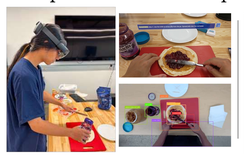
Mistake ATTribution: Fine-Grained Mistake Understanding in Egocentric Videos

Yayuan Li[1], Aadit Jain[1], Filippos Bellos[1], Jason J. Corso[1,2]
[1]University of Michigan, [2]Voxel 51

CVPR JUNE 3-7, 2026 DENVER COLORADO

Mistake Understanding — What? & Why?

- Envisioning Instructional AI:
  Upskill novice to complete professional tasks


Daily Cooking

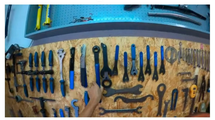

Professional Radiology

- Human mistake is inevitable:
  Beyond Recognition and Prediction

Existing Mistake Understanding

- Lack of diagnostic details

Attempt Video



Instruction Text: "Grab a chain whip"

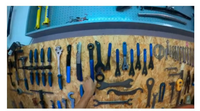
Mistake Detection: Yes

Mistake Recognition: "Modification Mistake"

Mistake ATTribution — Problem Formulation

Input

Grab a chain whip
Instruction Text


Attempt Video

Mistake ATTribution
What – When – Where

Semantic Attribution

Grab a chain whip
Correct Predicate | Mistaken Object

Temporal Attribution


Point-of-No-Return (PNR): t=3.2s

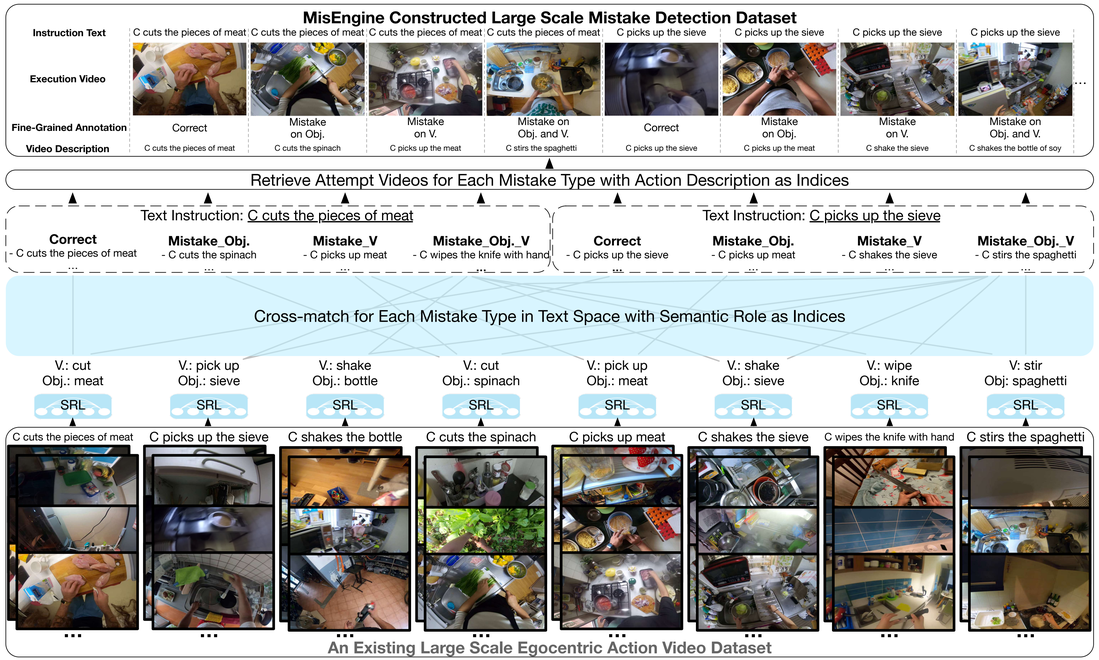
Spatial Attribution

Technical Challenges

- Benchmark: Hard to manually collect large scale mistake dataset and attribution annotation
- Method: Lack of a unified model to produce semantic (what), temporal (when), spatial (where) attribution

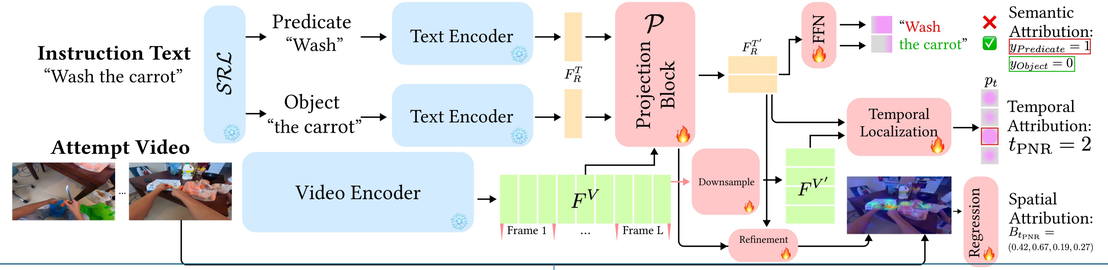
MisEngine – Automatic Dataset Construction

- The essence of mistakes is the divergence of video from text



MisFormer – Unified Structure for MATT

- Pretrained VLM for feature extraction
- Tailored attribution heads



Quantitative Results

Dataset	# Samples: Total	# Samples: By activity	Semantic Var.: Activ.	Semantic Var.: Dm.	Visual Var.: Env.	Visual Var.: Part.	Annotation: Det.	Annotation: Sem.	Annotation: Temp.	Annotation: Spa.
EgoPER [23]	599	9.7	62	1	2	11	✓	▲	✗	▲
Assembly101 [4, 40]	707	2.0	358	1	1	53	✓	▲	✗	✗
HoloAssist [46]	7,562	0.91	8,285	2	-	222	✓	▲	✗	✗
CaptainCook4D [37]	1,964	5.6	352	1	10	8	✓	▲	✗	✗
Epic-Tent [18]	626	52.2	12	1	-	24	✓	▲	✗	▲
EPIC-KITCHENS	89,975	5.0	18,003	1	45	37	✗	✗	✗	▲
Ego4D	155,367	2.1	75,423	14+	100+	931	✗	✗	▲	▲
EPIC-KITCHENS-M	221,094	18.0	12,283	1	45	37	✓	✓	✗	▲
Ego4D-M	257,584	16.0	16,099	14+	100+	248	✓	✓	✓	✓

Dataset	Method	Average: Acc. ↑	Average: F1 ↑	Predicate: Acc. ↑	Predicate: F1 ↑	Object: Acc. ↑	Object: F1 ↑
EK	LLaVA-Vid [49]	71.71	70.34	64.10	64.55	79.32	76.13
	Vid-Chat [29]	64.17	63.27	54.45	52.11	73.89	74.43
	ChatGPT [34]	77.66	77.23	66.23	65.11	89.09	89.35
	MisFormer (Ours)†	**84.91**	**84.78**	**76.35**	**75.31**	**93.47**	**94.25**
	MisFormer (Ours)	**84.13**	**83.89**	**76.83**	**76.43**	**91.43**	**91.34**
Ego	LLaVA-Vid [49]	42.83	40.12	48.10	47.95	37.56	32.29
	Vid-Chat [29]	36.47	34.36	36.73	36.84	36.21	31.88
	ChatGPT [34]	52.45	50.95	53.42	50.20	51.48	51.70
	MisFormer (Ours)†	**59.37**	**55.41**	**58.44**	**53.03**	**63.33**	**57.79**
	MisFormer (Ours)	**62.03**	**56.24**	**58.40**	**55.22**	**65.66**	**57.25**

Dataset Profile

Method	MAE (frames) ↓	MAE (seconds) ↓
EgoMotion-COMPASS [24]	48.96	1.632
EgoT2 [48]	24.48	0.816
MisFormer (Ours)	**19.14**	**0.638**

Temporal Attribution

Semantic Attribution

Method	mIoU (%) ↑	CD (%) ↓	BSE (%) ↓
MediaPipe-U [28]	49.88	13.47	20.98
SSDA [25]	**64.54**	**7.21**	**12.34**
MisFormer (Ours)	59.21	10.36	16.27

Spatial Attribution

Qualitative Results