VUDG: A Dataset for Video Understanding Domain Generalization

Ziyi Wang[1,2], Zhi Gao[1,2], Boxuan Yu[1], Zirui Dai[1], Peiyao Wang[1], Yuxiang Song[1], Qingyuan Lu[1], Jin Chen[1], Xinxiao Wu[1,2]

[1]Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University

BACKGROUND

Deploying Large Vision-Language Models (LVLMs) in real-world scenarios requires robustness against distribution shifts, known as Domain Generalization (DG). However, existing video benchmarks confound domain shifts with semantic differences, hindering rigorous DG evaluation. To bridge this gap, we introduce VUDG, the first dataset dedicated to evaluating domain generalization in video understanding. VUDG comprises 11 domains with diverse visual styles and conditions while maintaining semantic consistency, enabling a reliable assessment of model generalization capabilities.

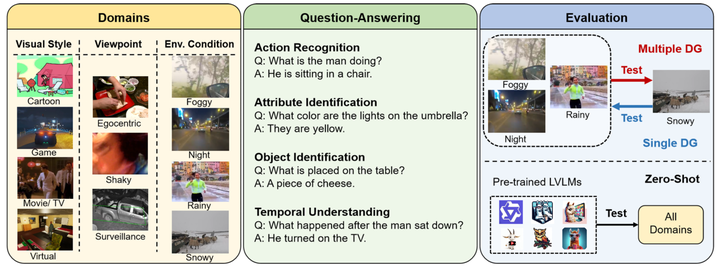
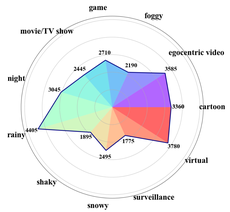


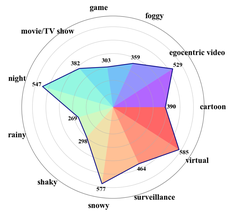
Figure 1: Overview of the proposed VUDG dataset.

STATISTICS

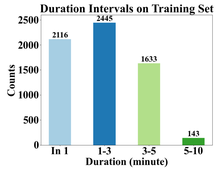
Our VUDG comprises 6,337 video clips (31,685 QA pairs) for the training set and 1,532 video clips (4,703 QA pairs) for the testing set.

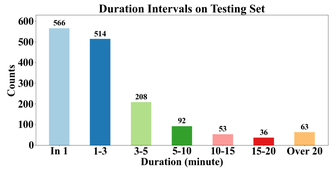
The following displays the distributions of videos across domains for the training set (left) and the testing set (right).



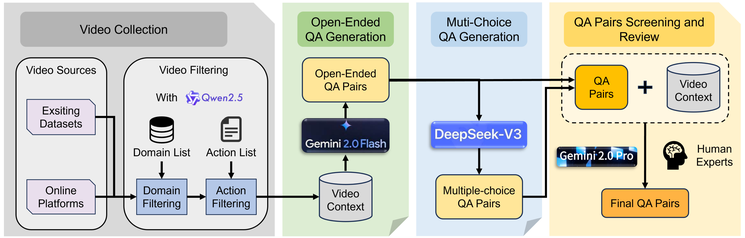


The following displays the duration distribution for the training set (left) and the testing set (right).





ANNOTATION



To construct VUDG, we introduce a progressive multi-expert annotation framework. A key feature of our framework is ensuring semantic consistency across all 11 domains by pre-defining a shared space of daily human activities. The framework employs a cascade of distinct large models for automated QA generation and verification, mitigating self-reinforcement bias and streamlines the final human review process to ensure data quality.

RESULTS

Table 6: Multiple-choice zero-shot test results on VUDG. Performance across 11 domains.

Model	Visual Style					Viewpoint				Env. Condition					D-Avg
	CA	GA	MO	VI	Avg	EG	SU	SH	Avg	FO	NI	RA	SN	Avg	
Video-ChatGPT-7B (Maaz et al., 2024)	14.1	12.5	14.4	9.7	12.7	12.9	11.6	14.1	12.9	14.5	16.3	13.8	15.8	15.1	13.6
MiniGPT4-Video (Ataallah et al., 2024)	13.6	12.9	13.9	14.2	13.7	12.7	13.2	13.8	13.2	15.9	15.7	13.0	13.0	14.4	13.8
VideoChat2-7B (Li et al., 2023)	16.2	9.6	14.1	10.3	12.6	14.6	13.8	13.4	13.9	16.7	15.4	17.8	11.6	15.4	14.0
Video-LLaVA-7B (Lin et al., 2023a)	23.3	23.1	21.2	29.9	24.4	22.3	26.1	19.5	22.6	22.6	26.0	20.1	25.0	23.4	23.5
VideoLLaMA2-7B (Cheng et al., 2024)	31.5	34.0	31.7	30.4	31.9	34.6	34.5	39.6	36.2	33.4	34.7	30.5	32.2	32.7	33.4
mPLUG-Owl3-7B (Ye et al., 2025)	50.0	50.8	49.7	61.0	52.9	53.5	46.8	56.7	52.3	51.0	49.2	48.7	48.2	49.3	51.4
Video-CCAM-7B (Fei et al., 2024)	55.6	40.9	60.0	52.7	52.3	54.8	47.0	57.1	53.0	51.0	51.0	48.3	47.8	49.5	51.5
VideoLLaMA3-7B (Zhang et al., 2025)	69.7	63.7	67.3	74.0	68.7	66.0	58.4	61.1	61.8	64.6	63.1	64.3	64.1	64.0	65.1
Tarsier2-7B (Yuan et al., 2025)	64.6	56.8	59.7	**75.4**	64.1	66.0	64.0	63.8	64.6	57.7	63.4	59.5	60.5	60.3	62.8
Qwen2.5VL-7B (Bai et al., 2025)	**71.3**	61.4	**72.3**	**75.4**	**70.1**	**79.8**	**73.3**	68.1	**73.7**	**77.2**	**69.7**	**71.0**	**73.7**	**72.9**	**72.1**
GPT-4o (16 frames)	64.6	**64.0**	68.6	73.2	67.6	66.5	60.8	**68.8**	65.4	61.3	62.0	61.0	59.8	61.0	64.6

Table 5: Multiple domain generalization results on multiple-choice QA under different domain shifts.

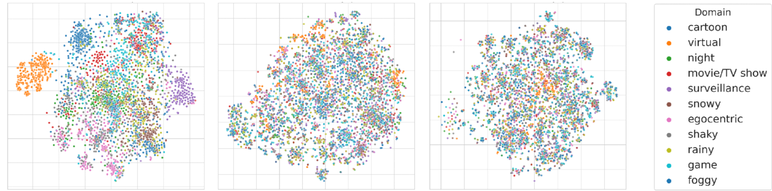
Model	Visual Style					Viewpoint				Env. Condition					D-Avg
	CA	GA	MO	VI	Avg^m	EG	SU	SH	Avg^m	FO	NI	RA	SN	Avg^m	
HBI (Jin et al., 2023)	14.9	18.2	17.2	16.4	16.7	17.4	16.7	18.5	17.5	17.7	18.9	16.9	17.6	17.8	17.3
EMCL4QA (Jin et al., 2022)	17.7	18.7	16.8	17.7	17.7	17.4	16.7	19.6	17.9	19.1	18.4	18.8	18.3	18.7	18.1
Qwen2.5VL-3B (Bai et al., 2025)	**70.5**	60.7	62.0	66.2	65.0	65.6	61.2	57.7	61.5	61.0	55.9	59.5	55.5	58.0	61.4
VideoLLaMA2-7B (Cheng et al., 2024)	61.6	**64.4**	**68.7**	**70.1**	**66.5**	**66.1**	**62.9**	**69.6**	**66.2**	**69.6**	**69.1**	**64.9**	**69.2**	**68.2**	**66.9**

Table 4: Single domain generalization results on multiple-choice QA under different domain shifts.

Model	Visual Style					Viewpoint				Env. Condition					D-Avg
	CA	GA	MO	VI	Avg^s	EG	SU	SH	Avg^s	FO	NI	RA	SN	Avg^s	
EMCL4QA (Jin et al., 2022)	18.4	16.7	18.5	17.9	17.9	19.0	17.4	16.4	17.6	18.8	18.2	17.1	16.8	17.7	17.7
HBI (Jin et al., 2023)	18.9	16.9	17.9	18.2	18.0	18.4	16.9	16.8	17.4	17.9	19.3	18.6	16.8	18.2	17.9
VideoLLaMA2-7B (Cheng et al., 2024)	53.2	48.7	44.9	47.8	48.6	52.4	54.4	52.5	53.1	**63.1**	56.6	**60.4**	53.0	**58.3**	53.4
Qwen2.5VL-3B (Bai et al., 2025)	**63.3**	**66.5**	**66.1**	**64.6**	**65.1**	**59.8**	**61.7**	**63.4**	**61.6**	57.2	**58.4**	57.3	**58.4**	57.8	**61.5**

We provide the multiple-domain and single-domain generalization results on multiple-chocie QA, displaying that Qwen2.5VL shows relatively stronger generalization ability than other models.

We also present the t-SNE visualization of visual frame features, question embeddings and GT answer embeddings, demonstrating the cross domain visual differences and semantic consistency of our dataset.



(a) CLIP frame features (b) Question embeddings (c) GT answer embeddings

CONCLUSTION

Core Contribution: We propose VUDG, a new video DG benchmark with semantically aligned content across 11 diverse domains for rigorous fair cross-domain evaluation, built with a progressive multi-expert annotation pipeline (multi-LLM collaboration + human expert refinement) to generate high-quality DG-tailored QA pairs.

Key Findings: Existing models struggle to mitigate domain shifts for downstream tasks; zero-shot LVLM performance varies drastically across models, with even SOTA LVLMs exhibiting unstable cross-domain performance.

Outlook: VUDG serves as a valuable foundation for generalizable video understanding research. We will extend it to cover textual domain shifts and audio modalities to support comprehensive multimodal DG evaluation.