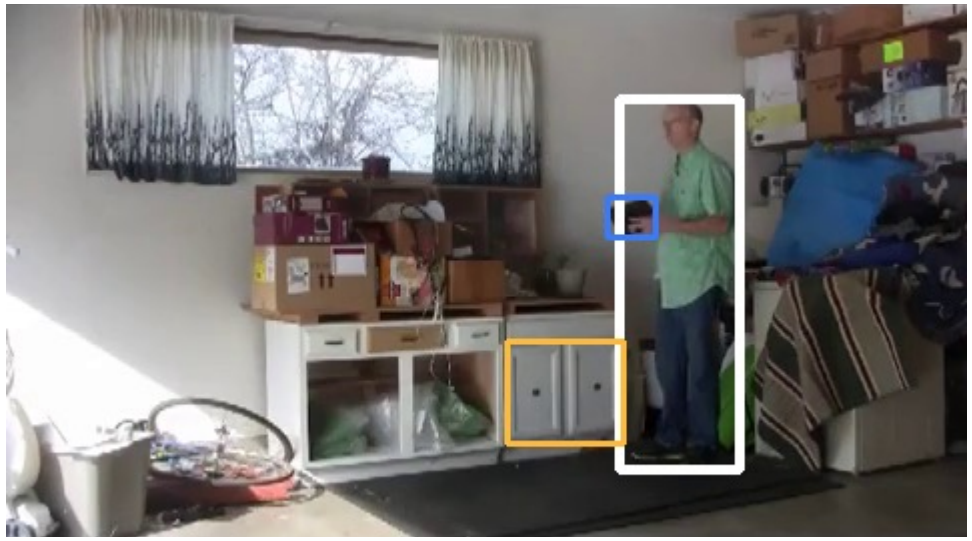
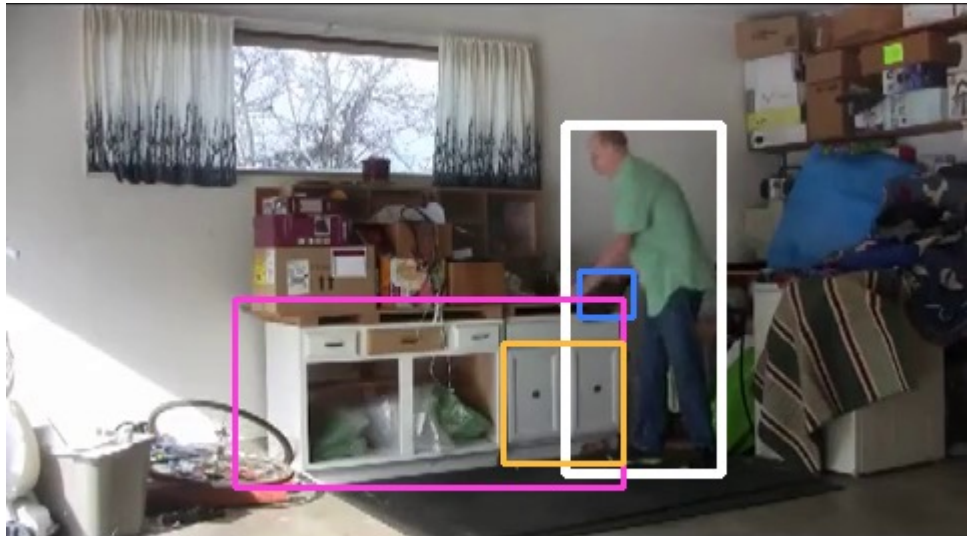
International Conference on Pattern Recognition 2026
Oral 3: Motion, Tracking And Video Analysis

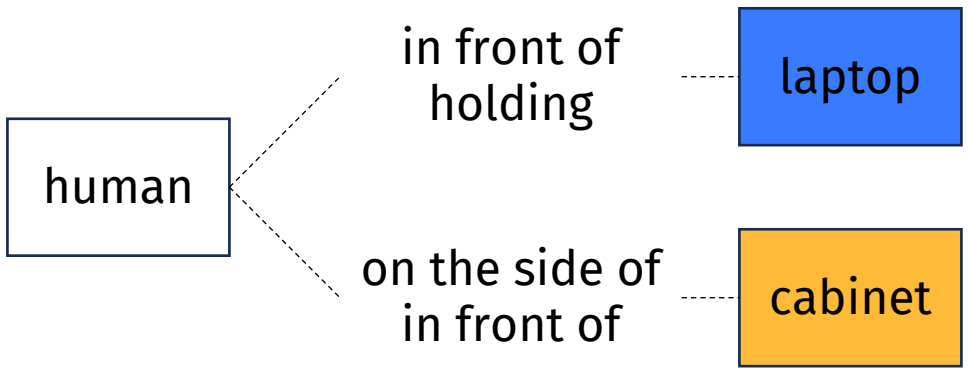
# Interaction-Centric Scene Graph Generation via Intended Interaction Target

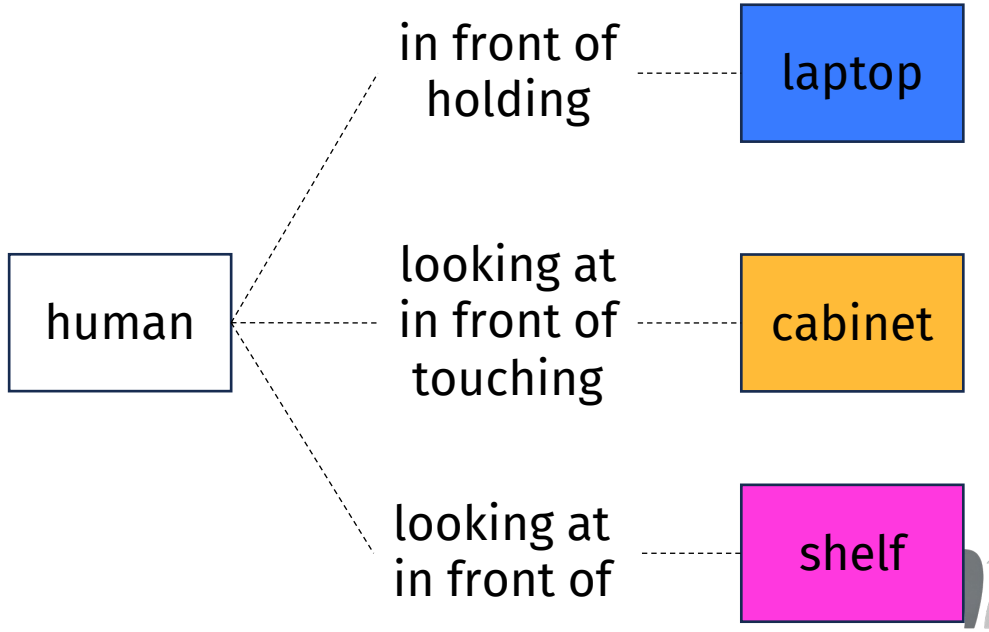
**YeEun Joo, Soon Ki Jung**
*School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Korea*

# Background: Video Scene Graph Generation







# Existing Paradigms

## Two stage

- Detect objects → construct object pairs → classify relationship

## One stage

- Directly predict ⟨object–relation–object⟩ triplets from visual features

## Common Limitation : Object-Centric Reasoning

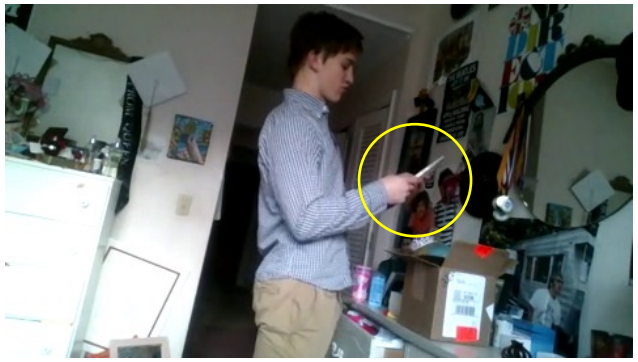
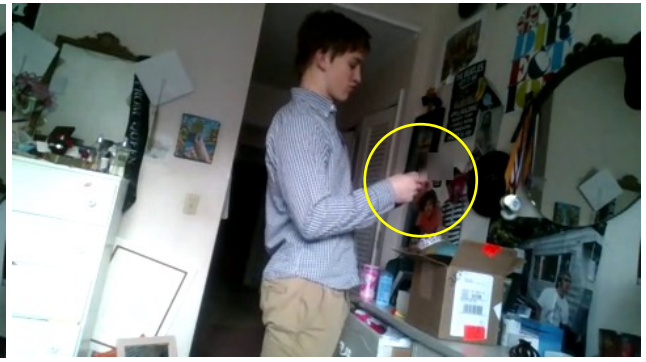
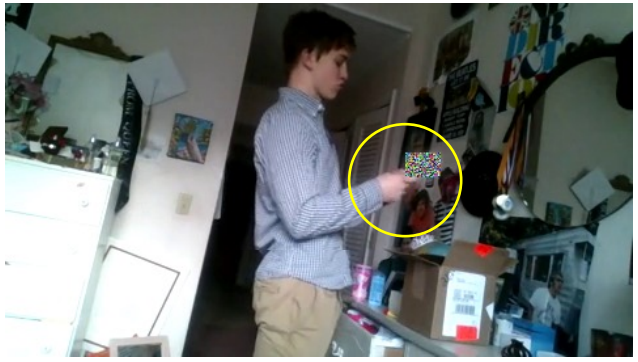
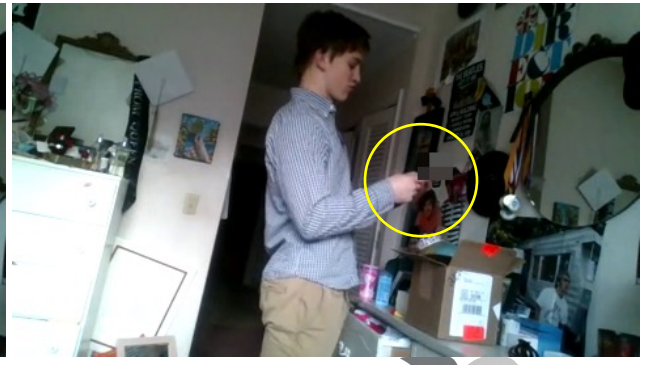
- Interaction reasoning is highly dependent on object visibility
- Performance degrades when interacting objects are poorly visible

# What Happens When Objects Become Unreliable?

## Impact of Object Appearance Degradation

- Existing VidSGG models show severe performance degradation as object appearance becomes unreliable
- Performance drops are particularly pronounced in the *SGDET* setting, which evaluates the full pipeline from object detection to relation prediction

| model | mode | Dataset | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Action Genome | | | AG-Blur | | | AG-Random noise | | | AG-Downsampling | | |
| | | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 |
| STTran [1] | predcls | 77.9 | 94.2 | 99.1 | 75.7 | 92.8 | 98.9 | 66.6 | 87.9 | 98.2 | 65.0 | 84.6 | 96.2 |
| TEMPURA [2] | predcls | 80.4 | 94.2 | 99.4 | 78.0 | 92.4 | 98.8 | 66.1 | 84.7 | 97.3 | 65.4 | 82.8 | 96.8 |
| OED [3] | predcls | 83.3 | 95.3 | 99.2 | 81.0 | 94.0 | 99.1 | 74.5 | 89.8 | 97.9 | 75.2 | 90.3 | 98.1 |
| STTran | sgdet | 24.6 | 36.2 | 48.8 | 13.6 | 21.9 | 31.0 | 2.1 | 3.6 | 5.0 | 1.2 | 1.7 | 2.2 |
| TEMPURA | sgdet | 29.8 | 38.1 | 46.4 | 17.1 | 22.6 | 28.3 | 3.1 | 4.1 | 5.0 | 1.3 | 1.8 | 2.1 |
| OED | sgdet | 35.3 | 44.0 | 51.8 | 20.2 | 25.9 | 32.4 | 4.2 | 5.6 | 7.6 | 3.2 | 4.3 | 5.8 |


*[1] Cong et al., Spatial-Temporal Transformer for Dynamic Scene Graph Generation, ICCV 2021.*
*[2] Li et al., Unbiased Scene Graph Generation in Videos, CVPR 2022.*
*[3] Tang et al., OED: Towards One-stage End-to-End Dynamic Scene Graph Generation, CVPR 2023.*

# Key Idea: Interaction-Centric Reasoning

**“A person's intention to interact manifests through observable behavior, even when the target object is poorly visible”**

# Dataset Reformulation I: Relation Taxonomy

## Action genome – relation classes

- Remove relation labels indicating the absence of interactions

### [Contact relation]

| | |
|---|---|
| Carrying | Carrying |
| Drinking from | Drinking from |
| Eating | Eating |
| Holding | Holding |
| Touching | Touching |
| Twisting | Twisting |
| Wearing | Wearing |
| Wiping | Wiping |
| Writing on | Writing on |
| Standing on | Standing on |
| Have it on the back | Have it on the back |
| Leaning on | Leaning on |
| Lying on | Lying on |
| Sitting on | Sitting on |
| Covered by | Covered by |
| Not contacting | ~~Not contacting~~ |
| Other relationship | ~~Other relationship~~ |

### [Spatial relation]

| | |
|---|---|
| Above | Near |
| Beneath | |
| Infront of | |
| Behind | |
| On the side of | |
| In | |

### [Attention relation]

| | |
|---|---|
| Looking at | Looking at |
| Not looking at | ~~Not looking at~~ |
| Unsure | ~~Unsure~~ |

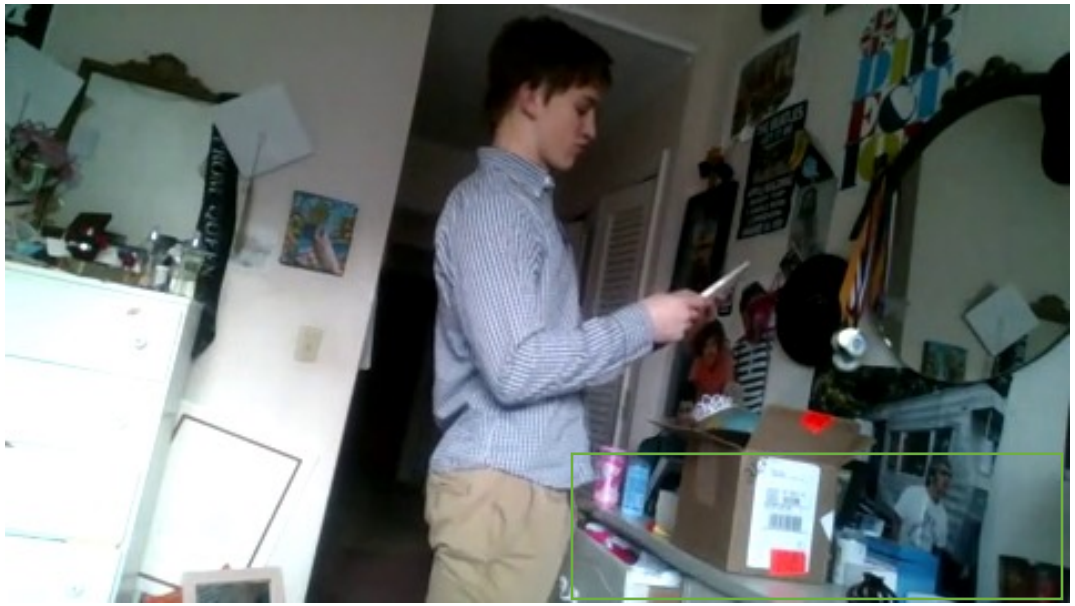
KYUNGPOOK NATIONAL UNIVERSITY

Visual AI

# Dataset Reformulation II: Interaction-Focused GT

## Training target

- Objects without interaction relations are excluded from the GT
- Objects associated solely with spatial relations are excluded from the GT

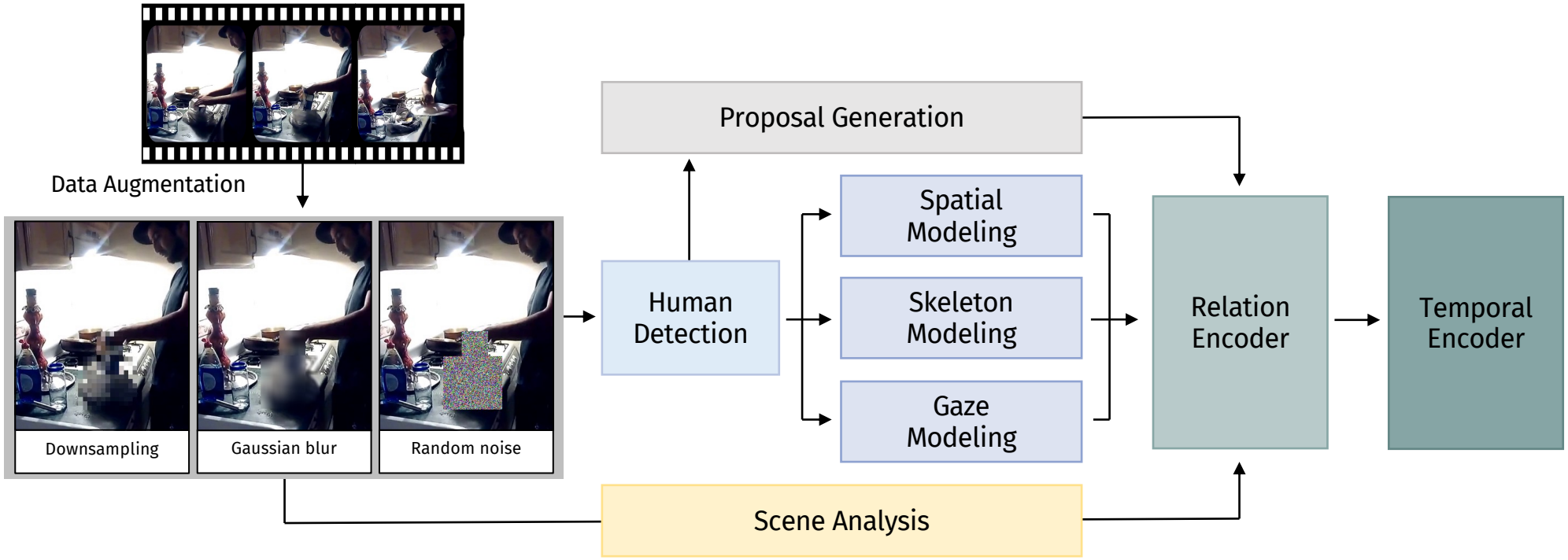


| GT | | |
|---|---|---|
| Person | looking at<br>in front of<br>holding | picture |
| Person | not looking at<br>in front of<br>not contacting | closet/cabinet |
| Person | looking at<br>in front of<br>holding | paper/notebook |

| GT | | |
|---|---|---|
| Person | looking at<br>in front of<br>holding | picture |
| | not looking at<br>not contacting | |
| Person | looking at<br>in front of<br>holding | paper/notebook |

# Method Overview

- Training with artificially degraded appearances of interaction targets
- Interaction proposal generation and refinement using hand and gaze cues
- Infer interaction types and targets using pose, gaze, spatial, and VLM-based contextual cues



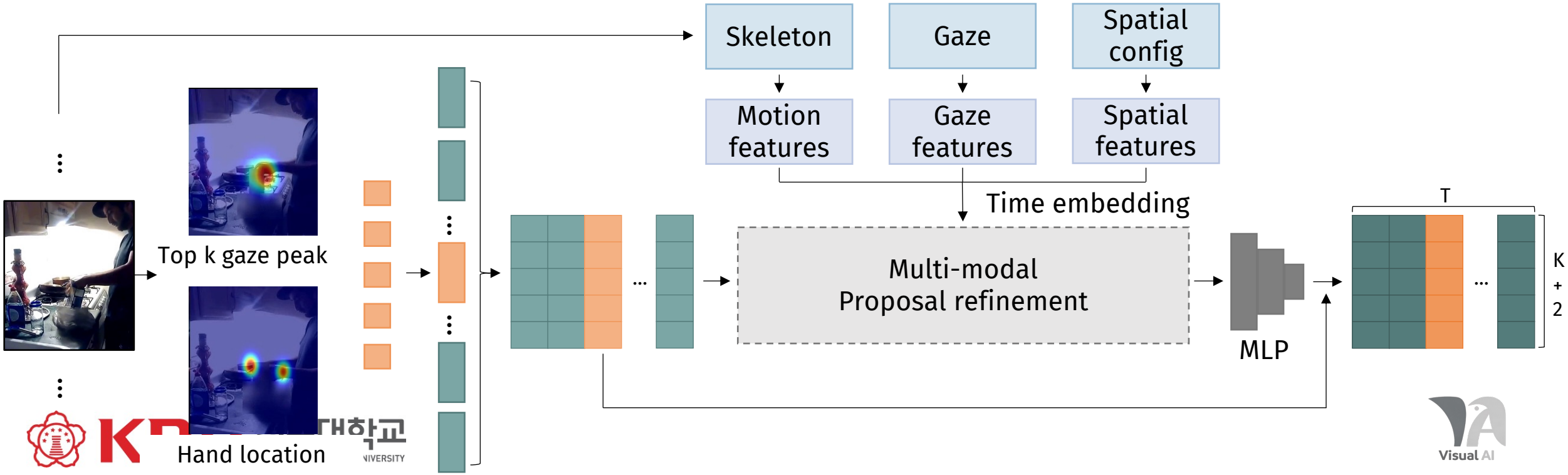
# Interaction-Centric Proposal Generation

## Proposal Initialization

- Generate fixed-size proposals from both hand locations and the gaze target
- 2 hand proposals + 1 gaze proposal per frame

## Temporal Proposal Refinement

- Refine proposal locations using human behavioral and spatial cues
- Leverage temporal consistency across consecutive frames



# Relation & Object Reasoning

## Proposal-wise Relation Encoding

- Different cues are adaptively weighted for each interaction proposal
- Cross-attention with learnable proposal embedding as query over spatial, skeleton, gaze, and context features

## Context-guided Object Reasoning

- Inject object identity using scene-level semantic priors rather than object appearance
- Semantic attention over the top-3 object embeddings proposed by the VLM

## Temporal Relation Modeling

- Refine interaction representations by modeling global and local temporal dependencies
- Global self-attention followed by sliding-window attention across frames

# How Do We Evaluate It?

## Evaluation Metric - Recall@K

- Proportion of ground-truth relations retrieved within the top-K predictions

## Decomposed Evaluation

- Proposal Matching : Can the model localize the intended interaction target?
- Object Pair Recognition : Can the model identify what object the person interacts with?
- Relation Pair Recognition : Can the model identify what interaction is occurring?
- Triplet Recognition : Can the model correctly predict the full ⟨human, relation, object⟩ triplet?

# Experimental Results: Decomposed Analysis

## Interaction Target Localization

- Stable proposal localization (63–66%) across all degradation types
- Hand and gaze targets cover a substantial portion of interaction regions

| Dataset | AG-Blur | AG-Random Noise | AG-Downsampling |
|---|---|---|---|
| Localization (%) | 66.4 | 63.3 | 65.6 |

## Relation vs. Object Recognition

- Relation recognition remains more stable than object recognition
- Behavioral cues provide strong signals for interaction semantics

| Dataset | Object Pair | | | | Relation Pair | | | |
|---|---|---|---|---|---|---|---|---|
| | R@5 | R@10 | R@20 | R@50 | R@5 | R@10 | R@20 | R@50 |
| AG-Blur | 16.8 | 27.1 | 40.4 | 59.7 | 41.0 | 50.6 | 55.9 | 57.0 |
| AG-Random Noise | 16.9 | 26.4 | 38.4 | 57.4 | 37.2 | 47.6 | 53.5 | 54.8 |
| AG-Downsampling | 16.3 | 25.9 | 38.8 | 58.6 | 40.6 | 49.7 | 55.4 | 56.7 |

# Experimental Results: Decomposed Analysis

## Full Triplet Recognition

- Triplet prediction remains challenging due to compounded errors
- Performance remains consistent across different degradation types

| Dataset | R@10 | R@20 | R@50 |
|---|---|---|---|
| Blur | 14.4 | 21.3 | 31.0 |
| Random Noise | 12.8 | 18.8 | 28.2 |
| Downsampling | 13.1 | 19.6 | 29.4 |

# Ablation: What Actually Matters?

## Skeleton Motion

- Largest performance drop when removed
- Key cue for relation and triplet recognition

## Gaze

- Significant impact on proposal localization
- Relatively small impact on triplet recognition

## VLM Context

- Provides mild but consistent improvements
- Acts as a complementary scene-level prior

| | Proposal Localization (%) | Triplet Recognition | | |
|---|---|---|---|---|
| | | R@10 | R@20 | R@50 |
| Full model | 66.4 | 14.4 | 21.3 | 31.0 |
| w/o vlm | 66.5 | 13.7 | 20.2 | 29.5 |
| w/o Gaze | 57.2 | 13.9 | 20.4 | 29.4 |
| w/o skeleton | 53.2 | 9.8 | 14.6 | 21.8 |

# Can It Complement Existing VidSGG?

## Integration with Existing Models

- Fuse predictions using a lightweight confidence-based MLP
- Both baseline and interaction-centric models remain frozen
- Evaluated on the original Action Genome setting

## Performance Gains

- STTran (2-stage) : R@10 24.6 → 29.1
  - Improves ranking among filtered candidates
- OED (1-stage) : R@50 51.8 → 54.4
  - Promotes relevant candidates in a larger candidate space

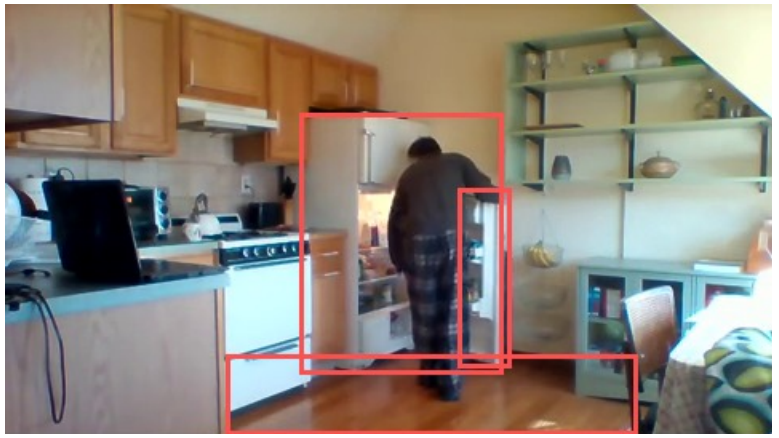
| Model | R@10 | R@20 | R@50 |
|---|---|---|---|
| STTran | 24.6 | 36.2 | 48.8 |
| STTran + Ours | 29.1 | 39.4 | 48.7 |
| OED | 35.3 | 44.0 | 51.8 |
| OED + Ours | 36.9 | 46.1 | 54.4 |

# Qualitative Results

## GT

person
- looking at / in front of / holding — food
- looking at / in front of / holding — bag

## OED



person
- looking at / in front of / touching — refrigerator
- not looking at / beneath / standing on — bag
- unsure / in front of / not contacting — door

→ Captures interactions with visible objects, but misses poorly visible interaction targets

## Ours

person — looking at / near / holding — food

→ Identifies low-visibility interaction targets through behavioral cues

KNU 경북대학교 KYUNGPOOK NATIONAL UNIVERSITY

# Limitation & Future Works

## Limited Interaction Coverage

- Currently focused on hand-related interactions

→ Extend to indirect interactions and broader spatial relations

## Lack of Appearance Cues

- Target appearance features are not explicitly incorporated

→ Jointly integrate appearance- and interaction-centric cues based on their reliability

# Thanks for your attention