X-CoT: Explainable Text-to-Video Retrieval via LLM-based Chain-of-Thought Reasoning

EMNLP 2025 Suzhou, China | 中国苏州

Prasanna Reddy Pulakurthi♠, Jiamian Wang♠, Majid Rabbani♠, Sohail Dianat♠, Raghuveer Rao♦, Zhiqiang Tao♠

♠Rochester Institute of Technology ♦DEVCOM Army Research Laboratory

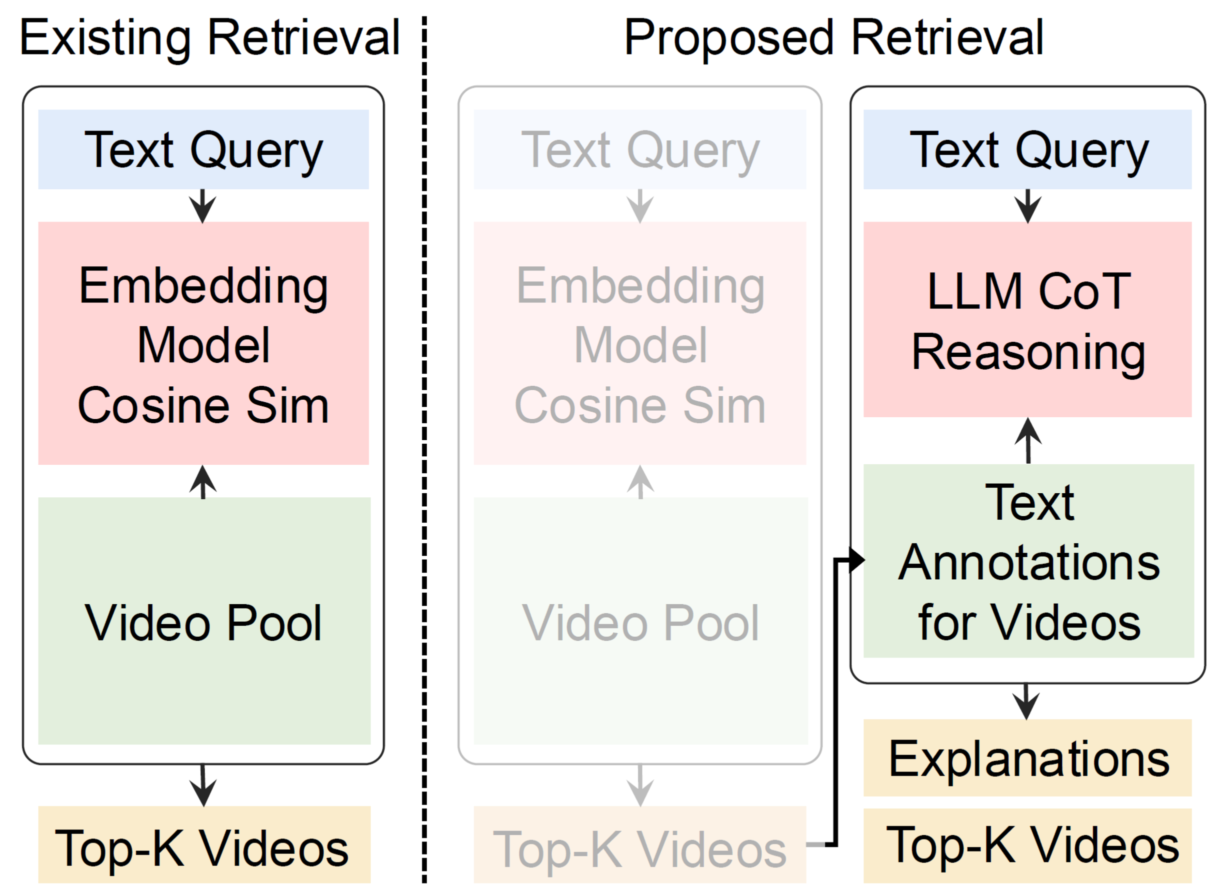
Motivation & Background

Motivation:

- Existing text-to-video retrieval relies on embedding models and cosine similarity, which lack interpretability and are sensitive to low-quality text-video pairs.

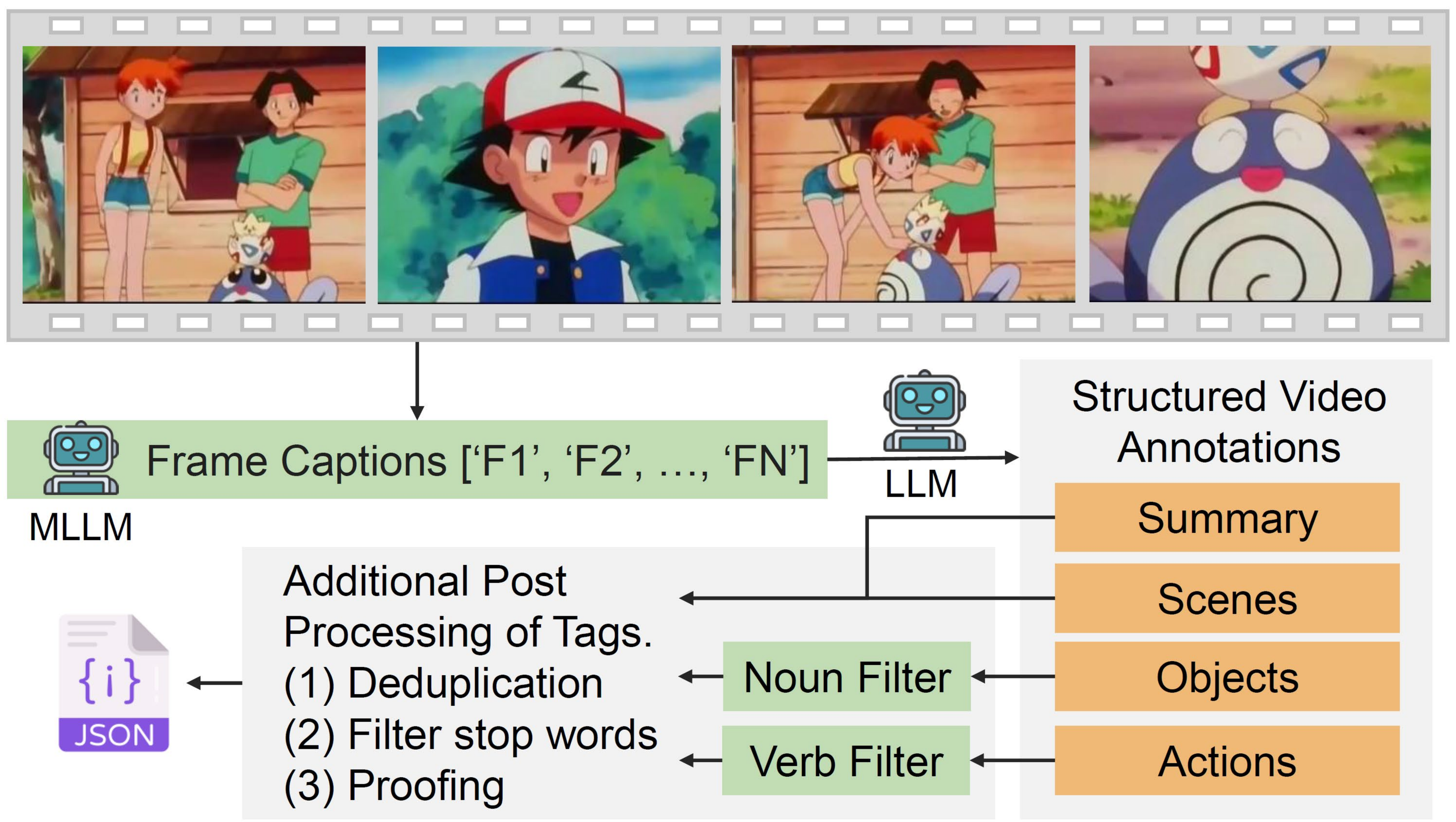
Contributions:

- Introduces X-CoT, an explainable retrieval framework using LLM Chain-of-Thought reasoning, advancing trustworthy and trackable retrieval.
- Expands benchmarks with structured video annotations (objects, actions, scenes, summaries, frame captions) for richer semantics.
- Employs pairwise LLM comparisons with Bradley-Terry aggregation to produce both rankings and natural-language explanations.
- Achieves consistent performance gains across all benchmarks (e.g., +5.6 % R@1 on MSVD) while enabling model and data quality analysis.



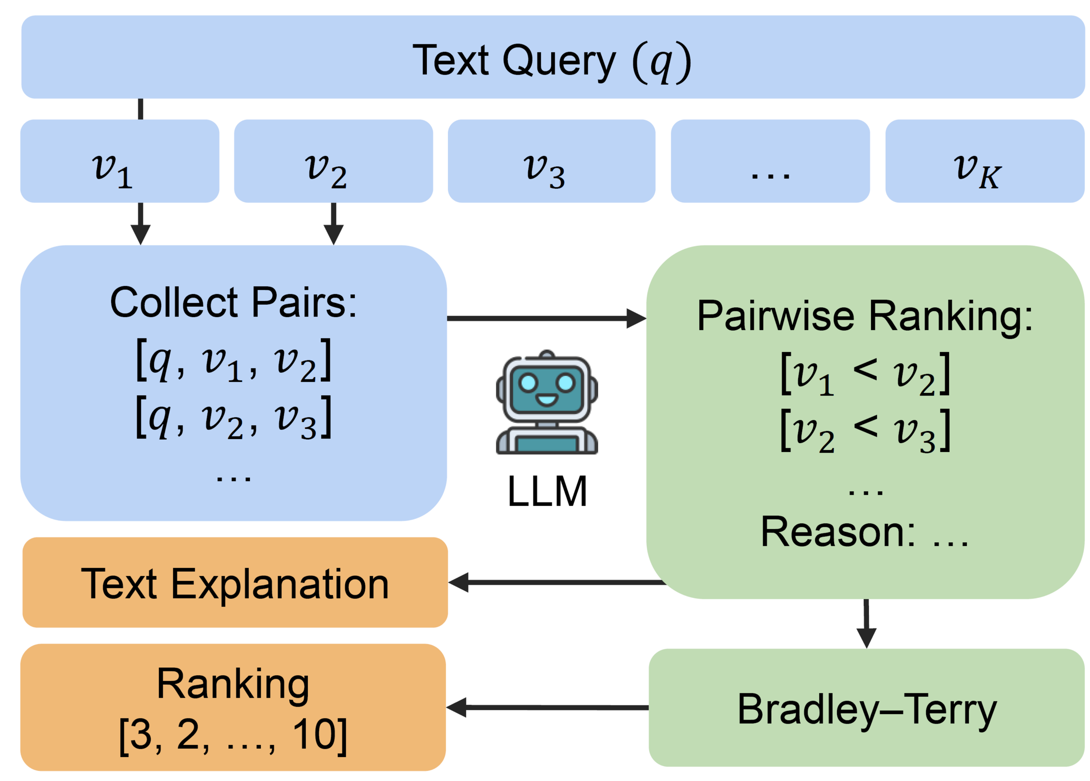
Method: Structured Video Annotations

- Generates frame-level captions using an MLLM to describe each sampled frame with fine-grained visual details.
- Uses an LLM to produce structured annotations containing objects, actions, scenes, and summaries for richer semantics.



Method: X-CoT Framework

1. Use an embedding model (e.g., CLIP, X-Pool) to obtain a top-K video pool ($\mathcal{V} = \{v_1, v_2 \dots v_k\}$) for a text query ($q$).
2. Perform pairwise LLM comparisons between candidate videos using structured annotations.
3. Aggregate results via the Bradley-Terry model to obtain a refined ranking and natural-language explanation.



Results & Explainability

- **Retrieval Gains: +5.6%** R@1 gain (CLIP, MSVD) and **+1.9 %** R@1 (X-Pool, MSVD); consistent improvements across MSR-VTT, MSVD, DiDeMo, and LSMDC.
- **Interpretability:** Generates **natural-language rationales** explaining why one video outranks another.

Methods	MSR-VTT					MSVD				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP (Radford et al., 2021)	31.6	53.8	63.4	4.0	39.0	36.5	64.0	73.9	3.0	20.8
X-CoT (ours)	**33.7**	**56.7**	**64.6**	**4.0**	**38.7**	**42.1**	**67.4**	**75.4**	**2.0**	**20.5**
VLM2Vec (Jiang et al., 2024)	36.4	60.2	70.7	3.0	27.3	46.7	73.8	82.6	2.0	12.8
X-CoT (ours)	**37.2**	**61.8**	**71.5**	**3.0**	**27.1**	**48.4**	**74.8**	**83.2**	**2.0**	**12.6**
X-Pool (Gorti et al., 2022)	46.9	73.0	82.0	2.0	14.2	47.2	77.2	86.0	2.0	9.3
X-CoT (ours)	**47.3**	**73.3**	**82.1**	**2.0**	**14.2**	**49.1**	**78.0**	**86.6**	**2.0**	**9.2**

Table 1: Text-to-video retrieval performance comparison on MSR-VTT and MSVD.

Methods	DiDeMo					LSMDC				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP (Radford et al., 2021)	25.2	49.4	59.0	6.0	49.7	15.9	28.4	35.3	31.0	129.6
X-CoT (ours)	**29.7**	**52.1**	**60.6**	**5.0**	**49.2**	**17.6**	**29.0**	**36.1**	**31.0**	**129.4**
VLM2Vec (Jiang et al., 2024)	33.5	57.7	68.4	4.0	34.1	18.2	33.6	41.4	23.0	119.1
X-CoT (ours)	**35.8**	**59.2**	**68.8**	**3.0**	**33.9**	**18.9**	**35.1**	**41.9**	**23.0**	**118.9**
X-Pool (Gorti et al., 2022)	44.6	72.5	81.0	2.0	15.1	23.6	42.9	52.4	9.0	54.1
X-CoT (ours)	**45.1**	**73.1**	**81.8**	**2.0**	**15.0**	**23.8**	**43.8**	**53.1**	**8.0**	**54.0**

Table 2: Text-to-video retrieval performance comparison on DiDeMo and LSMDC.

- **Insight & Diagnosis:** Enables identification of noisy captions or ambiguous text-video pairs and reveals the **retrieval model's behaviors** (e.g., semantic focus, missed concepts).

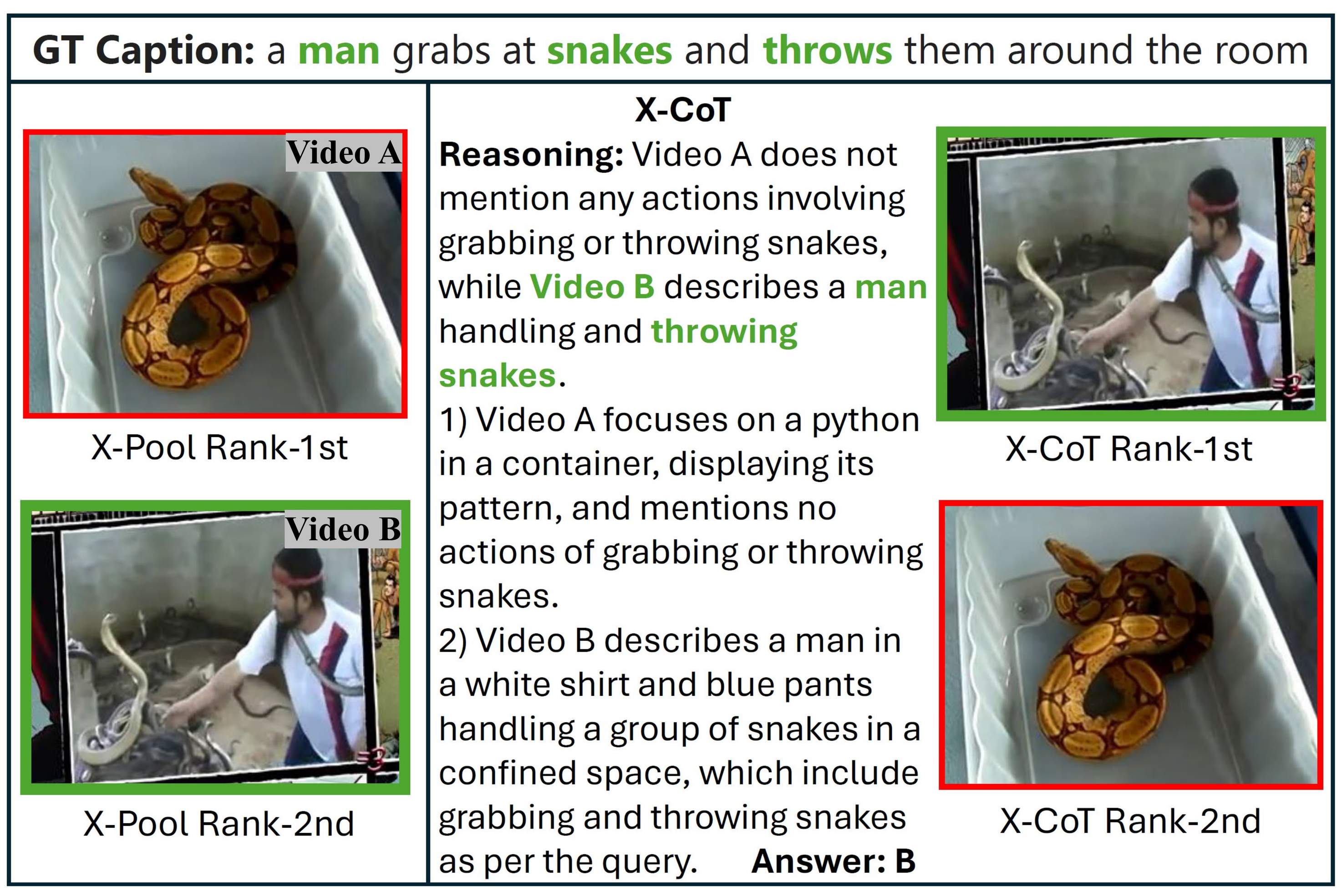


Fig 1. X-CoT provides human-readable explanations for ranking decisions.

Conclusion

X-CoT introduces explainable text-to-video retrieval by integrating Chain-of-Thought LLM reasoning into refined ranking, achieving consistent performance gains and generating natural-language explanations for transparent, trustworthy, and analyzable retrieval systems.

Acknowledgments: Supported by DEVCOM ARL (W911QX-21-D-0001), NSF (2502050), and NIH (R16GM159146).