Blocking Visual Leakage:

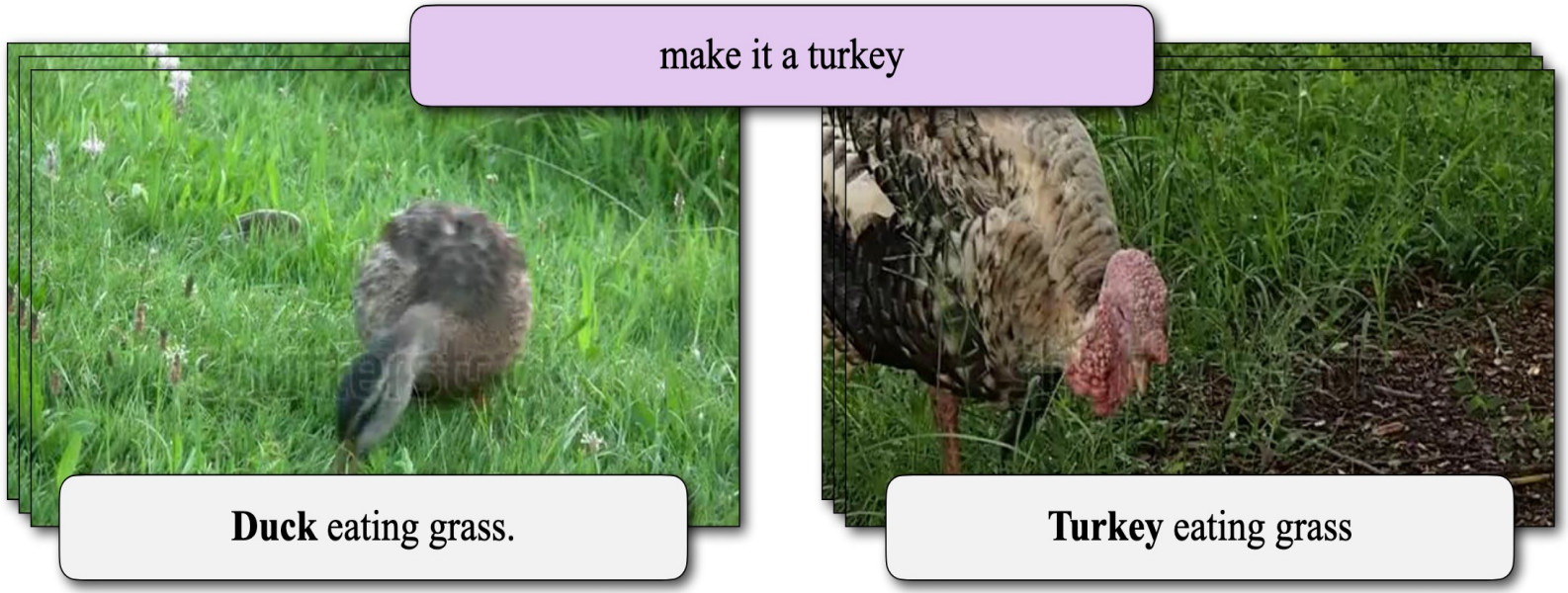
Visually-Agnostic Text Decomposition for Composed Video Retrieval

Jinkwon Hwang, Trung X. Pham, Junyeong Kim

Chung-Ang University · Van Lang University

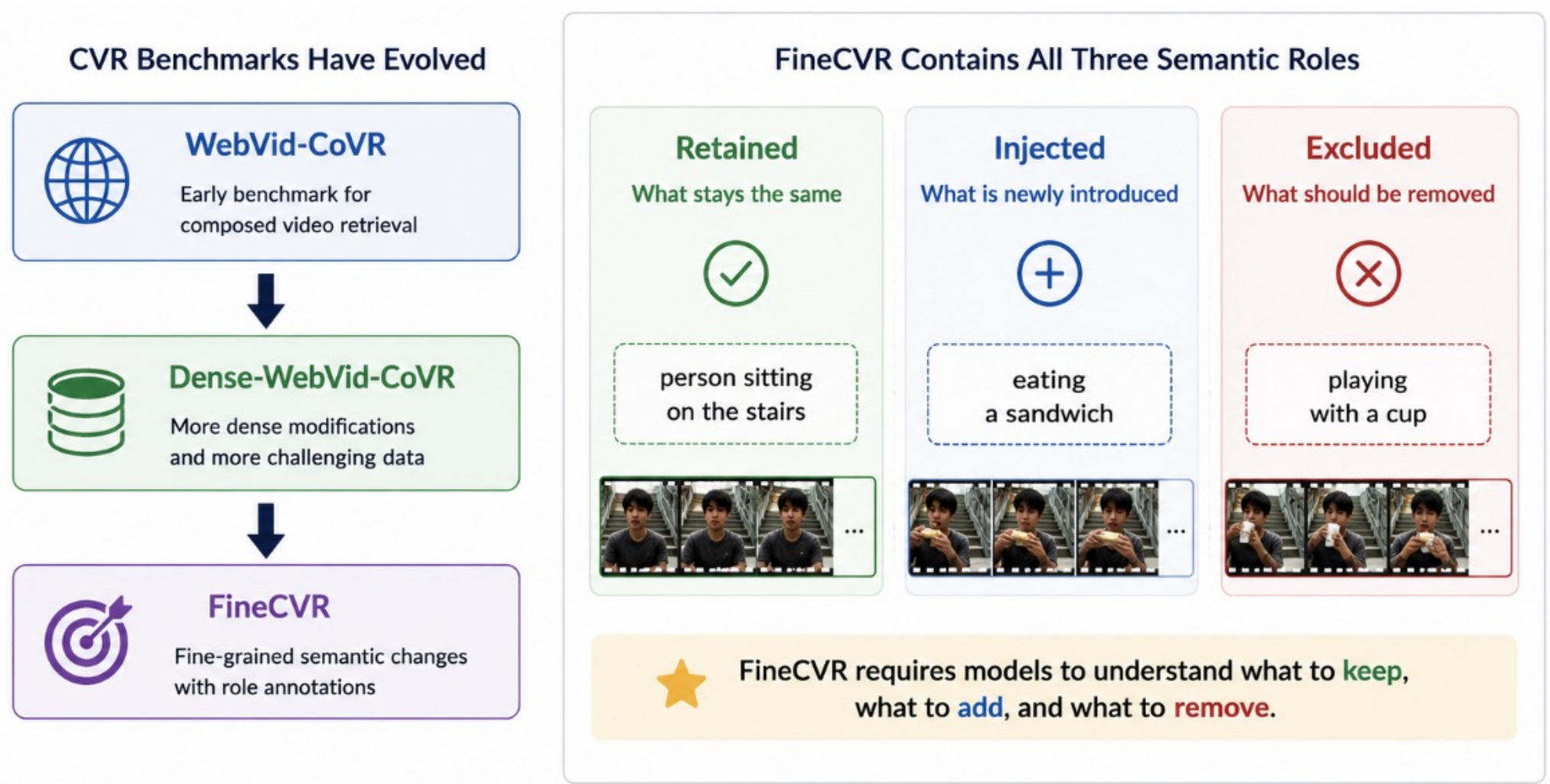
Index

Introduction



Triplet Data: **Reference Video + Modification Text → Target Video**

Evolution of CVR

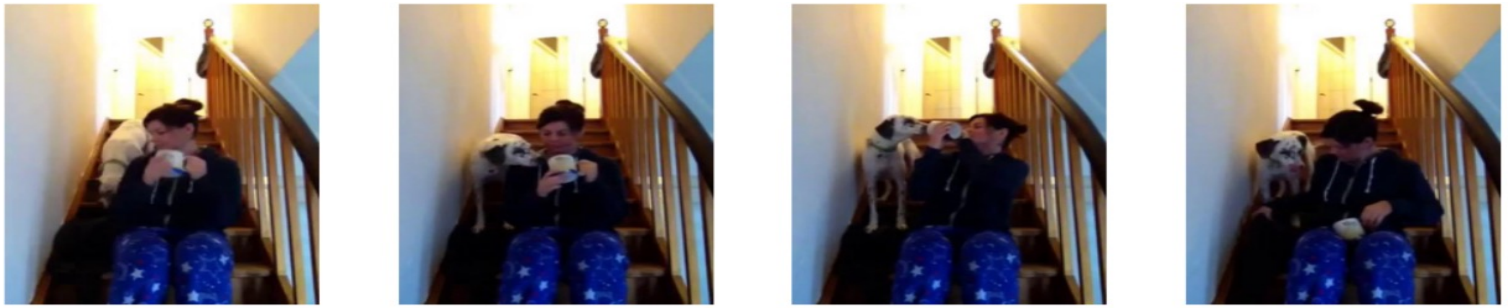


Motivation: Visual Leakage

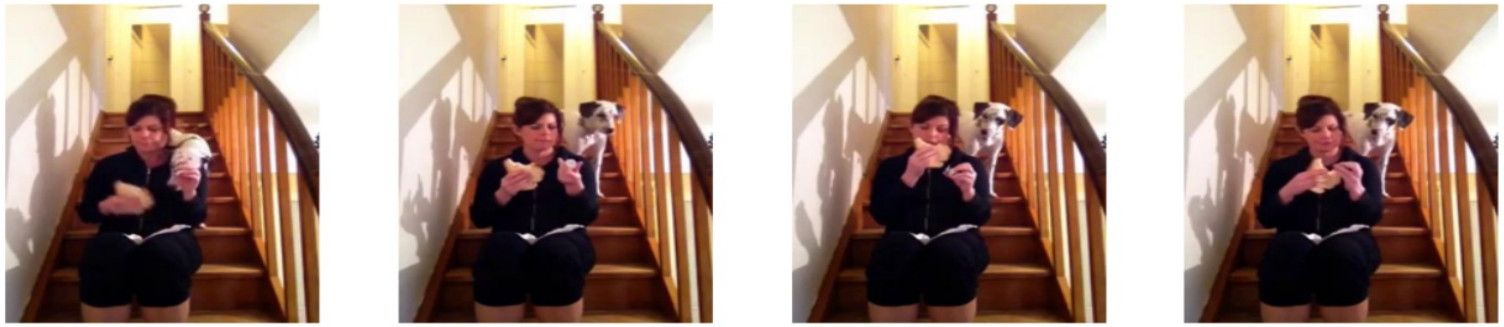
Caption: “the person is also sitting on the stairs, but they are eating a sandwich instead of playing with a cup”

Retain: the person is also sitting on the stairs	Inject: But they are eating a sandwich	Exclude: Instead of playing with a cup

Retain

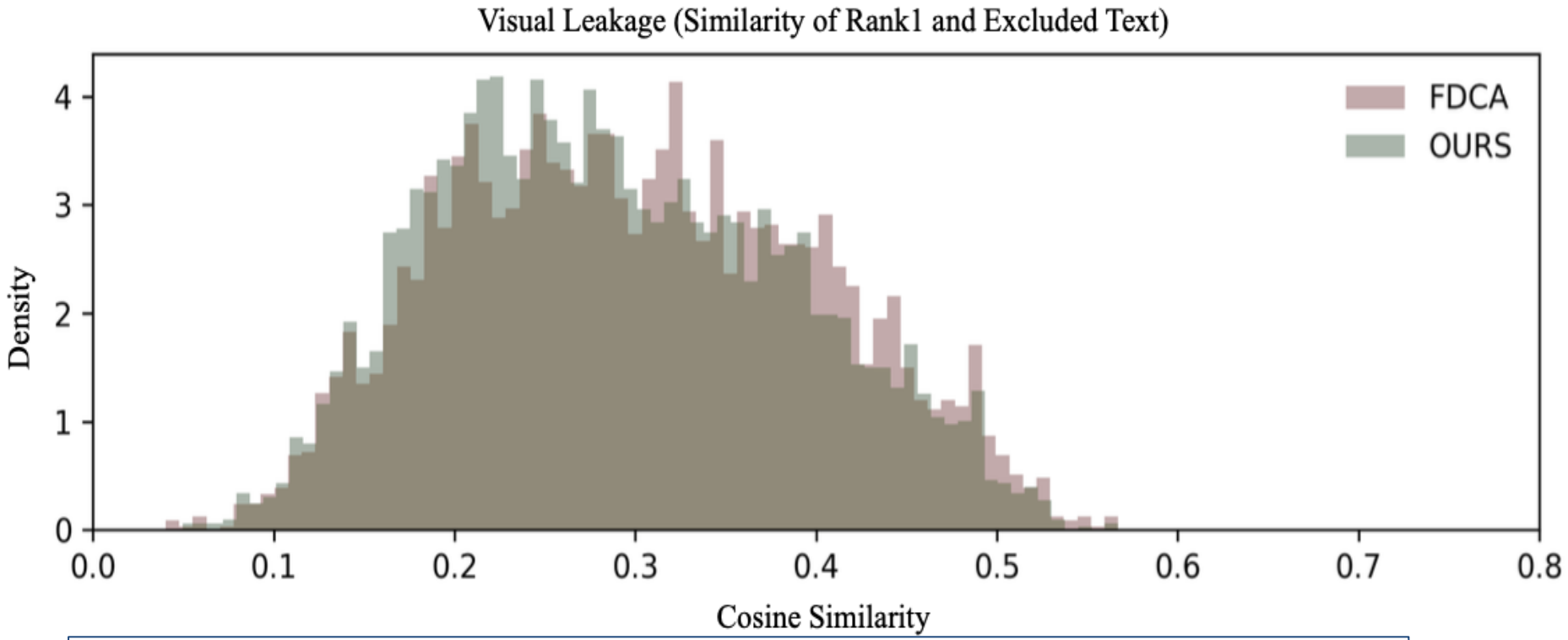


target



Motivation: Evidence

Quantitative Analysis



Existing methods remain highly similar to excluded concepts

Our method suppresses excluded concepts

Motivation: Key Idea

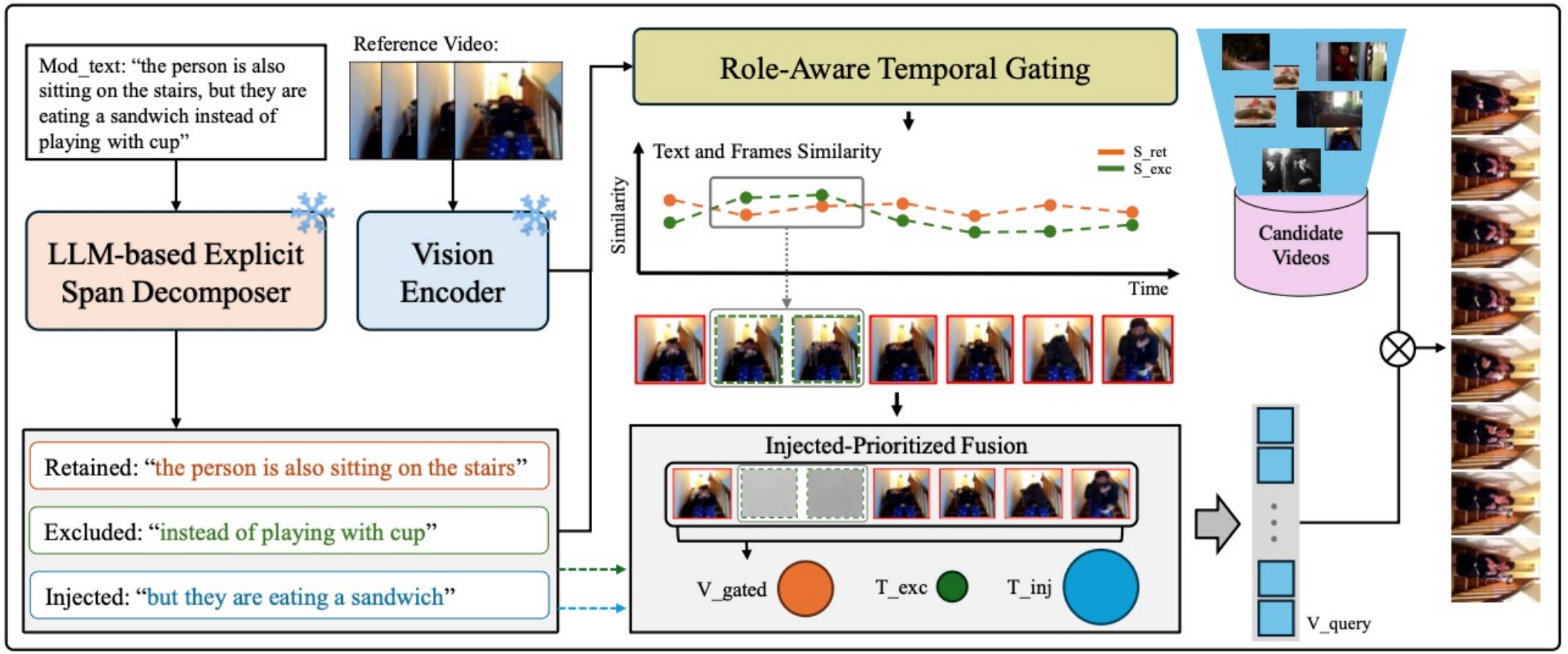
Interpret text before seeing the video

Decompose text

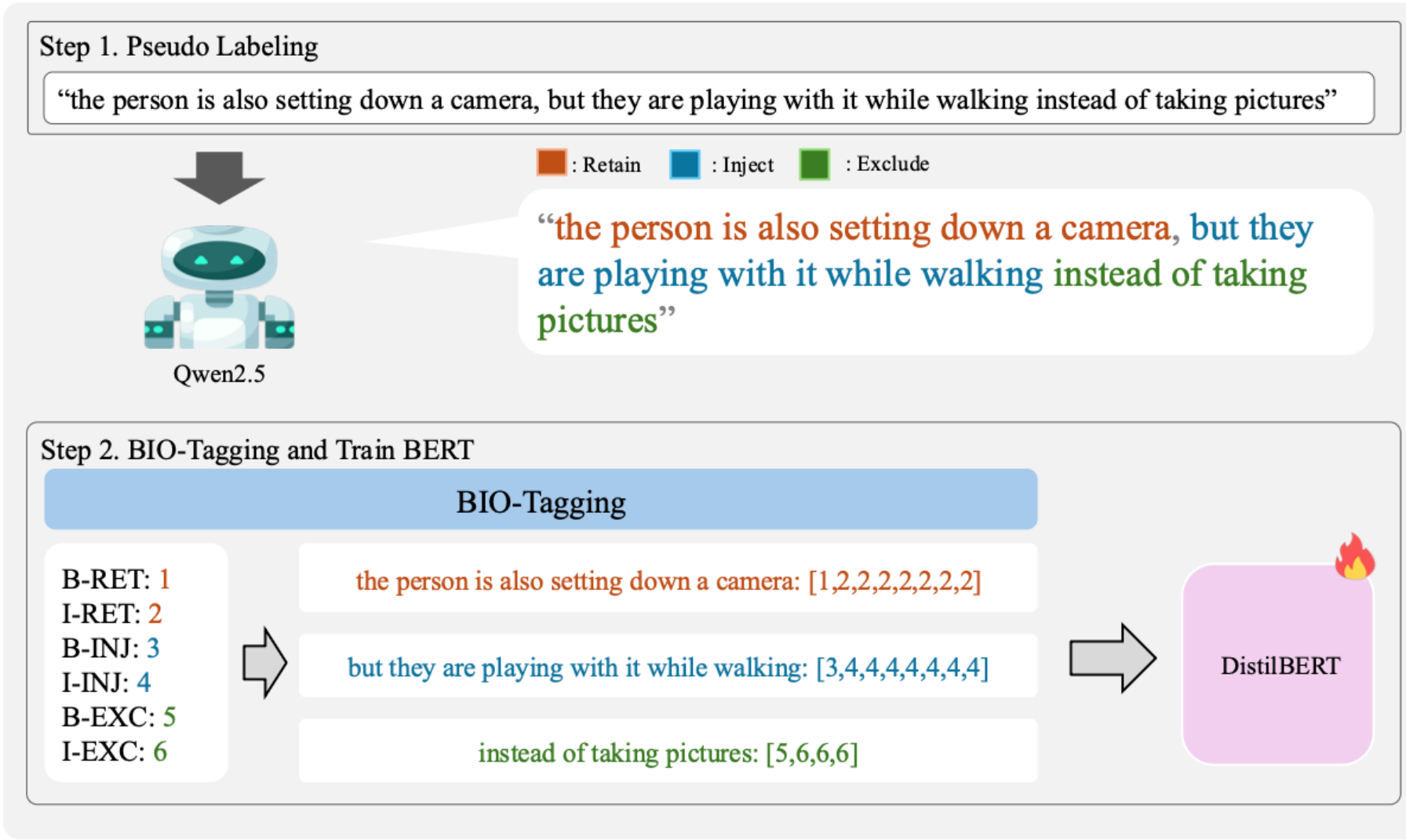
Remove excluded frames

Emphasize injected information

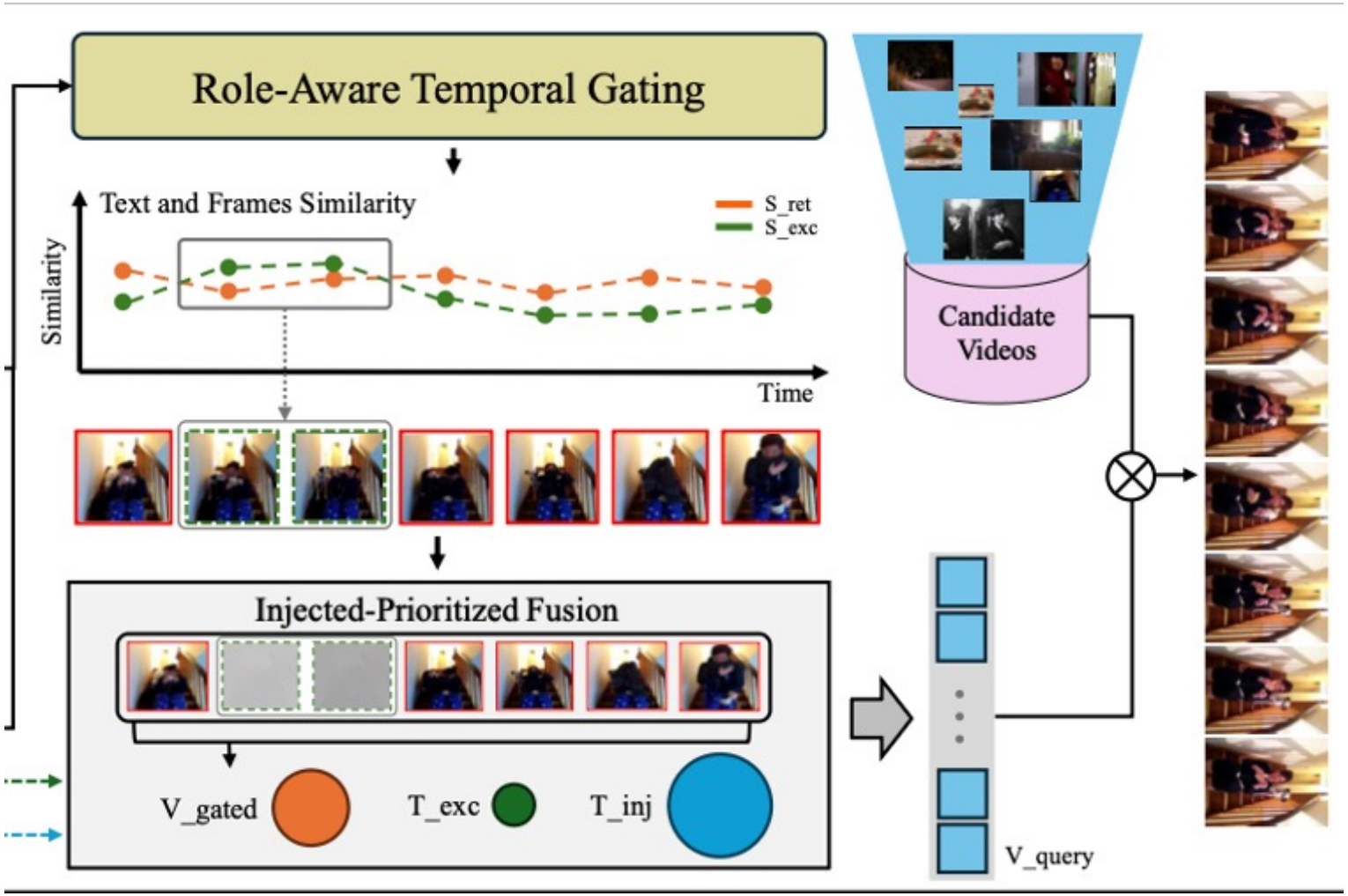
Method: Overall Framework



Method: Explicit Span Decomposer



Method: Role-Aware Temporal Gating



Experiment: Main Results on FineCVR-1M

Comparison with Strong Baseline (BLIP)

Method (BLIP Backbone)	R@1	R@5	R@10	R@50
FDCA-BLIP (Baseline)	26.79	63.21	78.65	97.25
VAD-CVR (Ours)	**29.15**	**65.84**	**80.56**	**97.01**
Improvement (vs. FDCA-BLIP)	**+2.36**	**+2.63**	**+1.91**	-0.24

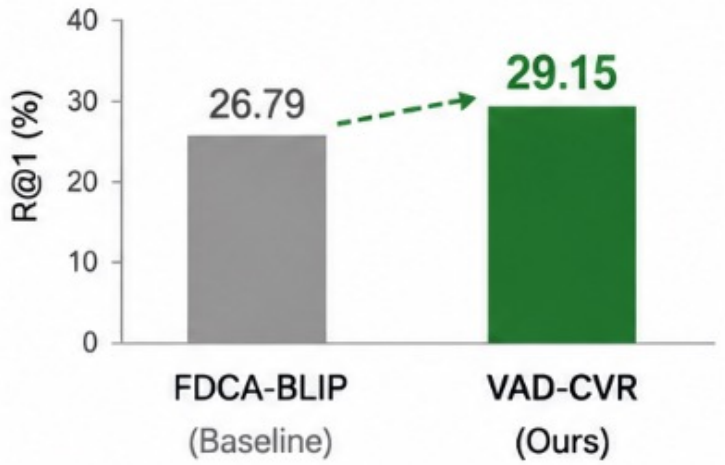
VAD-CVR improves all key metrics (R@1, R@5, R@10) and achieves the new **state-of-the-art** on FineCVR-1M.

Significant Gain at R@1

R@1 Improvement

+2.36%p

over FDCA-BLIP (Baseline)



Conclusion

1 **Visual Leakage**

Visual information biases modification-text interpretation

2 **We propose VAD-CVR**

Visually-agnostic text decomposition, Role-aware temporal gating

3 **We demonstrate**

- Effective removal of excluded visual concepts
- Better understanding of modification intent
- More accurate composed video retrieval

Takeaways

1

Visual Leakage Exists

Visual features bias the interpretation of modification text

2

Interpret Text First

Separate Retained, Injected and Excluded before visual interaction

3

Suppress, Then Retrieve

Suppress excluded visual evidence to retrieve the correct target video

Thank you!

Question & Contact: wlsrnjs905@cau.ac.kr