中文信息处理实验室-让机器理解语言
Chinese Information Processing Laboratory

快手

ISCAS

ICML 2026

MetaphorVU: Towards Metaphorical Video Understanding

Zhuoqun Li, Boxi Cao, Guiping Jiang, Fangrui Lv, Ruotong Pan, Jianan Wang, Xiangyu Wu, Hongyu Lin, Yaojie Lu, Yong Du, Ruyin Jia, Liyan, Tingting Gao, Han Li, Xianpei Han, Le Sun

Zhuoqun Li

2026.5.25

Background

- **Metaphorical videos serve as a crucial medium**
  - creators employ metaphorical content to guide viewers toward associations and interpretations
  - e.g., society criticism, life contemplation, Situation Portrayal
- **MetaphorVU is a high-order cognitive process**
  - transform perceived signals into deeper semantics, requiring of linking visual elements to underlying concepts

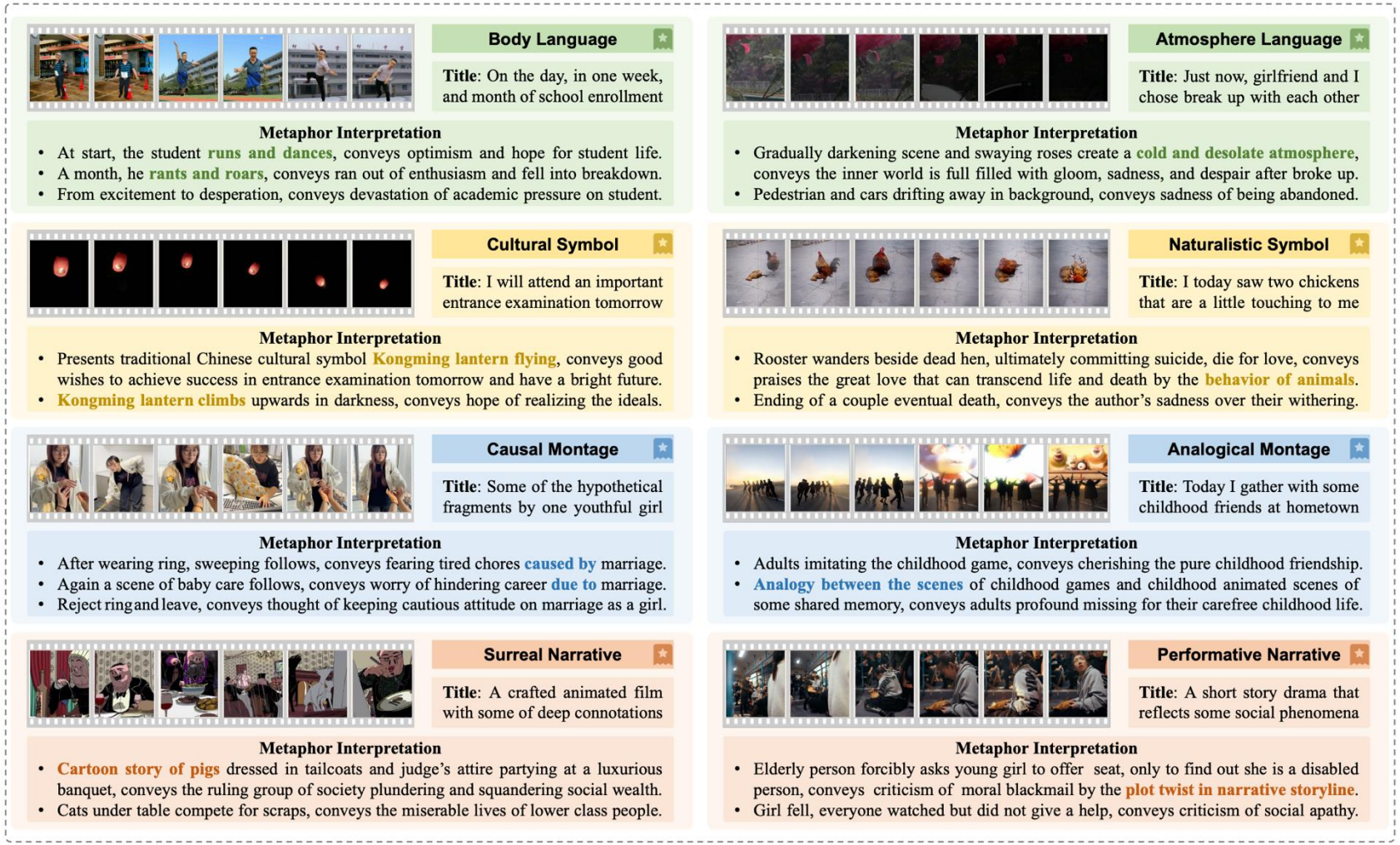
Motivation

- **Still lacking systematic study of MetaphorVU**
  - previous works mainly focus on object recognition and event description
  - limiting MLLMs application in complex scenarios and cognitive capability improvement

- This paper first construct **MetaphorVU-Bench**, then perform evaluation and analysis based on the benchmark, and finally propose **MetaphorBoost** to enhance MLLMs

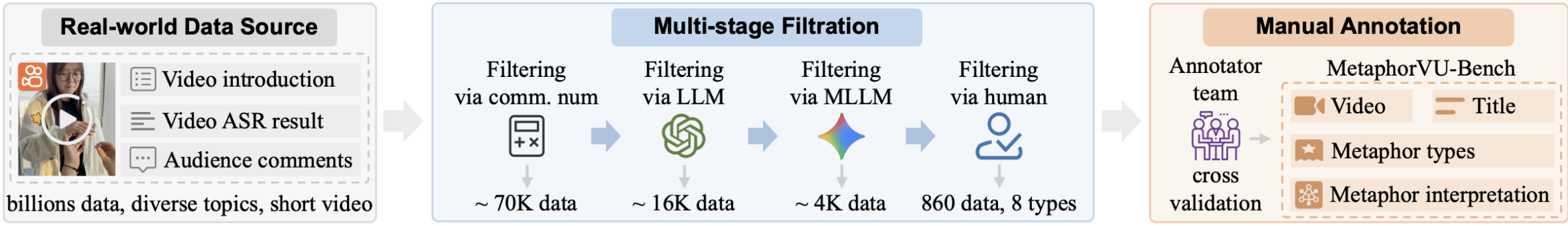
MetaphorVU-Bench: Video Metaphor Taxonomy

- **8 types of video metaphor** based on multimodal metaphor theory



MetaphorVU-Bench: Construction Pipeline

- **Real-world Data Source:** Kuaishou short-video platform
- **Efficient Multi-stage Filtration:** isolate the metaphorical videos from billions of raw videos in the data source
- **Reliable Manual Annotation:** human team perform cross validation

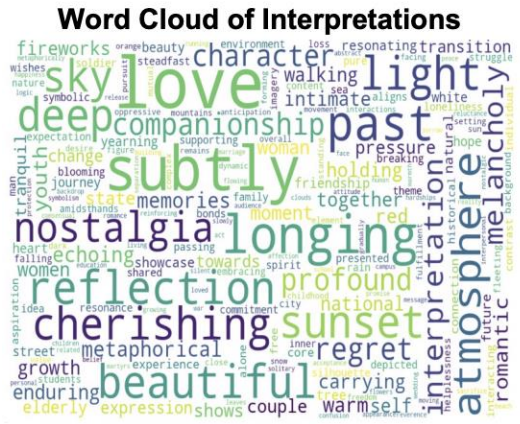


MetaphorVU-Bench: Statistic Information

- The first **systematic and comprehensive** MetaphorVU benchmark

Type	# Samples	Avg. Duration (s)	Avg. Tokens
Body Language (Body L.)	136	32.2	111.3
Atmosphere Language (Atmosp. L.)	150	13.1	104.5
Cultural Symbol (Cultural S.)	62	23.5	114.4
Naturalistic Symbol (Natural. S.)	113	17.3	108.8
Causal Montage (Causal M.)	54	57.7	108.9
Analogical Montage (Analog. M.)	171	58.7	124.8
Surreal Narrative (Surreal N.)	112	30.4	117.1
Performative Narrative (Perform. N.)	62	86.8	118.6
MetaphorVU-Bench	860	37.2	114.2

Based on above video metaphor taxonomy, **benchmark covers 8 video metaphor types, enabling the systematic evaluation**



MetaphorVU-Bench covers various of video topics, **enabling comprehensive evaluation of real-world metaphorical video understanding**

MetaphorVU-Bench: Evaluation Results

- MLLMs **struggle with** accurate metaphorical video understanding

Method	Body L.	Atmosph. L.	Cultural S.	Natural. S.	Causal M.	Analog. M.	Surreal N.	Perform. N.	Average
Upper-bound									
Human*	87.8	87.5	89.1	83.8	72.0	81.5	78.1	78.0	83.4
Close-source MLLMs									
GPT-5 (OpenAI, 2025)	69.9	**76.3**	77.4	66.6	45.0	55.4	54.9	46.1	63.7
GPT-4o (OpenAI, 2024)	63.4	70.5	70.3	62.6	39.1	48.2	45.7	37.9	56.8
Qwen3-VL-Plus (Bai et al., 2025a)	66.8	72.5	74.8	65.5	51.5	54.2	50.4	43.7	61.4
Gemini-2.5-Pro (Google, 2025a)	65.5	71.3	74.3	64.4	53.5	55.7	52.1	46.9	61.8
Gemini-3-Pro (Google, 2025b)	71.2	74.0	75.1	**66.9**	49.4	58.9	51.1	48.1	63.8
Doubao-1.5-Vision-Pro (Guo et al., 2025)	58.2	64.1	65.5	58.9	27.8	42.5	39.8	26.6	50.5
Open-source MLLMs									
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	36.0	49.9	46.1	42.1	12.4	23.5	28.6	16.1	33.8
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	56.0	66.1	68.8	60.8	33.2	45.0	39.3	29.2	52.0
LLaVA-onevision-1.5-8B-Instruct (An et al., 2025)	35.7	47.2	47.3	45.0	13.8	21.3	27.0	21.2	38.1
GLM-4.5V (Team et al., 2025)	62.7	67.9	71.9	62.1	37.6	50.1	46.1	38.4	56.8
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	65.4	70.4	71.9	58.1	43.2	54.6	46.1	38.1	58.6
Reasoning-enhanced Methods									
VideoRFT (Wang et al., 2025)	38.9	52.8	48.4	46.0	13.5	24.8	27.2	16.6	35.6
Vision-R1 (Huang et al., 2025)	39.3	45.1	42.0	42.4	19.4	23.2	25.0	18.6	33.1
ReAd-R (Long et al., 2025)	42.1	54.1	48.9	46.3	15.7	26.4	26.2	17.6	36.8
LTR (Liao et al., 2025)	54.1	44.7	56.2	47.4	27.8	44.6	31.9	36.1	44.5
ViTCoT (Zhang et al., 2025a)	58.8	47.7	59.2	48.7	26.1	45.1	34.0	32.1	46.2
Prompt Engineering (Wei et al., 2022)	57.8	66.3	67.9	59.2	36.1	42.7	41.6	32.6	52.4
Few-shot Example (Dong et al., 2024)	57.6	69.4	69.2	58.7	33.5	44.9	43.5	32.6	53.6

Human

Powerful MLLMs

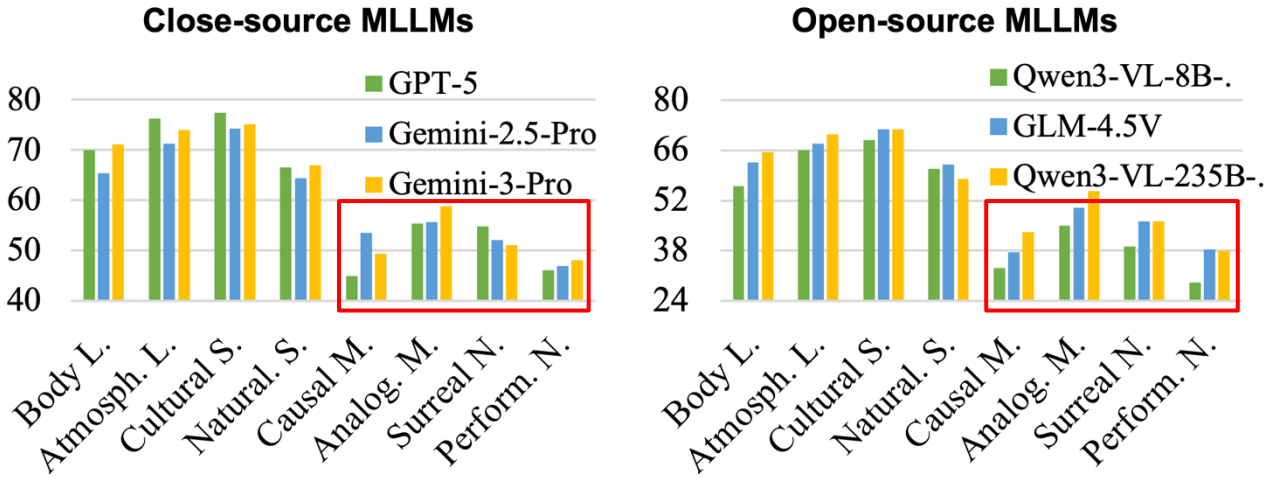
MetaphorVU-Bench: Ability Deficiency Analysis

- Main deficiency largely lies in **poor cross-domain mapping ability**,

  MLLMs fail to link visual elements to underlying concepts

Model	Wrong Recognition	Missing Mapping	Superficial Mapping	Improper Mapping
Gemini-3-Pro	10.7%	27.9%	33.7%	27.7%
Qwen3-VL-8B-Thinking	13.5%	28.1%	28.3%	30.1%

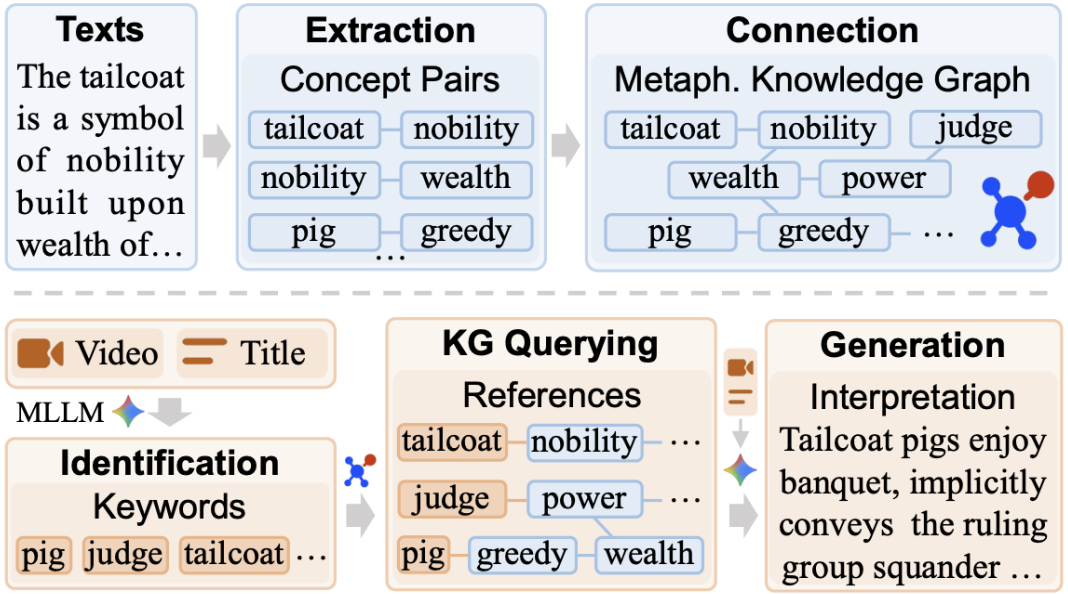
Wrong recognition is only a small proportion, **majority is missing, superficial and improper cross-domain mapping**



Across different metaphor types, **MLLMs perform worse when requiring more cross-domain mapping**

MetaphorBoost: Method

- Construct a **metaphorical knowledge graph** as external scaffold
- Improves MLLMs via **inference-time mapping augmentation** based on the constructed metaphorical knowledge graph



MetaphorBoost: Effectiveness

- **Consistently improve MLLMs** on metaphorical video understanding

Method	Body L.	Atmosph. L.	Cultural S.	Natural. S.	Causal M.	Analog. M.	Surreal N.	Perform. N.	Average
Upper-bound									
Human*	87.8	87.5	89.1	83.8	72.0	81.5	78.1	78.0	83.4
Close-source MLLMs									
GPT-5 (OpenAI, 2025)	69.9	**76.3**	77.4	66.6	45.0	55.4	54.9	46.1	63.7
GPT-4o (OpenAI, 2024)	63.4	70.5	70.3	62.6	39.1	48.2	45.7	37.9	56.8
Qwen3-VL-Plus (Bai et al., 2025a)	66.8	72.5	74.8	65.5	51.5	54.2	50.4	43.7	61.4
Gemini-2.5-Pro (Google, 2025a)	65.5	71.3	74.3	64.4	53.5	55.7	52.1	46.9	61.8
Gemini-3-Pro (Google, 2025b)	71.2	74.0	75.1	**66.9**	49.4	58.9	51.1	48.1	63.8
Doubao-1.5-Vision-Pro (Guo et al., 2025)	58.2	64.1	65.5	58.9	27.8	42.5	39.8	26.6	50.5
Open-source MLLMs									
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	36.0	49.9	46.1	42.1	12.4	23.5	28.6	16.1	33.8
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	56.0	66.1	68.8	60.8	33.2	45.0	39.3	29.2	52.0
LLaVA-onevision-1.5-8B-Instruct (An et al., 2025)	35.7	47.2	47.3	45.0	13.8	21.3	27.0	21.2	38.1
GLM-4.5V (Team et al., 2025)	62.7	67.9	71.9	62.1	37.6	50.1	46.1	38.4	56.8
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	65.4	70.4	71.9	58.1	43.2	54.6	46.1	38.1	58.6
Mapping Augmentation via Metaphorical Knowledge Graph									
MetaphorBoost (Gemini-3-Pro) (Ours)	**71.5**	**76.3**	**77.5**	**66.9**	**57.2**	**59.1**	**57.3**	**50.8**	**66.1**
Δ (vs Gemini-3-Pro)	+0.3	+2.3	+2.4	+0.0	+7.8	+0.2	+6.2	+2.8	+2.3
MetaphorBoost (Qwen2.5-VL-7B-Instruct) (Ours)	40.7	55.7	51.2	49.0	12.5	26.1	31.4	19.2	37.9
Δ (vs Qwen2.5-VL-7B-Instruct)	+4.6	+5.8	+5.1	+6.9	+0.1	+2.6	+2.9	+3.0	+4.1
MetaphorBoost (Qwen3-VL-8B-Thinking) (Ours)	61.8	71.0	71.8	61.3	36.7	47.1	45.7	31.5	55.9
Δ (vs Qwen3-VL-8B-Thinking)	+5.8	+4.9	+3.0	+0.5	+3.5	+2.1	+6.4	+2.3	+3.8

Gemini3

Qwen2.5

Qwen3

Summary

- We propose MetaphorVU-Bench, **the first systematic and comprehensive** benchmark dedicated to evaluation for metaphorical video understanding.
- We conduct extensive experiments and analysis, revealing the deficiencies of current MLLMs and **providing insights into the cross-domain mapping**.
- We construct MetaphorBoost, boosting MetaphorVU via inference-time **mapping augmentation based on a metaphorical knowledge graph**.

- **Paper**: https://openreview.net/forum?id=yKcBAJMPXZ
- **Code and Dataset**: https://github.com/Li-Z-Q/MetaphorVU

Author CV

Thanks