cause significant video quality disparities. Some videos are clear; others show motion blur, low-light blur, glare, or incomplete framing. Critically, while some blurry videos still contain discernible information (albeit more challenging to extract), others are completely missing key visual details. This challenges models twofold: low-quality frames need stronger associative reasoning, while missing information demands hallucination suppression.

Critical textual information: Text plays a key role in many videos, but the core challenge lies not in mere text recognition but in detecting the relationship between text and other visual elements. For example, a videographer may focus more on the text on a button pressed by a finger than on other incidental text in the scene.

Crucial spatial information: Spatial positioning is vital for tasks like object search or obstacle avoidance. This requires detecting target elements' positions relative to the observer and understanding spatial relationships among multiple elements. Some tasks also involve depth/distance estimation, adding model challenges.

Experiment

The study benchmarks mainstream VideoQA models using the VisAssist with open-ended annotations, focusing specifically on generative VideoQA models for evaluation.

Experimental Setups: We evaluate the zero-shot performance. We assess multiple closed-source models (including ChatGPT-4o (Hurst et al. 2024), Gemini-2.5-Pro (Comanici et al. 2025), and Gemini-2.5-Flash (Comanici et al. 2025)), as well as open-source models (such as TimeChat (Ren et al. 2024), VideoChat2 (Li et al. 2024)'s Vicuna and Mistral variants, VideoChatGPT (Maazi et al. 2024), VideoLLaMA2 (Cheng et al. 2024), Qwen2-VL-OCR(a fine-tuned version of Qwen2-VL (Wang et al. 2024) on OCR datasets), and Qwen-2.5VL (Bai et al. 2025)). Qwen2-VL-OCR has only 2B parameters, while the other open-source models all have 7B parameters. Prior research indicates that these models achieve SOTA performance on existing video QA benchmarks. Additionally, we evaluate the models' performance after fine-tuning to assess the dataset's difficulty. For fine-

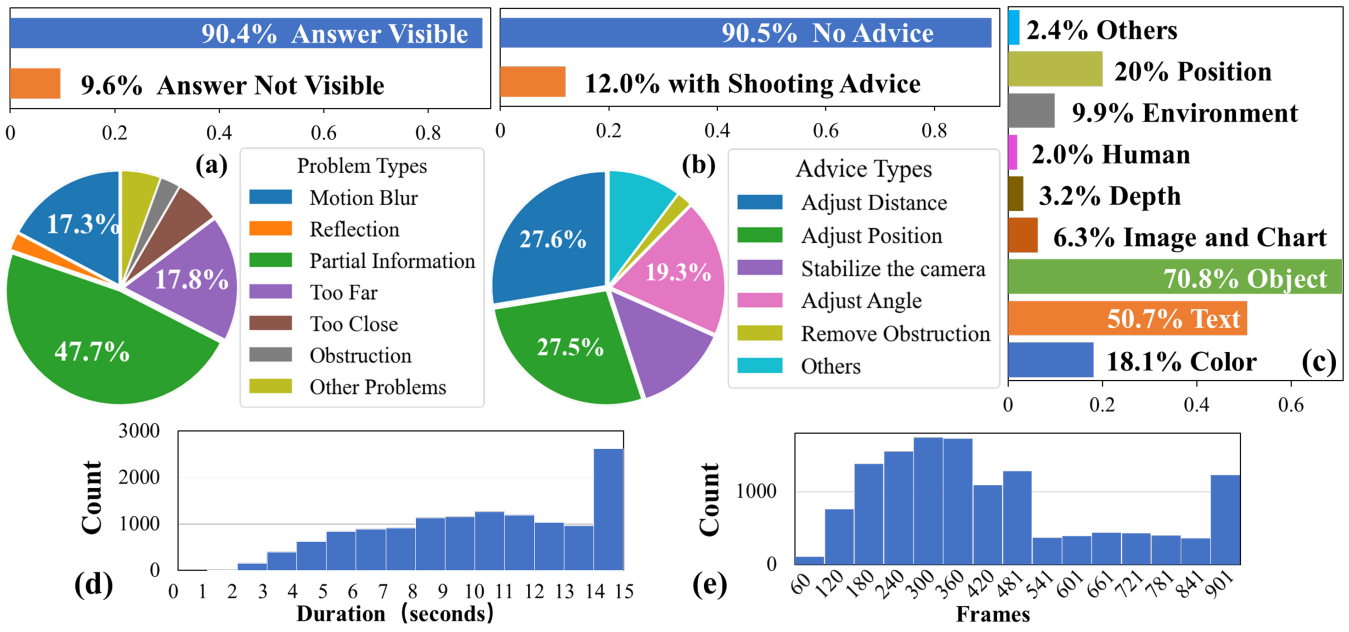


Figure 2: Data Statistics. (a) Presence of answers in videos. (b) Types of shooting adjustment suggestions. (c) Visual information required to answer questions. (d)Duration distribution. (e) Frame count distribution

Datasets	Video source	Clips	Format	Perspective	Content
MovieQA (Tapaswi et al. 2016)	Movie	6771	MC	Third Person	Long Story
MSVD-QA (Xu et al. 2017)	MSVD	1970	OE	Third Person	Records
MSRVTT-QA (Xu et al. 2017)	MSRVTT	10000	OE	Third Person	Records
TGIF-QA (Jang et al. 2017)	Social media	71741	MC&OE	Third Person	GIF
Pororo-QA (Kim et al. 2016)	Cartoon	16066	MC	Third Person	Long Story
TVQA (Lei et al. 2018)	TV show	21800	MC	Third Person	Long Story
ActivityNet-QA (Yu et al. 2019)	ActivityNet	5800	OE	Third Person	Activities
Social-IQ (Zadeh et al. 2019)	YouTube	1250	MC	Third Person	Social Support
EgoVQA (Fan 2019)	IU Multiview dataset	520	MC	First Person	Assistance
NExT-QA (Xiao et al. 2021)	YFCC-100M	5440	MC&OE	Third Person	Records
Valley (Wu et al. 2025)	WebVid2M	73000	OE	Third Person	Records
Video-ChatGPT (Maazi et al. 2024)	-	100000	OE	Third Person	Records
TimeIT (Ren et al. 2024)	Multiple datasets	125000	OE	Third Person	Records
FunQA (Xie et al. 2024)	YouTube	4365	MC&OE	Third Person	Records
EgoTextVQA (Zhou et al. 2025)	EgoSchema	1507	OE	First Person	Activities
VisAssist (Ours)	VIP volunteer	13413	OE	First Person	Assistance

Table 1: Video QA Dataset Comparison (VIP: Visually Impaired Persons; MC: multiple-choice; OE: open-ended)

tuning, we select the open-source TimeChat, VideoChat2's Vicuna variant, VideoLLaMA2, and Qwen-2.5VL.

Zero-shot testing uses raw answer annotations as ground truth, while fine-tuning incorporates supplementary shooting adjustment suggestions from the annotations. The goal is to enable the models not only to respond accurately but also to provide framing guidance when key information is absent from the video. The dataset splits into 8,037 training and 5,376 test videos (6:4 ratio), with all models undergoing 3-epoch fine-tuning across four NVIDIA A100 GPUs.

Metrics: The evaluation employs semantic metrics based on LLMs (Maazi et al. 2024), adapting COR (Correctness of Information) and DO (Detail Orientation) from existing frameworks (Maazi et al. 2024), while introducing SU (Spatial Understanding) to align with the spatial emphasis of the VisAssist. Model performance is assessed using COR, DO, and SU, each scored on a 0-5 scale (5 being optimal).

Zero-shot Evaluation for Generalization Capacity

Table 2 presents the results for Chinese versions. Most models demonstrate optimal performance on color-related questions, likely because color represents the most fundamental and distinguishable visual feature. Among open-source models (excluding Qwen), text-based questions prove most

Model	Color	Text	Object	Icon	Depth	Human	Env	Dir	COR	DO	SU	Avg
TimeChat	1.11	0.93	1.06	1.09	1.08	1.24	**1.28**	0.97	1.21	0.99	0.93	1.04
VideoChat2(vicuna)	1.73	1.02	1.34	1.32	1.26	**1.79**	1.53	1.22	1.34	1.24	1.27	1.28
VideoChat2(mistral)	**1.79**	0.93	1.22	1.27	0.82	1.29	1.14	1.00	1.27	1.07	1.15	1.16
VideoChatGPT	**1.46**	0.94	1.18	1.15	1.25	1.35	1.23	1.17	1.21	1.07	1.11	1.13
VideoLLaMA2	**2.11**	1.47	1.72	1.74	1.36	1.52	1.68	1.56	1.87	1.55	1.59	1.67
Qwen2-VL-OCR	**2.55**	2.15	2.18	2.22	1.77	2.26	2.03	1.93	2.31	2.07	2.18	2.19
Qwen2.5-VL	**2.58**	2.36	2.37	2.44	1.88	2.05	2.20	2.00	**2.54**	**2.29**	**2.29**	**2.37**
Gemini-flash	**3.13**	2.88	2.85	2.96	2.18	2.68	2.74	2.42	3.04	2.73	2.80	2.86
Gemini-pro	**3.45**	3.37	3.29	3.35	2.54	3.23	3.13	2.88	**3.47**	**3.14**	**3.30**	**3.30**
ChatGPT-4o	**3.13**	2.54	2.81	2.85	2.61	3.04	2.78	2.46	2.93	2.68	2.67	2.76

Table 2: Zero-shot performance of models on the dataset, using answer as ground truth. Avg denotes the mean of COR, DO, and SU metrics. The left column indicates the category of visual information required for answering; within each model, the strongest type is **bold**, while the weakest is underlined. *Dir* is the abbreviation for Direction; *Env* is the abbreviation for Environment.

Model	Visible	Vague	Reflect	Part Lost	Far	Close	Covered
TimeChat	1.03	1.07	**1.27**	1.09	1.14	1.10	1.13
VideoChat2(vicuna)	**1.31**	1.06	0.63	0.91	1.15	1.20	1.00
VideoChat2(mistral)	**1.19**	0.71	0.60	0.88	0.99	0.84	1.13
VideoChatGPT	1.15	0.93	0.73	0.96	**1.19**	0.80	1.36
VideoLLaMA2	1.68	1.48	1.00	1.56	1.56	1.62	**1.82**
Qwen2-VL-OCR	**2.24**	1.67	1.27	1.64	1.80	1.83	1.79
Qwen2.5-VL	2.37	2.50	2.00	2.36	2.24	2.79	**2.54**
Gemini-flash	**2.89**	2.53	2.77	2.47	2.39	2.81	2.69
Gemini-pro	**3.36**	2.90	2.83	2.72	2.48	2.73	2.95
ChatGPT-4o	2.78	2.62	2.47	2.47	2.36	**2.92**	2.74

Table 3: Zero-shot performance of models across video quality categories. Each model's top-performing category is **bold**, while the weakest is underlined.

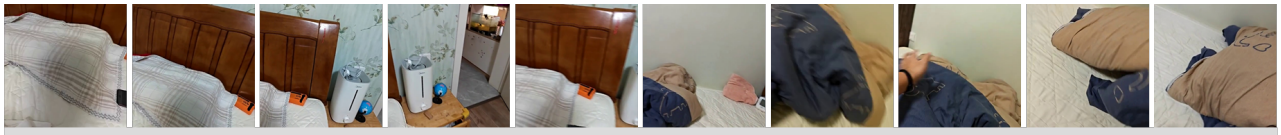
challenging. This difficulty may arise because the dataset contains primarily Chinese text that is often blurred due to recording conditions, combined with the inherently higher difficulty of text questions compared to conventional datasets and limited exposure during training.

All models perform poorly on depth-related questions, indicating significant limitations in estimating 3D spatial information. Directional understanding also remains weak across models. These limitations may stem from the dataset's unique characteristics: first-person videos recorded by visually impaired individuals require models to simultaneously infer spatial relationships, demanding exceptional scene comprehension capabilities.

Table 3 shows model performance on videos with varying degradation levels. Except for *Visible* cases, all categories involve unrecognizable critical visual information. A consistent weakness is handling reflective surfaces: though non-essential regions often remain clear, models tend to generate plausible but incorrect interpretations of reflective surfaces by over-relying on contextual cues, a behavior consistent with VLMs' instruction-following bias where models prioritize producing coherent responses over admitting uncertainty (Wang and Nalisnick 2025). Similarly, models show better adaptation to *Close* than *Far* scenarios. This aligns with the typical failure pattern of VLMs: *Close* cases, while severely blurred, provide limited but concentrated visual signals that constrain hallucination space, whereas *Far* scenarios with complex backgrounds offer more opportunities for erroneous contextual reasoning.

ZH Version				
Model	COR	DO	SU	avg
TimeChat	1.285	1.382	1.202	1.290
VideoChat2	1.225	1.245	1.163	1.211
VideoLLaMA	1.873	1.625	1.723	1.741
Qwen2.5-VL	1.944	1.798	1.852	1.865
EN Version				
Model	COR	DO	SU	avg
TimeChat	1.460	1.474	1.361	1.432
VideoChat2	1.378	1.414	1.304	1.365
VideoLLaMA	1.874	1.642	1.721	1.746
Qwen2.5-VL	1.825	1.694	1.761	1.760

Table 4: Cross-lingual evaluation results. Models are fine-tuned and evaluated using both answer annotations and shooting suggestions as ground truth. VideoChat is the vicuna version.



Question: Where is my charger?

GT: Your charger is on top of the humidifier on the nightstand

TimeChat: In the floor. (Score:0.0)

VideoChat2: Your charger is located on the bed. (Score:0.6)

VideoChatGPT: The charger is on the bed. (Score:0.6)

VideoLLaMA2: Your charger can be found on the bedside table. (Score:3.0)

Qwen2-VL-OCR (2B): It looks like you're in a bedroom. You're lying on a bed with a blanket and a pillow. There's a clock on the bed, and a phone on the nightstand. (Score:0.3)

Qwen2.5-VL (7B): It looks like your charger might be under the bed or in one of the drawers. The video shows a bed with a brown headboard and some items on it, but there's no clear view of a drawer or a place where a charger would typically be stored. If you're not sure where it is, you might want to check under the bed or in any other places where you keep your personal items. (Score:0.3)

Gemini-flash: The charging cables are on top of the Midea humidifier, which is on the nightstand next to the bed. (Score:4.3)

Gemini-pro: At the very beginning of the video (around 0:01-0:02), a white charging cable and what looks like a white charging brick are visible on top of the white Midea humidifier. The humidifier is on a wooden nightstand next to the bed. (Score:5.0)

ChatGPT-4o: The charger is located on the piece of furniture to the left. The white charging cable is plugged into the outlet, and the cable is visible coiled on top of a container. (Score:2.0)

Figure 3: Qualitative zero-shot results across models. Score represents the average of COR, DO, and SU metrics.

Fine-tuning for Domain Adaptation

During fine-tuning, the model is trained on concatenated text sequences where answers are directly followed by shooting adjustment suggestions, with the entire sequence serving as ground truth. Thus, evaluation requires the model to not only answer questions but also recommend shooting improvements. As shown in Table 4, TimeChat and VideoChat2 perform poorly. Although TimeChat leverages timestamps for long-video understanding, it fails to adapt to VisAssist's short, visually-oriented videos. VideoChat2 aligns modalities through multi-stage training but struggles with visual encoding, a critical requirement for VisAssist. By contrast, VideoLLaMA2 and the Qwen series show slightly better results. VideoLLaMA2's dual-branch design processes visual and auditory inputs jointly, while Qwen's 2D RoPE enhances spatial modeling. Despite these advantages, their accuracy remains low, revealing fundamental gaps in multimodal models for complex video tasks.

Impact of Visual Input Configuration

In the context of assisting the visually impaired, the model's response time is also a critical factor affecting its effectiveness. The number of frames and resolution are key factors influencing the model's latency and performance.

Frame Rate

The VisAssist's unique composition necessitates a evaluation of temporal sampling strategies. Unlike conventional video QA datasets where uniform high frame rates often suffice, our dataset contains three semantically distinct question types with divergent temporal demands: (1) single-frame questions (e.g., color identification) that require minimal visual sampling, (2) multi-frame relational reasoning tasks demanding sequential analysis, and (3) transient-keyframe questions (e.g., reading expiration dates) where critical information appears sporadically due to real-world filming instability. This tripartite structure creates an inherent tension between computational efficiency and information completeness.

To quantitatively characterize this trade-off, we implement four representative sampling strategies. The middle-frame selection serves as a baseline to establish the lower bound of temporal information requirements. Uniform sampling at 0.5 fps and 1 fps provides controlled points for evaluating the cost-accuracy curve under standardized conditions. Crucially, we implement a heuristic frame selection module (FSM) to study practical sampling strategies. This lightweight component operates without MLLM retraining.

As shown in Table 5, the four sampling strategies show a trade-off between accuracy and efficiency. Higher frame-rate sampling yields the best performance but is computationally expensive, while single-frame sampling is efficient but less accurate—underscoring the value of temporal information for VisAssist. Furthermore, the FSM achieves accuracy comparable to 0.5 fps sampling in most scenarios while significantly reducing computational overhead, with its TeraFLOPs performance substantially outperforming high-frame-rate sampling approaches.

The model's visual comprehension capability significantly impacts frame selection effectiveness. For instance, Gemini Pro with FSM slightly surpasses 0.5FPS uniform sampling, indicating stronger models utilize keyframe information more effectively. Conversely, weaker models like Qwen2-VL-OCR show marginally worse performance with FSM versus 0.5FPS sampling. This likely occurs because

Model	Method	COR	DO	SU	AVG	Latency(ms)	TFLOPs
VideoLLaMA2	1Frame	1.71	1.48	1.44	1.55	3525.77	-
	0.5FPS	1.77	1.47	1.55	1.60	3765.45	-
	1FPS	1.87	1.55	1.59	1.67	5258.08	-
	FSM	1.74	1.49	1.52	1.58	3615.23	-
Qwen2-VL-OCR	1Frame	2.17	1.89	2.02	2.03	1668.74	1.83
	0.5FPS	2.18	1.93	2.04	2.05	1890.79	6.32
	1FPS	2.31	2.07	2.18	2.19	2153.11	9.46
	FSM	2.10	1.94	1.96	2.00	1711.09	2.68
Qwen2.5-VL	1Frame	2.39	2.12	2.12	2.21	3559.41	4.97
	0.5FPS	2.33	2.12	2.11	2.19	4130.82	8.38
	1FPS	2.54	2.29	2.29	2.37	4590.12	12.64
	FSM	2.36	2.10	2.10	2.19	3663.22	6.17
Gemini pro	1Frame	3.05	2.76	2.77	2.86	-	-
	0.5FPS	3.32	3.03	3.05	3.13	-	-
	1FPS	3.47	3.14	3.30	3.30	-	-
	FSM	3.34	3.03	3.06	3.15	-	-

Table 5: Impact of input frame rates on model performance. The 1 frame baseline uses the temporally central frame of the video. FSM selects no more than 3 frames (averaging 2 frames). Latency is measured on a single NVIDIA 3090 GPU.

Resolution	Model	COR	DO	SU	AVG	Latency(ms)	TFLOPs
224x224	VideoLLaMA2	1.66	1.47	1.41	1.51	4931.34	-
	Qwen2-VL-OCR	2.17	1.93	2.06	2.05	1860.93	8.28
	Qwen2.5-VL	2.51	2.23	2.28	2.34	4446.48	10.92
	Gemini pro	3.37	3.04	3.19	3.20	-	-
448x448	VideoLLaMA2	1.87	1.55	1.59	1.67	5258.08	-
	Qwen2-VL-OCR	2.31	2.07	2.18	2.19	2153.11	9.46
	Qwen2.5-VL	2.54	2.29	2.29	2.37	4590.12	12.64
	Gemini pro	3.47	3.14	3.30	3.30	-	-

Table 6: Impact of Resolution on Model Performance

frame selection errors are amplified in low-capacity models, and keyframe information gains cannot compensate for limited visual understanding. Moreover, in complex temporal reasoning tasks (e.g., dynamic scene understanding), frame selection may hinder scene comprehension, resulting in performance degradation versus 1FPS sampling.

Resolution Beyond frame rate, input resolution is another critical factor influencing video QA performance. Higher resolutions preserve finer details, particularly important for VisAssist's fine-grained questions, but increase computational costs. Our experiments compare 224×224 and 448×448 resolutions. Table 6 shows 448×448 consistently improves accuracy across all models. Notably, gains correlate with models' visual capabilities: Gemini Pro shows the largest improvement, while VideoLLaMA2 has minimal gains, revealing model-dependent sensitivity.

Limitation

Due to geographical limitations in data collection, Chinese characters appear most frequently in video scenes involving text, resulting in a language bias. To mitigate this, we will incorporate multilingual data via international collaborations in future work. Additionally, while this study focuses on video frame selection, network design and temporal modeling also critically impact performance. Thus, a systematic study of these interdependent factors should be prioritized in follow-up research for more robust outcomes. Additionally, the multi-turn continuous dialogue task not considered in the study may be a direction worth exploring.

Conclusion

We introduces VisAssist, the first large-scale video QA dataset captured by visually impaired users, authentically reflecting blind photography challenges through first-person perspective. Results show SOTA models perform well on color recognition but struggle with spatial reasoning and text understanding, especially in challenging conditions. These results emphasize the need for specialized architectures in visual assistance systems.

Acknowledgments

This study was funded by the Natural Science Foundation of Shanghai (25ZR1401181), National Natural Science Foundation of China (62103269, 62073221), Med-X Research Fund of Shanghai Jiao Tong University (YG2022QN077), and New Faculty Initiation Program of Shanghai Jiao Tong University (23X010501996). We sincerely thank Libo

Zhang, Wa Gao, Liang Xie, Yanqin Wu, Xiying Zhao, and Qiyu Jiang for their assistance in volunteer coordination, dataset collection, and organization. We also extend our gratitude to Jing Zhou from Shanghai Jiao Tong University's Network Information Center for her support in model testing.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, C.; Anjum, S.; and Gurari, D. 2023. Vqa therapy: Exploring answer differences by visually grounding answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15315–15325.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *CoRR*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fan, C. 2019. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18995–19012.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2758–2766.

Kim, K.; Nan, C.; Heo, M.; Choi, S.; and Zhang, B. 2016. PororoQA: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, volume 15.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369. Association for Computational Linguistics.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Y.; Wang, X.; Xiao, J.; Ji, W.; and Chua, T.-S. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2928–2937.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Maazi, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *62nd Annual Meeting of the Association-for-Computational-Linguistics (ACL)/Student Research Workshop (SRW), Bangkok, THAILAND, aug 11-16, 2024*, 12585–12602. ASSOC COMPUTATIONAL LINGUISTICS-ACL.

Pan, J.; Lin, Z.; Ge, Y.; Zhu, X.; Zhang, R.; Wang, Y.; Qiao, Y.; and Li, H. 2023. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 272–283.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4631–4640.

Tseng, Y.-Y.; Sharma, T.; Zhang, L.; Stangl, A.; Findlater, L.; Wang, Y.; and Gurari, D. 2025. BIV-Priv-Seg: Locating Private Content in Images Taken by People With Visual Impairments. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 430–440. IEEE Computer Society.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; and Nalisnick, E. 2025. Are vision language models robust to uncertain inputs? *arXiv preprint arXiv:2505.11804*.

Wu, Z.; Chen, Z.; Luo, R.; Zhang, C.; Gao, Y.; He, Z.; Wang, X.; Lin, H.; and Qiu, M. 2025. Valley2: Exploring Multimodal Models with Scalable Vision-Language Design. *arXiv preprint arXiv:2501.05901*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xie, B.; Zhang, S.; Zhou, Z.; Li, B.; Zhang, Y.; Hessel, J.; Yang, J.; and Liu, Z. 2024. FunQA: Towards Surprising Video Comprehension. In *European Conference on Computer Vision (ECCV)*.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually

refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9127–9134.

Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8807–8817.

Zhou, S.; Xiao, J.; Li, Q.; Li, Y.; Yang, X.; Guo, D.; Wang, M.; Chua, T.-S.; and Yao, A. 2025. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3363–3373.