

ClearText-Video: A Large-Scale Text-Centric Video Dataset Bridging Video Restoration and Scene-Text Enhancement

Jinlong Li^{1*}, Jiaming Ding^{1*}, Dingfu Lu^{1,2‡}, Malcolm Hsiu^{1,3‡}, Chuang Ke¹, Kangning Yang¹, Bochen Guan¹, Lan Fu¹, Jie Cai¹, Huiming Sun¹, and Zibo Meng¹

¹ OPPO US AI Center, USA

² University of Wisconsin–Madison, USA ³University of California San Diego, USA
{jinlong.li1,jiaming.ding,chuang.ke3,kangning.yang,bochen.guan,lan.fu,
jie.cai,huiming.sun2,zibo.meng}@oppo.com

Abstract. Multimodal Large Language Models (MLLMs) have recently demonstrated strong progress in visual–linguistic understanding, yet their performance on text-centric video reasoning remains highly sensitive to input quality. Real-world user-provided videos frequently contain motion blur, compression artifacts, noise, and low-resolution text, substantially impairing reliable text reading and downstream reasoning. Whether MLLMs can robustly read and reason over in-the-wild text under diverse quality conditions remains an unanswered fundamental question. We introduce ClearText-Video (CTVid), a large-scale, scene-text-aware benchmark for studying text-centric video understanding under controlled quality variation. CTVid contains 4,639 real-world text-rich egocentric videos, 550K+ frames, 1.6M human-verified scene-text annotations, and 220K+ spatial/temporal question–answer pairs in Chinese and English. For each high-quality video, CTVid provides content-matched Degraded-Quality and Restored-Quality variants, supporting two task families: Text-Centric Video Restoration and Multi-Video Quality VideoQA. We evaluate 18 representative restoration methods and 16 state-of-the-art MLLMs on CTVid. The results show that visual enhancement does not guarantee textual fidelity or downstream reasoning gains: blur is more damaging than low resolution, restored videos can alter the textual evidence used by MLLMs, and OCR-only pipelines remain far below direct multimodal reasoning. CTVid exposes the gap between video restoration and text-grounded understanding, providing a rigorous foundation for restoration-aware, quality-robust text-centric video systems. Benchmark: <https://github.com/jinlong17/CTVid-Bench>.

1 Introduction

Recent Multimodal Large Language Models (MLLMs) [63,29,15,75,12] have made rapid progress, enabling joint visual–linguistic reasoning and the integrated

* Equal contribution. † Corresponding author.

‡ Work done during internships at OPPO US AI Center.

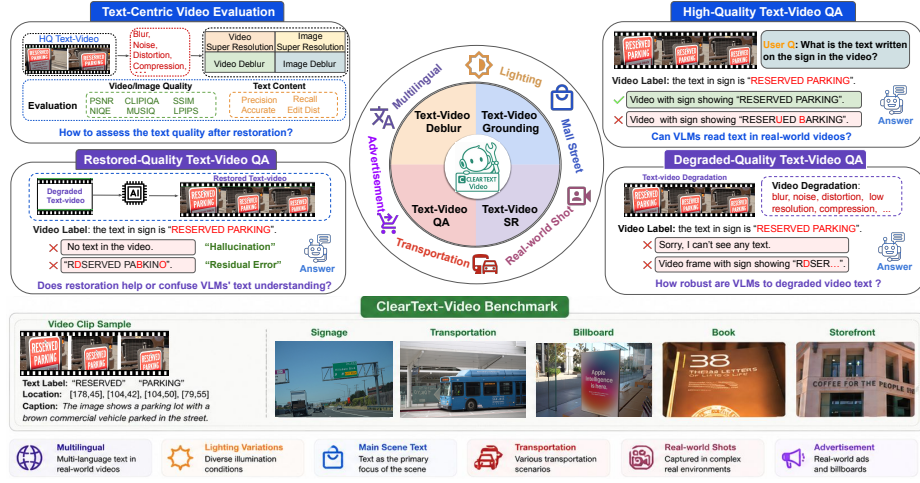


Fig. 1: Overview of the **ClearText-Video (CTVid) Benchmark**. CTVid is built from bilingual (Zh/En) egocentric videos with scene-text boxes, transcripts, captions, and spatial/temporal QA pairs. Its key design is a quality-controlled video triplet: each content instance is evaluated as High-Quality (HQ), Degraded-Quality (DQ), and Restored-Quality (RQ), allowing direct analysis of how degradation and restoration change the textual evidence available to MLLMs. Unlike restoration-only datasets or single-quality TextVQA benchmarks, CTVid links Text-Centric Video Restoration (image SR, video SR, and deblurring) with Multi-Video Quality VideoQA, testing whether models can preserve, recover, and reason over the same scene text across changing video quality.

processing of information across various modalities (*e.g.*, text, images, and video). These capabilities now underpin a wide range of real-world applications including autonomous driving, healthcare, and personal assistant service [17,30,73,46,19]. Amid this progress, Text-based Visual Question Answering (TextVQA) has emerged as a central yardstick for grounded multimodal intelligence, as it demands precise localization of visual evidence, accurate reading of scene text, and stepwise reasoning in the presence of real-world dynamics. VQA outcomes are sensitive to the quality of the input modalities and suffer when inputs exhibit poor resolution, compression artifacts, or blur. In real deployments, when users provide images or videos to general-purpose systems (*e.g.*, ChatGPT, Gemini), input quality is often suboptimal, which materially degrades answer reliability and user satisfaction.

Existing TextVQA benchmarks [73,42,74,19,42] have substantially advanced the field. As shown in Table 1, EgoTextVQA [73] introduces a text-aware videoQA benchmark spanning outdoor driving and indoor housekeeping activities; MTVQA [42] provides an image-based QA benchmark for multilingual text scenarios; Real-CE [31] establishes an image-based Chinese-English benchmark targeting text-image recovery; NewsVideoQA [19] poses questions over the textual content in news videos, requiring models to read and reason to obtain

answers; and ViTXT-GQA [74] enhances M4-ViteVQA [42] via spatiotemporal grounding annotations, enabling unified evaluation of answer grounding and QA. Taken together, these benchmarks emphasize multilingual contexts and the spatial-temporal text extraction, which is crucial for robust Text-VideoQA. However, existing datasets predominantly contain *medium-to-low-resolution* imagery/video, constraining systematic investigations into text video quality and limiting analyses of how resolution affects the text-centric capabilities of MLLMs. Crucial questions remain unanswered: *Can MLLMs reliably read and reason over in-the-wild text under diverse quality conditions?* To our knowledge, studies explicitly addressing different resolutions/qualities are limited, and current TextVQA benchmarks mostly overlook resolution variation.

To address the question above, we introduce ClearText-Video (CTVid), a novel, large-scale, high-quality, scene-text-aware video question answering benchmark. ClearText-Video is designed to advance research on input-modality quality and mainstream bilingual (Chinese/English) egocentric QA in real-world settings. We collect 4.6K+ real-world text-centric high-quality videos, covering 550K+ frames and 1.6M annotated text instances. As illustrated in Fig. 1, ClearText-Video explicitly supports controlled study of input video quality through three regimes: High-Quality (HQ), Degraded-Quality (DQ), and Restored-Quality (RQ), within a unified QA benchmark that reflects common daily scenarios. HQ features clean on-screen text and rich detail that favors MLLM encoding and reasoning; DQ is affected by blur, noise, and low resolution—typical in the wild and likely to impair text legibility and model performance; and RQ is produced by restoration models that enhance video quality and recover unreadable text, whose side effects on representation and reasoning must be systematically assessed. Unlike prior Text-VideoQA datasets that focus solely on perceptual quality or on generic text-centric reasoning, CTVid is curated to reflect real-world degradations and multilingual scene-text challenges, serving as a unified, text-centric benchmark that bridges low-level restoration and high-level video understanding. We further evaluate 16 recent state-of-the-art MLLMs and observe substantial headroom for improvement on CTVid, especially under quality variation conditions. Our main contributions are summarized as follows:

- We introduce ClearText-Video (CTVid), a large-scale, high-quality, scene-text-aware video QA benchmark targeting input-modality quality and bilingual (Zh/En) egocentric QA in real-life scenarios, with rich annotations (detection boxes, transcriptions, captions).
- We benchmark a range of MLLMs on CTVid and show persistent gaps in multilingual, quality-varying, text-rich settings, indicating significant room for advancement.
- Building on CTVid, we define two families of text-centric tasks: Text-Centric Video Restoration (including image super-resolution, video super-resolution, and video deblurring) and Multi-Video Quality Video Question Answering (covering spatial and temporal understanding). Together, these tasks establish a unified framework for evaluating textual preservation and interpretability under visual degradation, from low-level restoration to high-level reasoning.

Table 1: Comparison of **ClearText-Video** with representative video datasets. “Text in Video” indicates whether textual content appears in video frames, and “Content Language” reports the languages of embedded text.

Task	Dataset	Venue	GT Resolution	Caption Text in Video	Content Language	Videos/Img	Video Source
Video Super-Resolution	VideoLQ [4]	CVPR 2022	640×480	✗	✗	0.05K/4.9K	Online
	SPMCS [43]	ICCV 2017	1920×1080	✗	✗	0.03K/0.93K	Manual
	Vimeo-90k [59]	IJCV 2019	448×256	✗	✗	4.28K/0.6M	Online
	RealVSR [62]	ICCV 2021	1024×512	✗	✗	0.5K/2.5K	Manual
	YouHQ [72]	CVPR 2024	1920×1080	✗	✗	37K/-	Online
	MVSR4x [49]	CVPRW 2023	1920×1080	✗	✗	0.2K/2K	Manual
	Real-CE [31]	CVPR 2023	1920×1080	✗	✗	English, Chinese	Manual
Video Deblurring	GOPRO [34]	CVPR 2017	1280×720	✗	✗	-/3.2K	Manual
	BSD [71]	IJCV 2022	640×480	✗	✗	0.08K/11K	Manual
	DVD [41]	CVPR 2017	960×540	✗	✗	0.07K/6.7K	Manual
	REVD [21]	CVPR 2024	1024×768	✗	✗	0.01K/6.3K	Manual
	REDS [33]	NTIRE 2019	1280×720	✗	✗	0.3K/30K	Manual
Visual Question Answering	HD-VILA-100 [58]	CVPR 2022	1280×720	✓	✗	-/3.3M	YouTube
	InternVid [52]	ICLR 2024	1280×720	✓	✗	-/7.1M	YouTube
	TVQA [22]	EMNLP 2018	1280×720	✓	✗	-/21.7K	TV shows
	RoadTextVQA [46]	ICDAR2023	1280×720	✓	✗	English	-/3.2K
	M4-ViteVQA [70]	NeurIPS 2022	1280×720	✓	✓	English	7.6K/1.3M
	NewsVideoQA [19]	WACV 2023	1280×720	✓	✓	English	3K/0.9M
	EgoTextVQA [73]	CVPR 2025	960×540	✗	✓	English	1.5K/-
	ViT-T-GQA [74]	TMM 2025	1280×720	✓	✓	English	7.6K/1.3M
	MTVQA [42]	ACL 2025	1280×720	✓	✗	9 languages	-/8.9K
	MME-VideoOCR [39]	ArXiv 2025	1280×720	✓	✓	English	1.4K/8.9K
All	ClearText-Video (Ours)	ECCV 2026	1920×1080	✓	✓	English, Chinese	4.6K/550K



Fig. 2: Example videos and their annotated questions from the ClearText-Video benchmark. Note: **[BBOX]** denotes bounding-box coordinates, visualized as the red detection boxes in the images.

2 Related Work

Real-world Video Restoration. Video restoration aims to recover high-quality video from low-quality footage degraded by imperfect capture, motion blur, or compression [50,62,55,27]. Building on the success of diffusion models in image restoration, recent methods extend them to video [72,26,61,10]. For example, Upscale-A-Video [72] employs optical-flow-guided propagation, MGLD-VSR [61] introduces motion-aware objectives, FlashVSR [76] proposes a one-step streaming framework for real-time restoration, and DiffVSR [26] adopts staged optimization; DOVE [10] leverages large-scale video-diffusion priors to create an efficient one-step diffusion model. While these approaches markedly improve temporal consistency and perceptual quality, they provide limited treatment of semantics crucial for downstream tasks. Scene-text legibility in particular remains under-addressed, with a heightened tendency toward hallucinations.

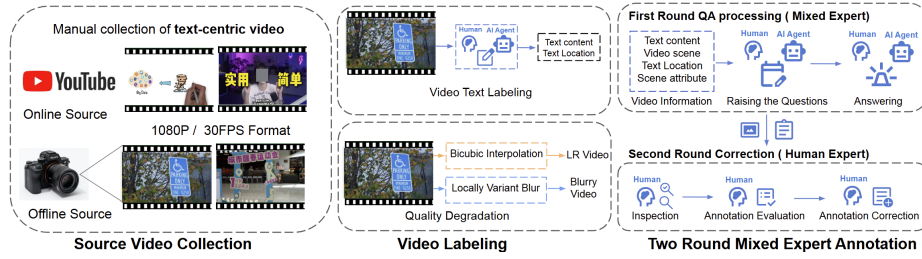


Fig. 3: A brief pipeline of the annotation process.

Multimodal Large Language Models. MLLMs exhibit strong text capabilities, making them well suited for text VQA tasks, including document understanding and text recognition [17,30]. Building on this foundation, recent MLLMs [63,29,15,75,12] extend these capabilities to video, enabling the processing of dynamic visual information. Consequently, they can not only read text in static images but also extract text signals from videos for more effective understanding. However, performance still heavily relies on dataset diversity and scale. Furthermore, there is still no comprehensive, systematic evaluation of how input video quality impacts text-centric performance—despite its central importance in real-world scenarios.

Text-aware VQA Benchmark. In scene-text VQA, numerous video-based datasets have been proposed. For example, TextVQA [40], ST-VQA [4], and ESTVQA [51] provide high-quality images with questions explicitly grounded in scene text; however, their relatively simple settings limit practical applicability. Recent work also broadens language coverage via multilingual resources, such as MTVQA [42], EgoTextVQA [73]. On the video side, NewsVideoQA [19] requires models to read and reason over on-screen text in news footage, ViTXT-GQA [74] extends M4-ViteVQA [70] with spatiotemporal grounding to jointly evaluate answer grounding and QA, and MME-VideoOCR [39] covers a broad spectrum of video OCR scenarios to support deeper comprehension and reasoning. Meanwhile, RoadTextVQA [46] provides text-rich videos, yet its questions remain largely simple and tightly focused on well-localized text, offering limited challenge for compositional reasoning and robustness.

3 ClearText-Video Dataset

Our benchmark targets multilingual, high-quality, text-rich video settings and is built from manually collected, text-centric footage rather than repurposing existing benchmarks. This section details the dataset’s construction process, annotation schema, statistics, and the unified evaluation protocol.

Video Collection. We capture videos using a Sony A7R V with a 28–70 mm f/3.5–5.6 FE lens. The variable focal length supports multi-zoom recording for diverse perspectives. Each 10 second raw video is center-cropped into a 4 second

clip, from which we extract 120 frames for text annotation. To further increase diversity, we additionally source 15% of videos from online text-centric content.

Labeling and QA Generation. The full annotation pipeline is briefly illustrated in Fig. 3. Starting from more than 6.4K candidate videos, 16 annotators conduct two rounds of annotation and verification. After collecting source videos and clipping them, we annotate each HQ video frame with text detection boxes and text transcripts. To reduce manual effort, we first use PaddleOCR [13] tool to generate coarse detection candidates, then annotators refine them with Labelme tool. The detection boxes indicate the spatial locations of text regions, while the text transcripts represent their semantic content. For video quality variants, following prior work [33], we apply bicubic downscaling ($\times 4$) to produce low-resolution videos. We apply local motion blur to produce blurry videos. These synthetic degradations do not exhaust all real-world artifacts, but they create reproducible, content-matched quality variants that isolate resolution and blur effects while keeping videos, text annotations, and QA pairs fixed. We construct text-centric VQA pairs in two rounds, which is similar to the annotation phase in [4, 73]. In the first round, we extract metadata for each text-rich video clip after labeling. They record the text content, the scene, the text locations, and relevant scene attributes. We then use GPT-4o with prompts to propose a set of candidate question types (*e.g.* multiple choice, true/false, fill-in-the-blank). Given the candidate questions, a rule-based pipeline derives answers consistent with the recorded text and fills in the selected questions with the corresponding text locations from metadata. ClearText-Video primarily focuses on two text-centric VideoQA settings: temporal understanding and spatial understanding. In the second round, to mitigate annotator and model biases and to maintain high data quality, the expert annotators visualize each clip and re-check text transcripts, text boxes, and their pairings. They verify answers and correct errors. They also screen for unethical content, including privacy-sensitive material. This round specifically removes ambiguous questions, fixes inaccurate answers, and raises task difficulty where needed.

Dataset Statistics. The ClearText-Video benchmark contains 4,639 text-centric HQ videos, 550K+ frames, 1.6M annotated text instances, and 220K+ question-answer pairs. The dataset is split into 4,327 videos for training and 312 videos for testing. For each HQ video, we also provide low-resolution and blurry counterparts. The test set contains 74% offline and 26% online clips, balanced English/Chinese coverage within each source type, and a difficulty distribution of 46% easy, 37% medium, and 17% hard questions. Fig. 4 summarizes the distributions of text carrier types and scene categories, together with word-frequency visualizations of captions and text annotations.

Dataset Comparison. We compare our ClearText-Video dataset with related datasets, as shown in Table 1. Restoration datasets in the first two domains lack text-bearing video content and captions, and thus cannot support text-aware restoration. ClearText-Video uniquely provides high-resolution videos with explicit text regions and frame-level captions, enabling restoration with both visual fidelity and textual semantics, while also offering much higher resolution

Table 2: **Text-Centric Visual Super-resolution Benchmark.** We report both general perceptual metrics and text-aware metrics. In text-aware metrics, precision ($P_{det}\%$), recall ($R_{det}\%$), and H-mean ($F1_{det}\%$) for detection, and accuracy ($Acc_{rec}\%$) and normalized edit distance ($NED_{rec}\%$) for recognition. The best and second-best results for each metric are highlighted in **red** and **blue**, respectively.

Input Type	Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	CLIPQA $\uparrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	FID $\downarrow$	$P_{det}\uparrow$	$R_{det}\uparrow$	$F1_{det}\uparrow$	$Acc_{rec}\uparrow$	$NED_{rec}\downarrow$
Image-based	Real-ESRGAN [50]	27.85	0.8748	0.1638	0.1032	0.4790	4.877	64.65	0.5718	35.00	89.06	57.78	70.09	34.82	53.51
	SwinIR [27]	28.18	0.8788	0.1613	0.1012	0.5029	4.877	65.73	0.5705	32.15	88.78	58.47	70.50	36.76	54.96
	SeeSR [55]	27.89	0.8529	0.1824	0.1161	0.6471	5.121	69.92	0.6196	27.66	87.27	68.72	76.89	35.08	54.95
	OSDiff [54]	25.67	0.8289	0.1946	0.1151	0.6219	4.852	69.11	0.6168	32.67	87.59	66.35	75.50	34.25	54.06
	S3Diff [65]	26.05	0.8033	0.1617	0.09611	0.5764	4.517	64.52	0.5799	29.16	86.72	69.52	77.17	38.42	58.03
	AdcSR [7]	25.84	0.8239	0.2024	0.124	0.6448	4.941	69.82	0.6083	37.52	88.84	57.72	69.98	32.02	51.26
Video-based	RealBasicVSR [6]	28.26	0.8924	0.1639	0.1126	0.4586	4.413	65.40	0.6042	29.91	87.78	57.15	69.23	41.97	59.30
	BasicVSR++ [5]	22.26	0.6214	0.4012	0.2429	0.5808	5.901	67.49	0.6358	18.02	91.11	54.50	68.20	35.21	55.34
	RealViFormer [69]	27.95	0.8821	0.1577	0.1198	0.4032	5.049	60.33	0.5521	32.56	88.74	48.98	63.12	29.89	46.37
	MGLD-VSR [61]	28.98	0.8813	0.1596	0.1084	0.4192	4.510	63.4	0.5819	19.66	90.21	61.23	72.95	39.66	58.66
	Upscale-A-Video [72]	27.64	0.8481	0.1875	0.1167	0.5505	4.693	66.96	0.5868	26.58	90.62	58.86	71.36	33.40	52.96
	DOVE [10]	30.24	0.9047	0.1267	0.08996	0.4185	5.210	63.91	0.5479	25.22	90.5	63.36	74.53	44.72	63.14

Table 3: **Text-Centric Video Deblurring Benchmark.** We use the same metrics as the Super-resolution Benchmark above. Best and second-best results are highlighted in **red** and **blue**, respectively.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	CLIPQA $\uparrow$	NIQE $\downarrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	FID $\downarrow$	$P_{det}\uparrow$	$R_{det}\uparrow$	$F1_{det}\uparrow$	$Acc_{rec}\uparrow$	$NED_{rec}\downarrow$
MIMO-UNet+ [11]	29.26	0.8657	0.2053	0.1393	0.3324	6.483	46.02	0.4760	47.15	88.58	65.02	74.99	51.75	69.54
NAFNet [8]	28.21	0.8716	0.2461	0.1749	0.3487	7.614	48.35	0.4253	48.65	90.81	66.37	76.69	50.77	68.30
Restormer [64]	31.38	0.8863	0.1904	0.1349	0.3404	6.552	47.31	0.4864	41.50	88.28	67.08	76.23	54.99	71.67
Stripformer(GoPro) [47]	31.13	0.8832	0.1754	0.1191	0.3348	6.386	48.65	0.4955	38.78	88.42	66.69	76.03	55.05	72.26
Stripformer(Realblur_J) [47]	27.04	0.8655	0.1837	0.1182	0.3657	6.240	54.44	0.4922	32.36	89.18	69.99	78.43	58.93	76.51
Stripformer(Realblur_R) [47]	27.36	0.8667	0.1905	0.1166	0.3396	6.490	50.41	0.4668	33.46	89.45	69.71	78.36	58.57	75.71
RVRT(DVD) [28]	28.83	0.8489	0.2671	0.1843	0.2902	7.158	34.37	0.4138	58.32	90.61	61.32	73.14	44.19	60.61
RVRT(GoPro) [28]	28.77	0.8553	0.2380	0.1622	0.2959	6.879	38.92	0.4407	52.91	89.63	61.25	72.77	46.23	63.31
ShiftNet(DVD) [24]	28.47	0.8445	0.2771	0.1913	0.3077	7.334	33.52	0.4065	58.76	91.05	61.15	73.16	43.10	59.56
ShiftNet(GoPro) [24]	28.20	0.8434	0.2570	0.1763	0.3004	6.965	35.69	0.4244	56.64	90.47	60.54	72.54	43.88	61.40

than existing Video QA datasets for fine-grained visual and textual understanding. Its central distinction is the quality-controlled setting: the same underlying videos support paired HQ/DQ/RQ evaluation, allowing us to test whether restoration preserves the textual evidence required for downstream reasoning rather than merely improving perceptual appearance.

Data Release and Licensing. We will release train/test splits, human-verified boxes and transcripts, captions, temporal trajectories, spatial/temporal QA pairs, degradation scripts, evaluation code, prompts, and baseline outputs. Self-captured clips and annotations will be distributed under a research license. For online-source clips whose redistribution is restricted, we will release source IDs/URLs where allowed, clip ranges, annotations, and preprocessing scripts, and redistribute raw videos only when permission or compatible licenses allow it.

4 ClearText-Video Benchmark

Our ClearText-Video benchmark is designed to support text-centric, video-related tasks across both low-level restoration and high-level understanding. It provides paired, temporally aligned videos for restoration tasks—specifically,

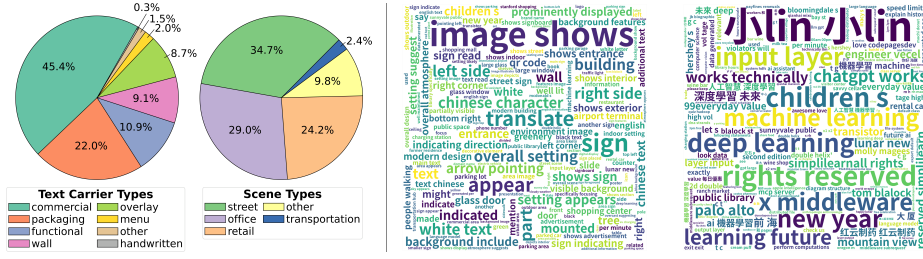


Fig. 4: Statistics of the CTVid dataset. The left panels summarize text carrier types and scene categories, while the right panels show word-frequency visualizations for video captions and frame-level text annotations.

image and video super-resolution as well as video deblurring. Building on this foundation, we introduce a multi-quality VideoQA benchmark for text-centric visual understanding, constructed from triplets of High-Quality (HQ), Degraded-Quality (DQ), and Restored-Quality (RQ) videos. This design enables rigorous evaluation of MLLMs under varying input quality and supports analyses of their text-semantic understanding in both spatial and temporal domains.

4.1 Text-Centric Video Restoration

Text-Centric Image SR task. Text-centric image super-resolution aims to reconstruct high-resolution frames from $4\times$ bicubic down-sampled inputs [33]. Each CTVID frame includes a caption, enabling research on semantics-aware or language-guided super-resolution, where text aids in restoring degraded visuals.

Text-Centric Video SR task. Text-centric video super-resolution reconstructs high-resolution videos from low-resolution inputs. If semantic information is needed, the caption of the middle frame serves as the video’s representative caption.

Text-Centric Video Deblurring task. The Text-Centric Video Deblurring task seeks to recover sharp images from synthetically blurred inputs. Blur is synthesized by applying $16\times$ frame interpolation with RIFE [18] to convert 30 fps clips to 480 fps, followed by temporally fusing every 65 frames to emulate camera or object motion.

4.2 Text-Centric Multi-Video Quality VideoQA

Text-Centric VideoQA for Spatial Understanding task. This task focuses on understanding scene text within its spatial context in video frames. For each frame, we generate spatially grounded questions, multiple-choice, true/false, and fill-in-the-blank, based on text that appears inside a specific bounding box region. A rule-based spatial partitioning method identifies meaningful regions by analyzing the layout of detected text and selecting cutting planes to form bounding boxes that isolate specific text groups. the GPT-4o then formulates

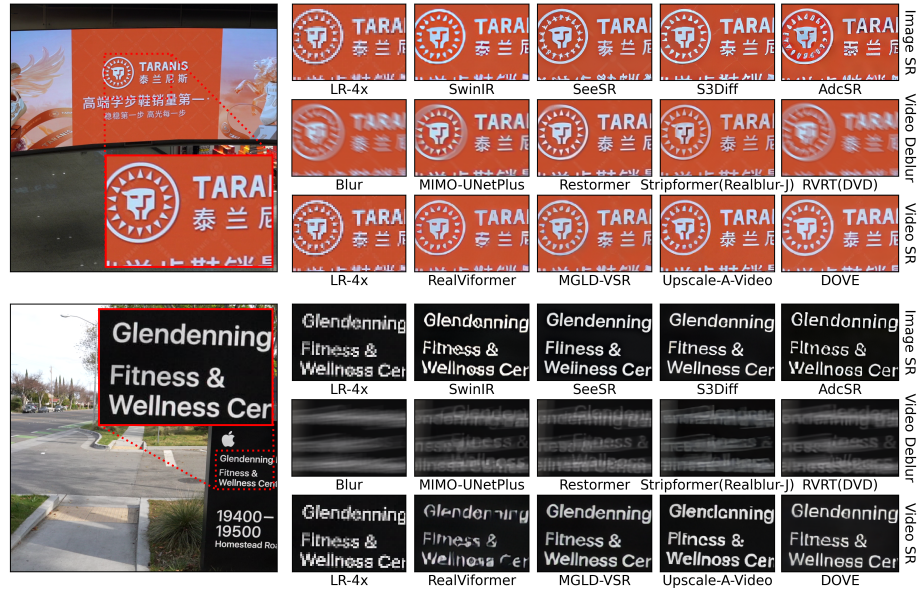


Fig. 5: Qualitative results on the Text-Centric Video Restoration benchmarks using the CTVid testing set. Results are shown for both English and Chinese text instances. The examples include video super-resolution and deblurring outputs, highlighting differences in text fidelity and legibility across methods.

natural-language questions using these regions, while a validation module ensures the questions are spatially correct, well-structured, and answerable. Then we manually ensure that answering requires genuine spatial reasoning rather than simple text matching.

Text-Centric VideoQA for Temporal Understanding task. This task evaluates a model’s ability to understand how scene text moves and changes over time within a video. Given 120-frame clips, questions require reasoning about temporal visibility patterns, motion behavior, spatial transitions, and size variations of text instances. All questions are generated deterministically through rule-based analysis of text trajectories. For each text instance, we track frame-by-frame properties such as position, area, speed, and acceleration, and use these temporal features to construct questions across five categories: visibility, spatial localization over time, motion and trajectory shape, size and scale changes, and boundary interactions. After the rule-based generation step, annotators manually further screen and refine the outputs to ensure that only meaningful and high-quality questions are retained. This hybrid pipeline guarantees consistency and reproducibility while avoiding the impact of LLM hallucinations.

4.3 Evaluation Protocol.

A) *Evaluation on visual Restoration.* We evaluate restoration effectiveness with both image-quality and text-centric measures. For image quality assessment, we report both full-reference and no-reference metrics. For full-reference IQA, we use PSNR, SSIM [53], LPIPS [68], DISTS [14], and FID [16]; For no-reference IQA, we use NIQE [67], MANIQA [60], MUSIQ [20], and CLIPQA [48]. To quantify text-aware semantic quality in restored regions, we employ a unified PaddleOCR [13] pipeline for detection, and we assess text-aware semantic restoration quality indirectly via text detection and recognition results. We adopt Precision (P_{det}), Recall (R_{det}), and H-mean ($F1_{det}$) to measure text detection performance, while Text recognition accuracy (Acc_{rec}) and Normalized Edit Distance (NED_{rec}) are used to measure text recognition performance.

B) *Evaluation on visual question answering.* We evaluate VideoQA with four metrics: accuracy (Acc), uncertainty-aware accuracy (UAcc), over-confidence ratio (OC), and abstention rate (Abs). UAcc and OC follow the definitions in [56], assessing reliability by coupling correctness with confidence. Abs measures the fraction of queries on which the model withholds an answer [57].

Table 4: **Text-Centric VideoQA for Spatial Understanding.** We evaluate accuracy (Acc), uncertainty-aware accuracy (UAcc), over-confidence ratio (OC), and abstention rate (Abs) across six input-quality conditions: HQ (High-Quality), DQ-Low_res [33] (low-resolution degradation), DQ-Blur [33] (blur degradation), RQ-DOVE [10] (video-based SR), RQ-MIMO [11] (blur restoration), and RQ-S3DIFF [65] (image-based SR). “Qwen2.5-VL-7B-SFT” denotes Qwen2.5-VL-7B instruction-tuned on the ClearText-Video training set. For Acc, OC, and UAcc, a single best and second-best result per column is highlighted in **red** and **blue**, respectively, across all models. All reported values are percentages (%).

Model	HQ				DQ-Low_res [33]				DQ-Blur [33]				RQ-DOVE [10]				RQ-MIMO [11]				RQ-S3DIFF [65]			
	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs	Acc $\uparrow$	OC $\downarrow$	UAcc $\uparrow$	Abs
GPT-5.4 [36]	60.00	15.00	55.00	0.00	53.33	10.00	58.33	6.70	56.67	8.33	60.00	18.30	48.33	15.00	55.00	3.30	55.00	10.00	58.33	6.70	53.33	13.33	53.33	0.00
GPT-5.4-mini [36]	46.67	6.67	65.00	0.00	48.33	18.33	51.67	0.00	43.33	5.00	63.33	3.30	46.67	3.33	66.67	0.00	41.67	11.67	56.67	0.00	45.00	8.33	56.67	0.00
Claude-Sonnet-4.6 [2]	70.00	21.67	66.67	0.00	63.33	20.00	70.00	0.00	60.00	20.00	70.00	13.30	65.00	18.33	70.00	0.00	70.00	20.00	70.00	3.30	61.67	16.67	71.67	0.00
Gemini-2.5-pro [12]	71.67	16.67	65.00	0.00	65.00	8.33	71.67	1.70	60.00	13.33	66.67	23.30	65.00	13.33	68.33	1.70	63.33	20.00	56.67	10.00	70.00	11.67	65.00	0.00
Gemini-2.5-flash [12]	60.00	21.67	61.67	1.70	65.00	10.00	61.67	1.70	48.33	10.00	76.67	30.00	66.67	10.00	66.67	3.30	60.00	16.67	61.67	11.70	61.67	13.33	58.33	1.70
InternVL2.5-8B [9]	39.65	59.09	40.76	9.80	38.67	59.74	40.13	10.15	36.82	60.49	39.35	12.58	39.12	59.40	40.49	9.84	37.63	60.28	39.56	11.70	38.31	60.20	39.68	10.05
InternVL3-8B [75]	40.78	57.94	41.85	0.85	39.80	58.85	40.86	0.46	36.89	61.42	38.37	0.41	40.19	58.49	41.22	0.75	37.88	60.63	39.14	0.50	39.44	59.24	40.49	1.02
Llama3.2-11B [32]	34.21	57.70	35.66	0.28	31.70	60.47	33.50	0.38	30.22	59.29	32.48	0.42	32.38	59.62	33.79	0.26	31.29	58.90	33.14	0.38	32.13	60.38	33.08	0.19
Llama3-llava-next-8b [25]	30.89	50.61	36.81	0.42	30.40	52.62	36.51	0.44	29.81	51.08	37.30	1.63	30.49	51.12	36.45	0.58	29.88	51.46	36.67	0.90	30.55	51.16	36.47	0.65
LLaVA-OneVision [23]	33.90	63.18	35.19	0.52	32.66	63.42	34.15	0.87	31.04	60.76	34.40	2.33	32.95	62.97	34.76	1.05	31.66	62.53	34.05	1.16	32.74	62.95	34.55	1.01
MiniCPM-O 2.6 [37]	39.47	57.87	39.37	1.57	36.00	63.21	35.84	1.85	32.48	55.42	32.72	2.61	37.48	60.05	37.30	1.82	33.26	55.80	33.30	1.84	36.97	60.22	36.79	1.66
MiniCPM-V 4.5 [45]	44.26	49.32	45.82	0.12	40.90	50.18	42.88	0.08	35.69	48.11	37.89	0.95	41.26	51.61	43.09	0.24	37.58	47.80	38.58	0.28	40.55	51.47	42.51	0.21
Phi-4-multimodal [1]	33.26	63.81	34.70	0.18	31.61	59.87	35.18	0.66	29.32	64.49	32.57	3.02	31.60	63.97	33.96	0.64	30.42	64.71	33.17	1.46	31.25	63.65	33.91	0.56
Qwen3-VL-8B [38]	50.56	46.21	52.61	0.63	45.43	50.52	49.00	0.87	44.46	46.50	52.18	3.87	47.70	46.94	51.80	1.30	46.91	47.89	51.03	1.73	46.64	47.93	51.03	1.05
Qwen2.5-VL-7B [3]	49.99	21.36	60.66	0.17	40.61	15.72	62.38	0.61	42.16	16.46	63.26	1.74	45.37	20.22	60.73	0.26	44.85	20.08	61.21	0.46	45.00	19.01	61.21	0.23
Qwen2.5-VL-7B-SFT	58.43	1.73	43.71	0.08	50.72	4.14	50.12	0.12	49.79	1.11	50.83	0.06	53.21	1.97	48.45	0.05	52.89	1.56	48.53	0.06	51.31	1.66	49.79	0.08

5 Experiments

5.1 Experimental setup

Training Details. We employ Qwen2.5-VL-7B [3] as the base model for all supervised fine-tuning experiments. Training parameters follow the default configuration of the official implementation [3]. Depending on the question type, task-specific system prompts are dynamically loaded to guide the model. To preserve pretrained visual representations, the vision encoder and multimodal projector are frozen during fine-tuning. The model is trained for 10 epochs on $8\times$ NVIDIA A100 GPUs.

Compared Methods. All evaluations are conducted on our CTVid test set, covering Text-Centric Video Restoration and Text-Centric Multi-Video Quality VideoQA tasks. For the restoration task, we evaluate both classic quality enhancement models and recent diffusion-based methods. For the QA task, we assess open-source and closed-source MLLMs.

5.2 Evaluation on Text-Centric Video Restoration

Text-Centric Image Super-Resolution. We benchmark six SOTA image SR methods on the CTVid test set as shown in Table 2. The evaluated models include classical approaches[50,27] and recent diffusion-based techniques[55,54,7,65]. Our evaluation employs both perceptual image quality metrics and text-aware measures. Good performance in perceptual metrics does not guarantee preserving textual fidelity. While SwinIR[27] and SeeSR[55] lead image-based methods according to perceptual metrics, S3Diff[65] achieves the highest text detection and recognition accuracy. Qualitative results in Fig. 5 show that all models still introduce distortions in challenging text regions. Classical methods tend to over-smooth small or degraded characters, whereas diffusion-based models produce sharper structures but may hallucinate fine details. In the second example, SeeSR misrecognizes “Fitness” as “Fliness” and “Wellness” as “Weiiness”, while SwinIR fails to produce legible letters. In the first example, the more complex Chinese characters undergo even more severe degradation.

Text-Centric Video Super-Resolution. We benchmark six SOTA VSR models on the CTVid dataset, including classical methods[6,5,69] and recent diffusion-based approaches[61,72,10]. As shown in Table 2, DOVE[10] achieves 63.36% in P_{det} , 74.53% in R_{det} and 74.53% in $F1_{det}$ as the best performance, highlighting its effectiveness in preserving textual content under temporal dynamics. RealViFormer[69] obtains the best NED_{rec} of 46.37%, while BasicVSR++[5] obtains highest P_{det} of 91.11%. Qualitative results in Figure 5 show that, all models introduce some text distortion. Nevertheless, video-based methods are able to leverage temporal information to enhance consistency and legibility. In the first example, DOVE successfully restores complex Chinese characters with almost zero distortion.

Text-Centric Video Deblurring. We evaluate text-centric video deblurring using both image- and video-based SOTA models[11,8,47,28,24]. The evaluation protocol mirrors that of the super-resolution tasks. As shown in Table 3,

Stripformer[47] achieves 74.53% in R_{det} and 78.43% in $F1_{det}$, and 58.93% in Acc_{rec} as the best scores. And RVRT[28] yields the highest P_{det} of 91.05% and the lowest NED_{rec} of 59.56%, demonstrating superior preservation of fine-grained structure under motion blur. Figure 5 illustrates that, with effective deblurring, all methods introduce only minor text distortions largely due to input frames sharing resolution with the ground truth. However, failure cases emerge under large motion, including partial restoration or complete breakdown when motion exceeds model capacity.

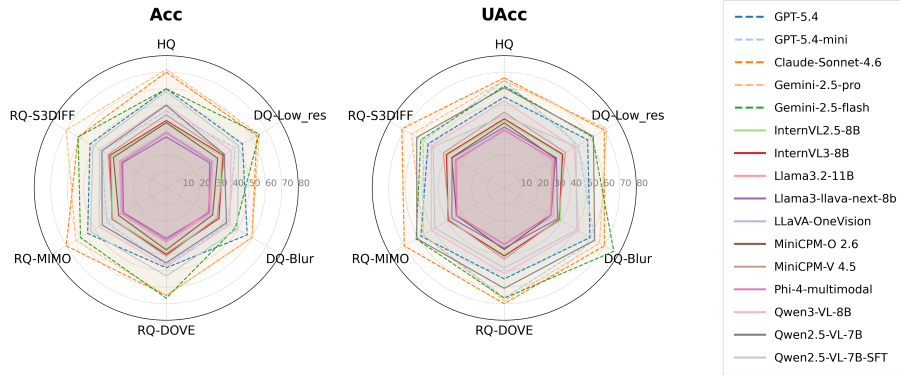


Fig. 6: Accuracy (Acc) and uncertainty-aware accuracy (UAcc) radar plots for multiple MLLMs across six video-quality conditions. The top plot reports Acc, and the bottom plot reports UAcc.

5.3 Evaluation on Multi-Video Quality VideoQA

Text-centric VideoQA for Spatial Understanding. We benchmark 16 MLLMs on the CTVID test set, including five proprietary models [36,2,12] and ten open-source baselines [3,9,75,32,25,23,37,45,1,38], together with our Qwen2.5-VL-7B-SFT variant. As shown in Table 4, proprietary MLLMs still achieve the strongest spatial VideoQA accuracy, but their robustness varies across quality conditions. Gemini-2.5-pro obtains the best accuracy on HQ, DQ-Low_res, and RQ-S3DIFF inputs, with scores of 71.67%, 65.00%, and 70.00%, respectively. Claude-Sonnet-4.6 leads on DQ-Blur and RQ-MIMO, reaching 60.00% and 70.00%, while Gemini-2.5-flash performs best on RQ-DOVE at 66.67%. These results show that restoration does not yield a uniformly monotonic improvement: the best model can change with the restoration source, and perceptually enhanced videos may still alter text evidence in ways that affect downstream reasoning. Among open-source models, Qwen3-VL-8B gives the strongest zero-shot accuracy on HQ input (50.56%), whereas Qwen2.5-VL-7B yields the highest open-source UAcc on all six conditions. After instruction tuning on ClearText-Video, Qwen2.5-VL-7B-SFT achieves the best open-source accuracy across every quality condition

(58.43%, 50.72%, 49.79%, 53.21%, 52.89%, and 51.31%), improving over the Qwen2.5-VL-7B base model by 6.31–10.11 percentage points. The tuned model also substantially reduces over-confidence, but its lower UAcc indicates that accuracy gains and uncertainty-aware reliability do not always move together. The radar chart in Fig. 6 summarizes these accuracy and UAcc trends across the six quality conditions. A controlled LR-vs-blur analysis further shows that blur is more damaging than low resolution: averaged over 16 MLLMs, LR reduces spatial QA accuracy by 3.14 points from HQ, whereas blur reduces it by 6.05 points. An OCR+LLM diagnostic baseline remains far below direct MLLM inference, confirming that CTVID requires spatial grounding and multimodal reasoning beyond extracted OCR text.

Text-centric VideoQA for Temporal Understanding. The temporal understanding VideoQA task follows the same input protocol as the spatial setting using five input video-quality conditions, which are uniformly $2\times$ downsampled to satisfy memory constraints. To better assess temporal reasoning, we adopt a fixed downsampling scheme and minimize frame sampling whenever possible, thereby preserving temporal continuity across sequences. As shown in Table 5, Claude-Sonnet-4.6 consistently ranks first across all five temporal conditions, obtaining 60.02% on HQ, 59.45% on DQ-Blur, 59.34% on RQ-MIMO, 58.88% on RQ-S3DIFF, and 60.37% on RQ-DOVE, and is the only model to exceed 58% under any condition. In contrast to the spatial setting, the proprietary advantage does not generalize here: the remaining proprietary models (Gemini-2.5-flash, Gemini-2.5-pro, and GPT-4o-mini) fall into the high-40% to low-50% band, on par with or below the open-source baselines, with Gemini-2.5-pro dropping to 45.83% on HQ. Among open-source models, performance is concentrated in the low-50% range. Kimi-VL-16B achieves the best open-source accuracy on HQ, DQ-Blur, RQ-S3DIFF, and RQ-DOVE, while InternVL3-8B is marginally better on RQ-MIMO. Claude-Sonnet-4.6 consistently leads the strongest open-source model by about six percentage points, with the margin reaching 6.30 points on RQ-DOVE and remaining above 5 points in all five temporal settings. This indicates that temporal text reasoning remains challenging even when individual frames contain readable text. Model rankings also vary across quality conditions: restoration improves some models but causes unstable changes for others, suggesting that temporal aggregation and restoration-induced artifacts both affect answer reliability.

6 Discussion and Limitations

Restoration and understanding are coupled. CTVID does not treat restoration as an independent leaderboard detached from VideoQA. Instead, restoration constructs the RQ regime and enables a diagnostic question that prior single-quality TextVQA benchmarks cannot isolate: whether visually improving degraded text videos preserves the exact textual evidence required for downstream reasoning. Our results show that this relationship is not monotonic. Some restored videos become sharper but still alter character strokes, introduce plausible-looking

Table 5: **Text-Centric Video Question Answering for Temporal Understanding.** We report accuracy across five video-quality conditions. The best and second-best results in each quality condition are highlighted in **red** and **blue**, respectively, across all models.

Accuracy%(↑)					
Models	HQ	DQ-Blur [33]	RQ-MIMO [11]	RQ-S3DIFF [65]	RQ-DOVE [10]
Gemini-2.5-flash [12]	51.55	51.20	52.12	51.32	53.04
Gemini-2.5-pro [12]	45.83	50.00	51.39	50.00	48.61
GPT-4o-mini [35]	50.06	50.52	50.29	50.13	49.94
Claude-Sonnet-4.6 [2]	60.02	59.45	59.34	58.88	60.37
Qwen2.5-VL-7B [3]	50.86	50.06	50.40	50.29	50.63
Kimi-VL-16B [44]	53.95	53.49	52.92	53.49	54.18
InternVL3-8B [75]	52.92	52.46	53.04	51.66	52.81
VideoLLaMA3-7B [66]	52.46	51.66	52.92	50.63	50.40

text, or shift spatial cues, which can leave MLLMs less accurate than on degraded inputs. This supports evaluating restoration with both visual/text-fidelity metrics and downstream reasoning metrics.

Controlled degradations and data release. Our DQ videos use controlled bicubic downsampling and motion blur to keep the underlying content, text annotations, and QA pairs fixed across HQ/DQ/RQ comparisons. This choice enables reproducible causal analysis of resolution and blur, but it does not cover every artifact present in user videos, such as exposure changes, rolling shutter, sensor noise, ISP effects, and complex compression. Future versions of CTVid can extend the same paired protocol to richer real-world degradations. The release will include annotations, QA pairs, code, prompts, and permitted video assets/metadata so that these extensions can be reproduced under the same benchmark contract.

OCR probes and calibration. PaddleOCR is not used as ground truth. It initializes coarse text candidates during annotation and serves as a fixed OCR probe for reproducible text-fidelity measurement, while human annotators verify boxes, transcripts, QA pairs, and ambiguity. The OCR+LLM diagnostic baseline further confirms that extracted text alone is insufficient: CTVid still requires spatial grounding and multimodal reasoning. Finally, supervised fine-tuning on CTVid improves answer accuracy but does not uniformly improve reliability; lower UAcc and changed over-confidence indicate an accuracy-calibration trade-off that future models should address explicitly.

7 Conclusion

In this work, we introduced ClearText-Video (CTVid), the first large-scale, scene-text-aware video QA benchmark designed to examine how input video quality affects text-centric multimodal reasoning. CTVid is built around a controlled

quality protocol: the same underlying text-rich videos are evaluated under High-Quality, Degraded-Quality, and Restored-Quality regimes. This design allows us to separate the effects of degradation, restoration, and model reasoning, which is difficult to study with existing single-quality TextVQA or restoration datasets. Beyond dataset construction, CTVid defines a unified evaluation suite that connects low-level restoration with high-level video understanding. The benchmark supports text-centric image super-resolution, video super-resolution, video deblurring, and multi-quality VideoQA, together with human-verified text annotations, captions, and spatial/temporal QA pairs. This setup makes it possible to evaluate not only whether a method improves visual quality, but also whether it preserves the textual evidence needed for downstream reasoning. Experiments with 18 restoration methods and 16 MLLMs show that current systems remain fragile under quality changes: restoration does not always improve reasoning accuracy, and visually enhanced videos can still alter textual evidence needed by MLLMs. We expect CTVid to support future research on text-faithful restoration, quality-robust multimodal reasoning, and evaluation protocols that jointly measure visual enhancement and text-grounded understanding.

References

1. Abouelenin, A., Ashfaq, A., Atkinson, A., others: Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743 (2025), <https://arxiv.org/abs/2503.01743>, covers Phi-4-multimodal-instruct; accessed 30 June 2026
2. Anthropic: Claude sonnet 4.6. <https://www.anthropic.com/claude> (2025), anthropic large multimodal model; accessed via the Claude API; accessed 30 June 2026
3. Bai, S., Chen, K., Liu, X., others: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025), <https://arxiv.org/abs/2502.13923>, for Qwen2.5-VL-7B-Instruct; accessed 30 June 2026
4. 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)
5. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
6. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5962–5971 (2022)
7. Chen, B., Li, G., Wu, R., Zhang, X., Chen, J., Zhang, J., Zhang, L.: Adversarial diffusion compression for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
8. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. arXiv preprint arXiv:2204.04676 (2022)
9. Chen, Z., Wang, W., Cao, Y., others: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024), <https://arxiv.org/abs/2412.05271>, internVL 2.5 series (incl. 8B variants); accessed 30 June 2026

10. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: Dove: Efficient one-step diffusion model for real-world video super-resolution. In: NeurIPS (2025)
11. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4641–4650 (2021)
12. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., others: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025), <https://arxiv.org/abs/2507.06261>, covers Gemini 2.5 Pro and 2.5 Flash; accessed 30 June 2026
13. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., Zhang, Y., Lv, W., Huang, K., Zhang, Y., Zhang, J., Zhang, J., Liu, Y., Yu, D., Ma, Y.: Paddleocr 3.0 technical report (2025), <https://arxiv.org/abs/2507.05595>, accessed 30 June 2026
14. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2022). <https://doi.org/10.1109/TPAMI.2020.3045810>
15. Fu, C., Lin, H., Long, Z., Shen, Y., Dai, Y., Zhao, M., Zhang, Y.F., Dong, S., Li, Y., Wang, X., et al.: Vita: Towards open-source interactive omni multimodal llm. arXiv preprint arXiv:2408.05211 (2024)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
17. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3096–3120 (2024)
18. Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Real-time intermediate flow estimation for video frame interpolation. In: Proceedings of the European Conference on Computer Vision (2022)
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. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5148–5157 (2021)
21. Kim, T., Cho, H., Yoon, K.J.: Frequency-aware event-based video deblurring for real-world motion blur. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24966–24976 (2024)
22. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
23. Li, B., Zhang, Y., Guo, D., others: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024), <https://arxiv.org/abs/2408.03326>, for llava-onevision-qwen2-7b-ov-hf; accessed 30 June 2026
24. Li, D., Shi, X., Zhang, Y., Cheung, K.C., See, S., Wang, X., Qin, H., Li, H.: A simple baseline for video restoration with grouped spatial-temporal shift. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9822–9832 (June 2023)

25. Li, F., Zhang, R., Zhang, H., others: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024), <https://arxiv.org/abs/2407.07895>, cites the llama3-llava-next-8b lineage; accessed 30 June 2026
26. Li, X., Liu, Y., Cao, S., Chen, Z., Zhuang, S., Chen, X., He, Y., Wang, Y., Qiao, Y.: Diffvsr: Enhancing real-world video super-resolution with diffusion models for advanced visual quality and temporal consistency. arXiv e-prints pp. arXiv-2501 (2025)
27. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
28. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Van Gool, L.: Recurrent video restoration transformer with guided deformable attention. arXiv preprint arXiv:2206.02146 (2022)
29. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
30. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutlm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024)
31. Ma, J., Liang, Z., Xiang, W., Yang, X., Zhang, L.: A benchmark for chinese-english scene text image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19452–19461 (2023)
32. Meta AI: Llama-3.2-11b-vision-instruct. Available at <https://huggingface.co/meta-llama/> (2024), model release page; accessed 30 June 2026
33. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
34. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3883–3891 (2017)
35. OpenAI: Gpt-4o system card. Tech. rep., OpenAI (May 2024), <https://cdn.openai.com/gpt-4o-system-card.pdf>, safety/system report describing GPT-4o; accessed 30 June 2026
36. OpenAI: Gpt-5. <https://openai.com/index/gpt-5/> (2025), covers GPT-5.4 and GPT-5.4-mini; accessed 30 June 2026
37. OpenBMB Team: Minicpm-o 2.6: A gpt-4o-level mllm for vision, speech, and multimodal live streaming on your phone. Online; OpenBMB Notion Page (2025), <https://github.com/OpenBMB/MiniCPM-o>, accessed 2025-11-12
38. Qwen Team, Alibaba Cloud: Qwen3-vl: Sharper vision, deeper thought, broader action. Blog post, Qwen Research (Sep 2025), <https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef>, announcement of the Qwen3-VL vision-language model family; accessed 30 June 2026
39. 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)
40. 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)

41. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1279–1288 (2017)
42. Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Wang, A.L., Lin, C., Feng, H., Zhao, Z., Wang, Y., et al.: Mtvqa: Benchmarking multilingual text-centric visual question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 7748–7763 (2025)
43. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE international conference on computer vision. pp. 4472–4480 (2017)
44. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
45. Team, O.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025), <https://arxiv.org/pdf/2509.18154>, details the 4.5 release; accessed 30 June 2026
46. 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)
47. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: Stripformer: Strip transformer for fast image deblurring. In: European conference on computer vision. pp. 146–162. Springer (2022)
48. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
49. Wang, R., Liu, X., Zhang, Z., Wu, X., Feng, C.M., Zhang, L., Zuo, W.: Benchmark dataset and effective inter-frame alignment for real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1168–1177 (2023)
50. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
51. 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)
52. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: The Twelfth International Conference on Learning Representations (2023)
53. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on* **13**, 600 – 612 (05 2004). <https://doi.org/10.1109/TIP.2003.819861>
54. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems* **37**, 92529–92553 (2024)
55. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
56. Xia, P., Chen, Z., Tian, J., Gong, Y., Hou, R., Xu, Y., Wu, Z., Fan, Z., Zhou, Y., Zhu, K., et al.: Cares: A comprehensive benchmark of trustworthiness in medical

- vision language models. *Advances in Neural Information Processing Systems* **37**, 140334–140365 (2024)
57. Xing, S., Hua, H., Gao, X., Zhu, S., Li, R., Tian, K., Li, X., Huang, H., Yang, T., Wang, Z., Zhou, Y., Yao, H., Tu, Z.: AutoTrust: Benchmarking Trustworthiness in Large Vision Language Models for Autonomous Driving. *arXiv* (Dec 2024). <https://doi.org/10.48550/arXiv.2412.15206>
 58. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5036–5045 (2022)
 59. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. *International Journal of Computer Vision* **127**(8), 1106–1125 (2019)
 60. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1191–1200 (2022)
 61. Yang, X., He, C., Ma, J., Zhang, L.: Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In: *Proceedings of the European Conference on Computer Vision* (2024)
 62. Yang, X., Xiang, W., Zeng, H., Zhang, L.: Real-world video super-resolution: A benchmark dataset and a decomposition based learning scheme. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4781–4790 (2021)
 63. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800* (2024)
 64. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5728–5739 (2022)
 65. Zhang, A., Yue, Z., Pei, R., Ren, W., Cao, X.: Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058* (2024)
 66. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106* (2025)
 67. Zhang, L., Zhang, L., Bovik, A.C.: A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* **24**(8), 2579–2591 (2015)
 68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
 69. Zhang, Y., Yao, A.: Realviformer: Investigating attention for real-world video super-resolution. In: *European Conference on Computer Vision*. pp. 412–428. Springer (2024)
 70. 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)
 71. Zhong, Z., Gao, Y., Zheng, Y., Zheng, B., Sato, I.: Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *International Journal of Computer Vision* **131**(1), 284–301 (2023)

72. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2535–2545 (2024)
73. 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)
74. Zhou, S., Xiao, J., Yang, X., Song, P., Guo, D., Yao, A., Wang, M., Chua, T.S.: Scene-text grounding for text-based video question answering. arXiv preprint arXiv:2409.14319 (2024)
75. Zhu, J., others: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025), <https://arxiv.org/abs/2504.10479>, accessed 30 June 2026
76. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)