UNIVERSITY OF MICHIGAN

VOXEL51

CVPR JUNE 3-7, 2026 DENVER COLORADO

Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos

Yayuan Li[1], Aadit Jain[1], Filippos Bellos[1], Jason Corso[1,2]

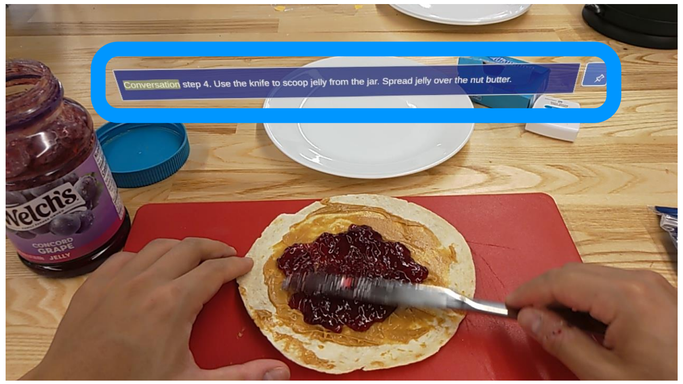
[1]University of Michigan, [2]Voxel51

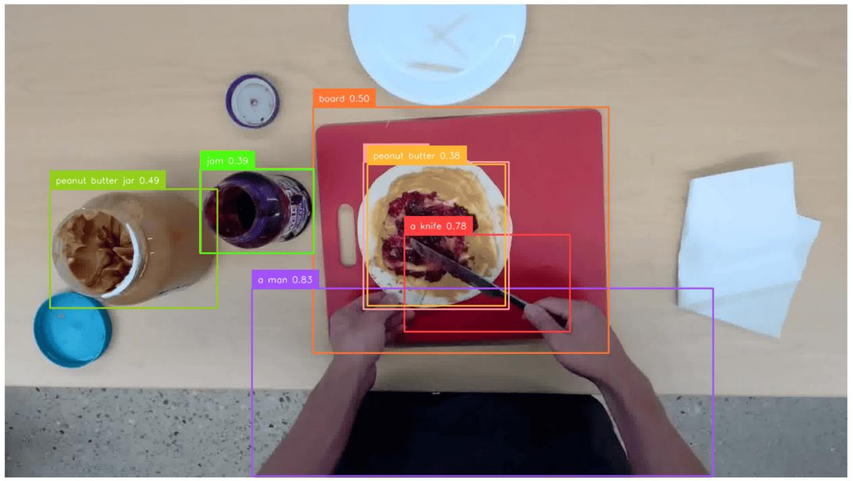
Page

Contact

Mistake Understanding — What? & Why?

Envisioning Instructional AI[1,2,3]:
Assist human to complete physical tasks





Understanding human activity is fundamental

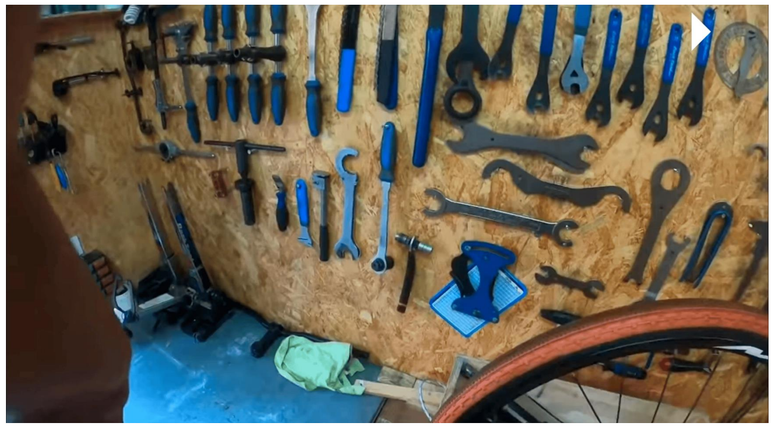
Human mistake is an under-explored activity in development of Instructional AI.

1. Bellos, Filippos*, Yayuan Li*, Cary Shu, Ruey Day, Jeffrey Siskind, and Jason Corso. "Towards Effective Human-in-the-Loop Assistive AI Agents." In Proceedings of the IEEE/CVF International Conference on Computer Vision (Workshop), pp. 2513-2522. 2025.
2. Castelo, Sonia, Joao Rulff, Erin McGowan, Bea Steers, Guande Wu, Shaoyu Chen, Iran Roman et al. "Argus: Visualization of ai-assisted task guidance in ar." IEEE transactions on visualization and computer graphics 30, no. 1 (2023): 1313-1323.
3. Bao, Yuwei, Keunwoo Yu, Yichi Zhang, Shane Storks, Xiao Zheng, and Joyce Chai. "Can Foundation Models Watch, Talk and Guide You Step by Step to Make a Cake?." the Association for Computational Linguistics: EMNLP 2023, pp. 12325-12341. 2023.

Existing Mistake Understanding

Lack of diagnostic details

Attempt Video



Instruction Text: “Grab a chain whip”

Mistake Detection: Yes

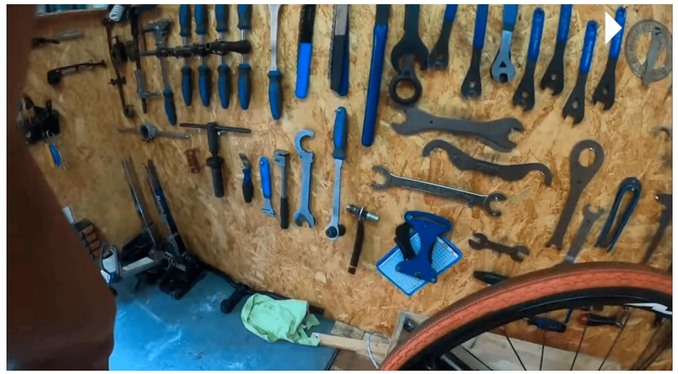
Mistake Recognition: “Modification Mistake”

Mistake Attribution — Problem Formulation

Input

Grab a chain whip

Instruction Text



Attempt Video

Mistake ATTribution

Semantic Attribution

Grab a chain whip

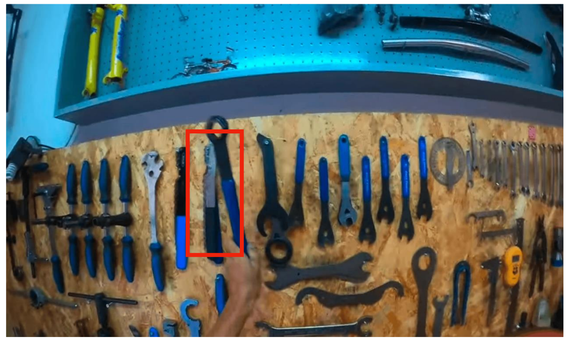
Correct Predicate

Mistaken Object

Temporal Attribution

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

Spatial Attribution



Challenges

- Benchmark:
  Hard to manually collect large scale mistake dataset and attribution annotation

- Method:
  Lack of a unified models to produce semantic (what), temporal (when), spatial (where) attribution

Challenges

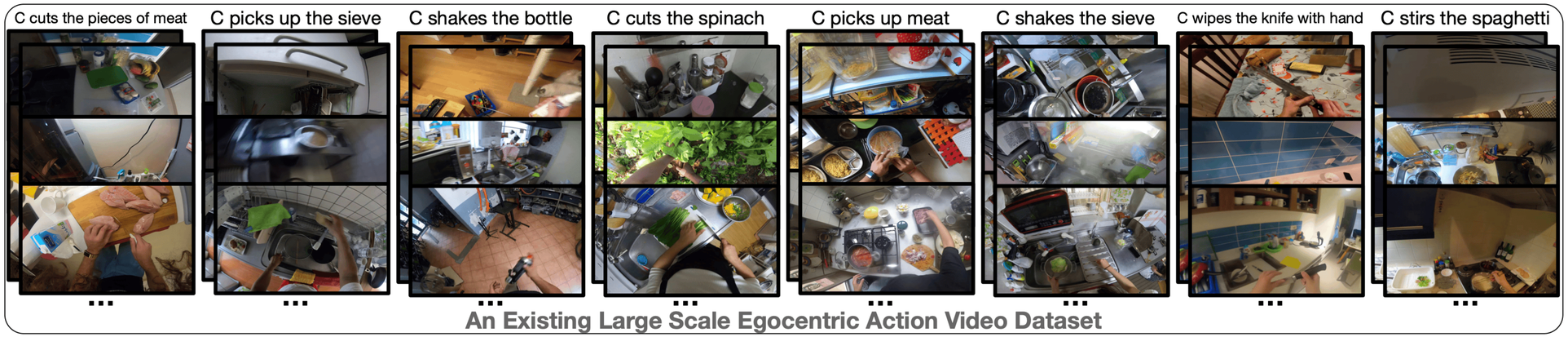
- Benchmark:
  Hard to manually collect large scale mistake dataset and attribution annotation

	# Samples		Semantic Var.		Visual Var.		Annotation			
Dataset	**Total**	**By activity**	**Activ.**	**Dm.**	**Env.**	**Part.**	**Det.**	**Sem.**	**Temp.**	**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	✓	▲	▲	✗
CaptionCook4D [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	✗	✗	▲	▲

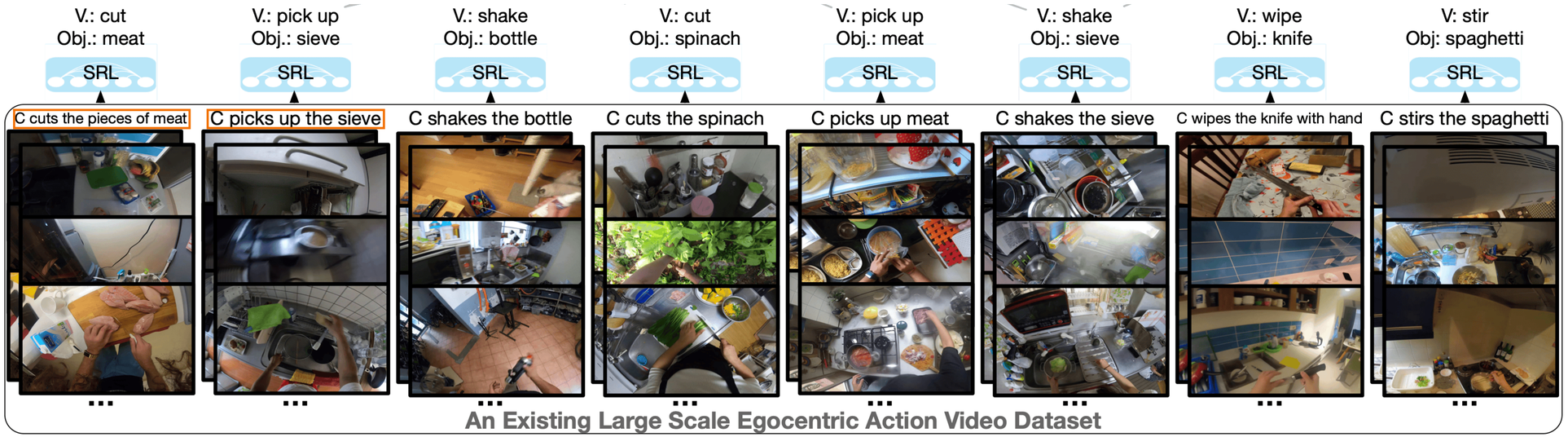
MisEngine

Automatically compose mistakes from existing action datasets

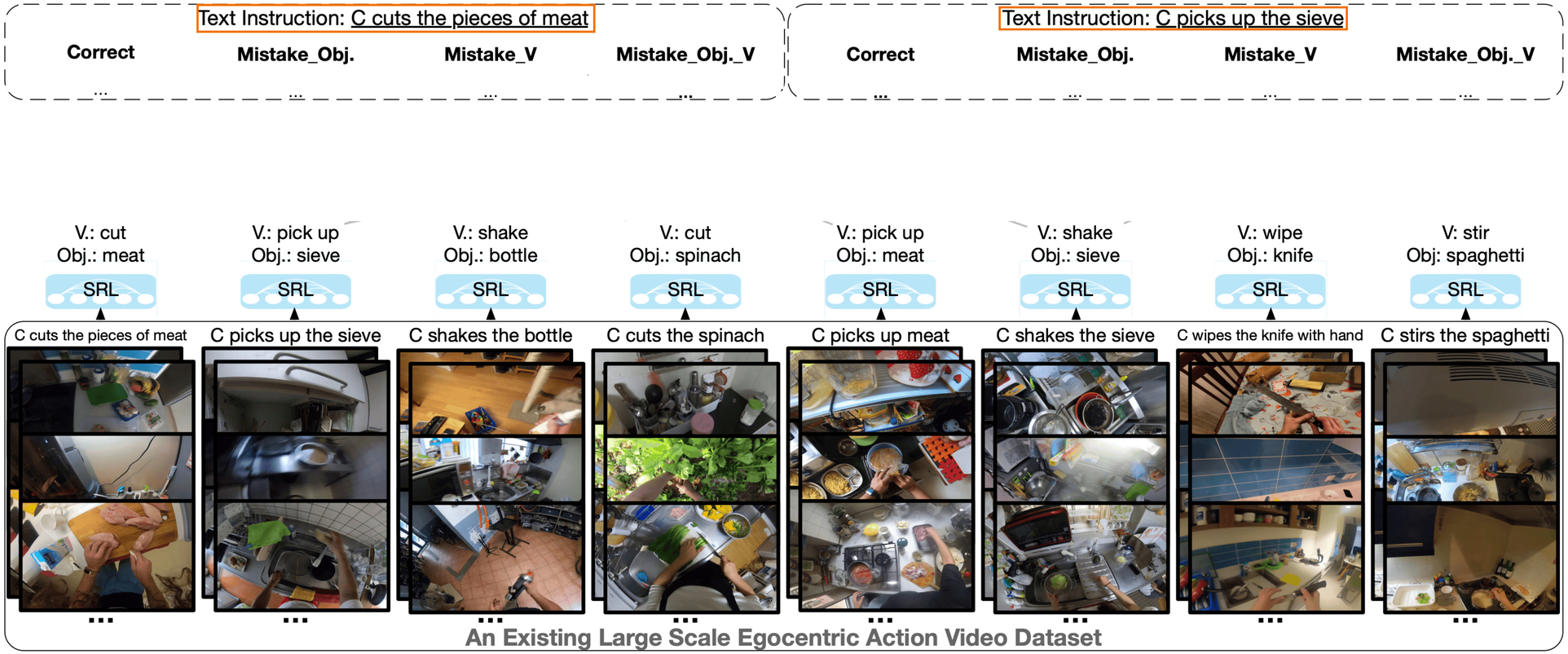
The essence of mistakes is the <u>divergence</u> of video from text



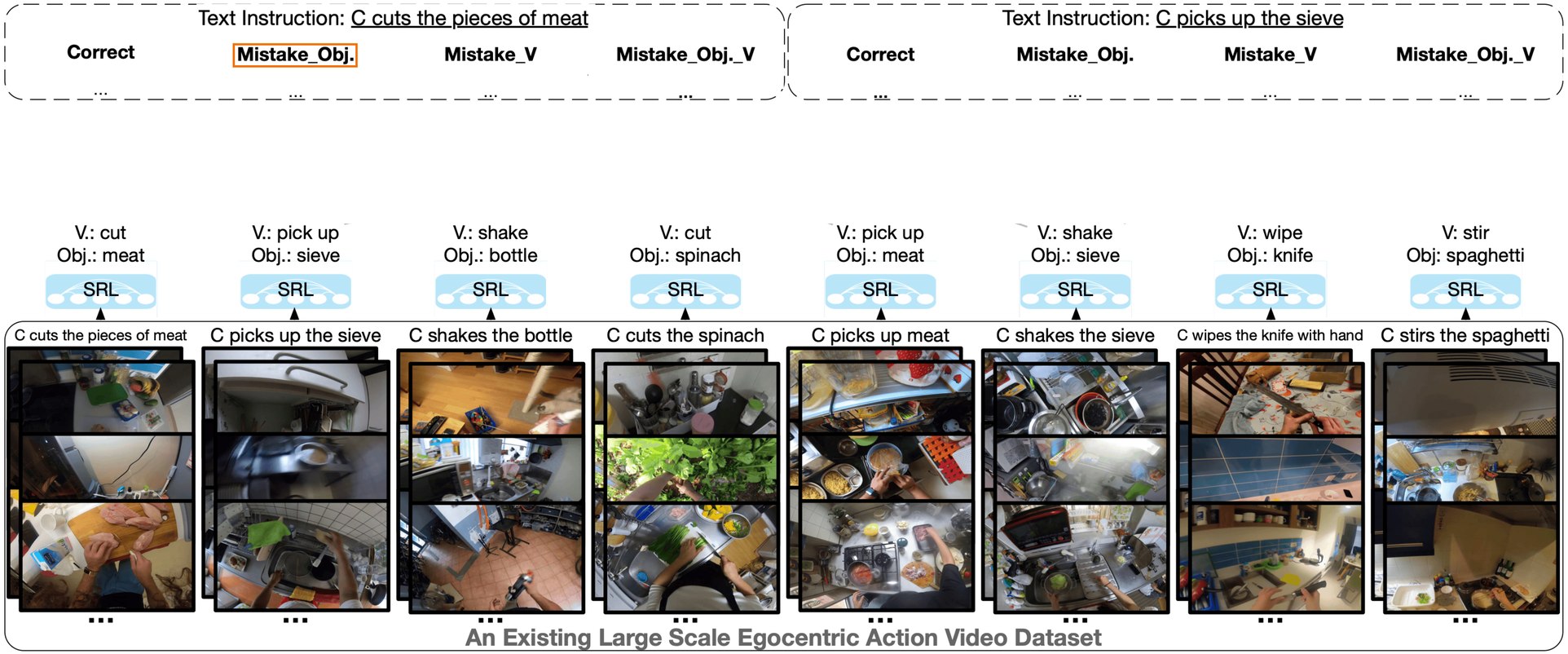
MisEngine


V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

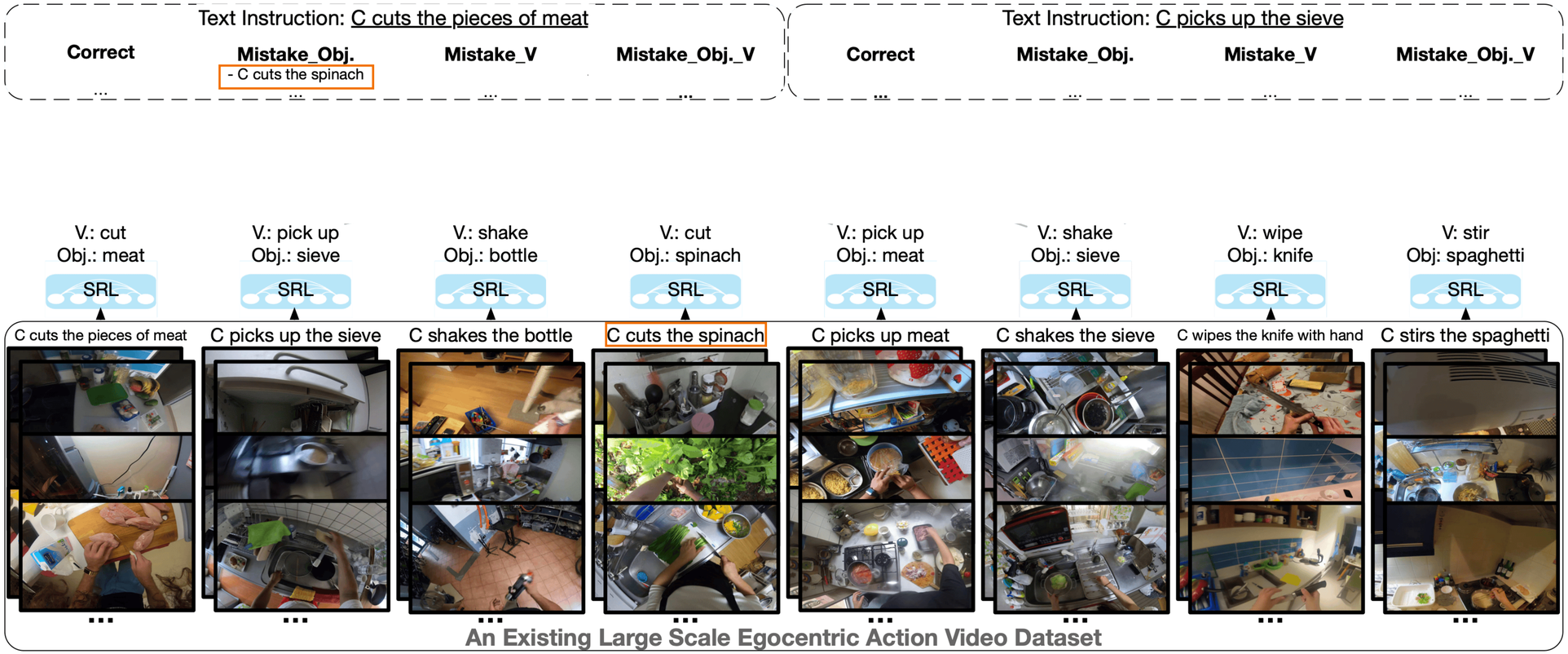
MisEngine


Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

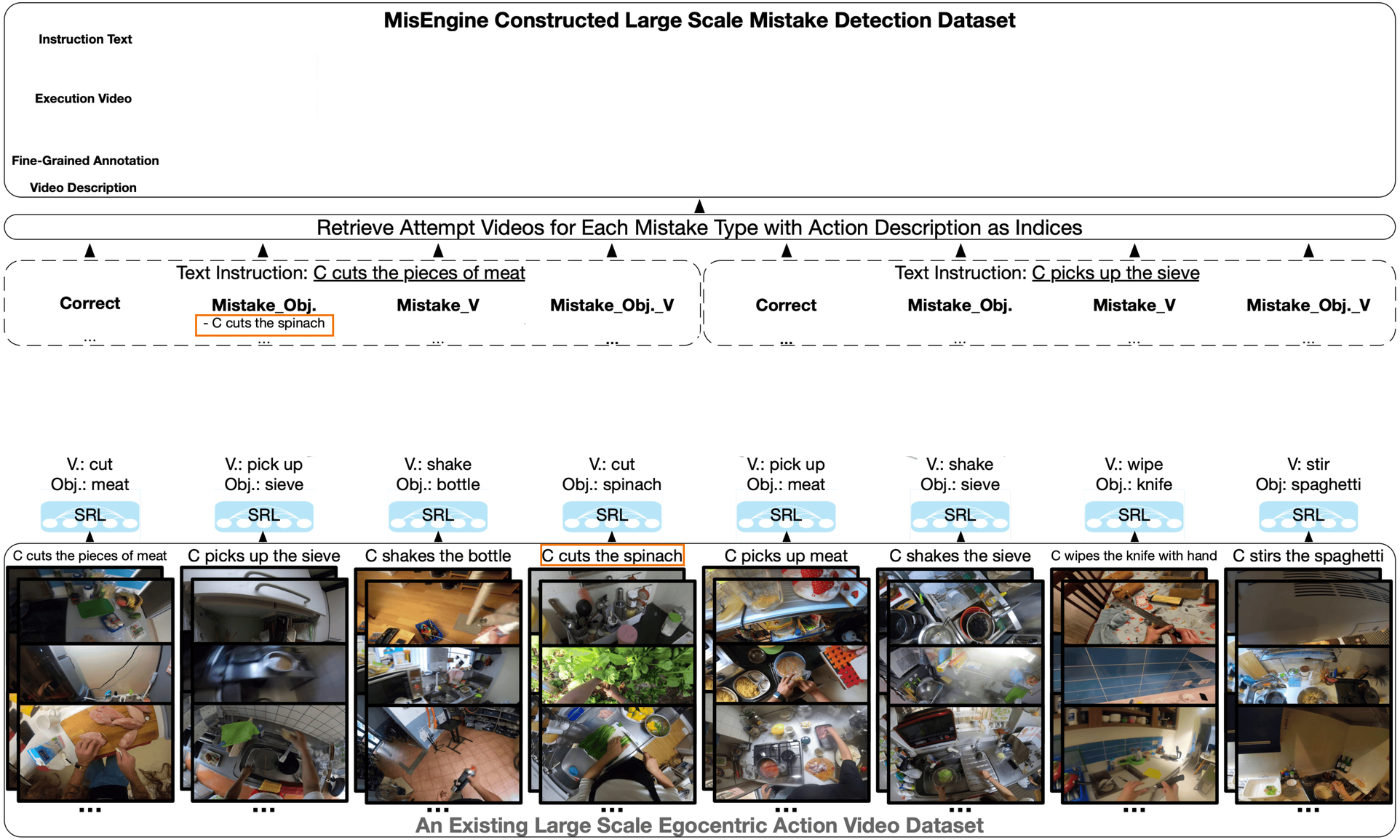
MisEngine


Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

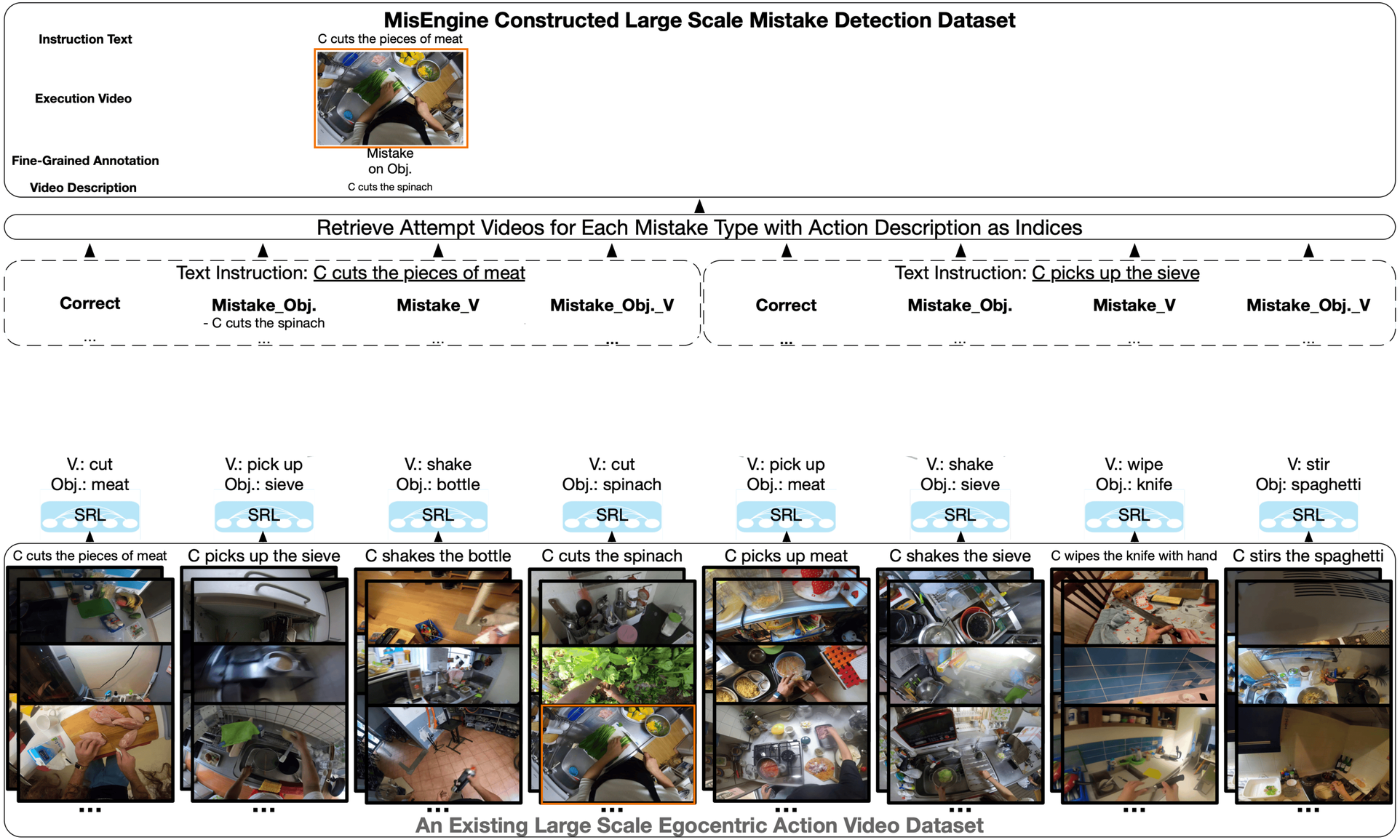
MisEngine


Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
- C cuts the spinach
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

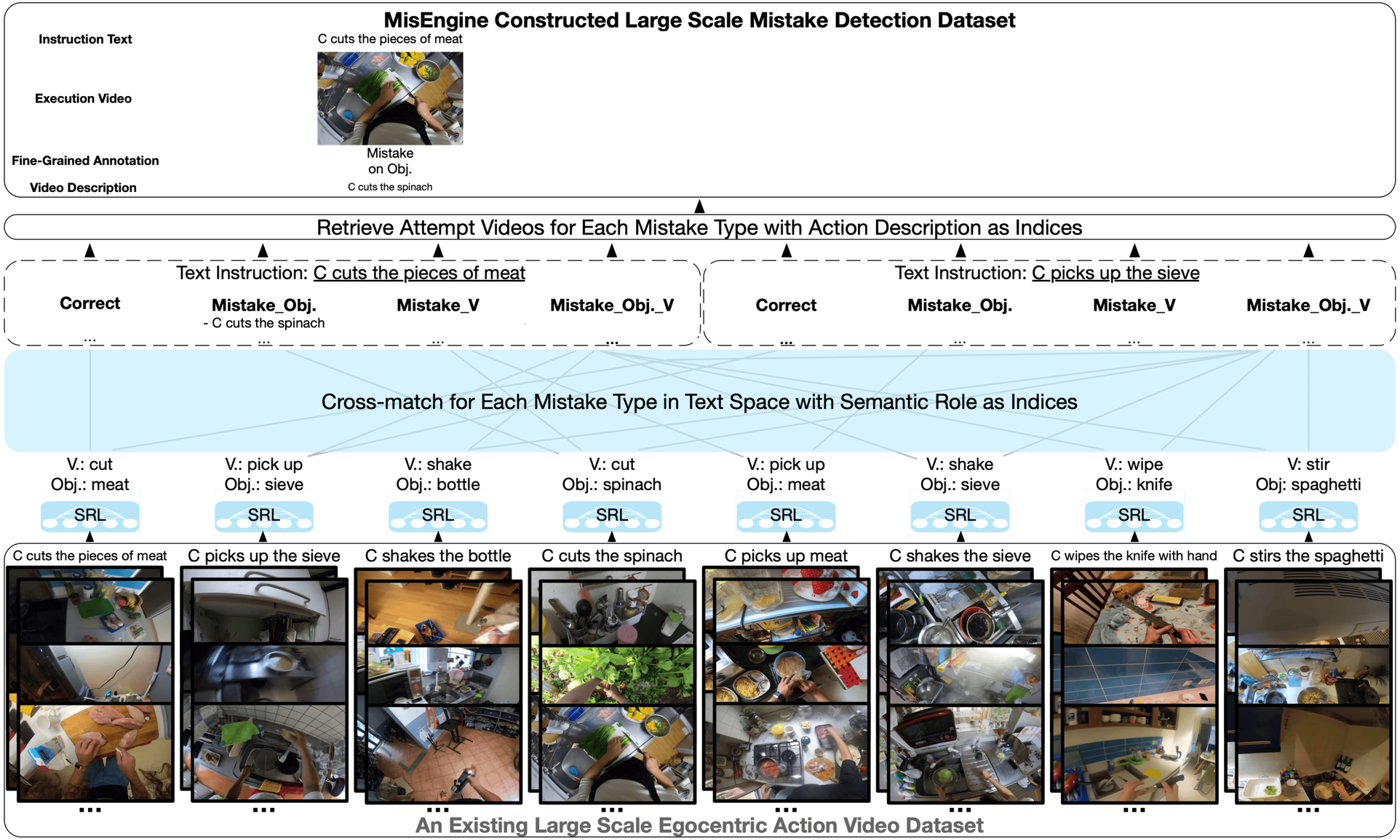
MisEngine


MisEngine Constructed Large Scale Mistake Detection Dataset
Instruction Text
Execution Video
Fine-Grained Annotation
Video Description
Retrieve Attempt Videos for Each Mistake Type with Action Description as Indices
Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
- C cuts the spinach
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

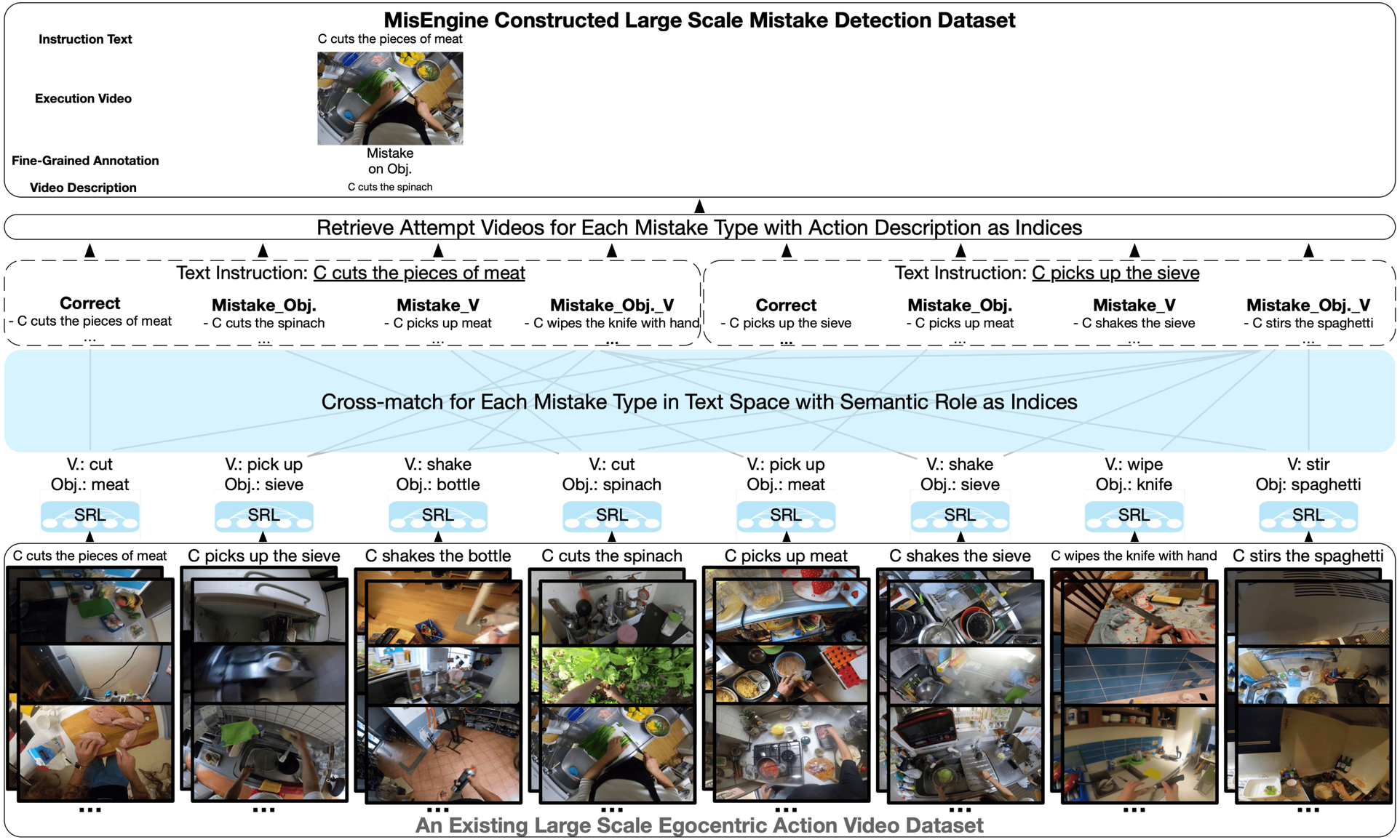
MisEngine


MisEngine Constructed Large Scale Mistake Detection Dataset
Instruction Text
C cuts the pieces of meat
Execution Video
Fine-Grained Annotation
Mistake
on Obj.
Video Description
C cuts the spinach
Retrieve Attempt Videos for Each Mistake Type with Action Description as Indices
Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
- C cuts the spinach
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

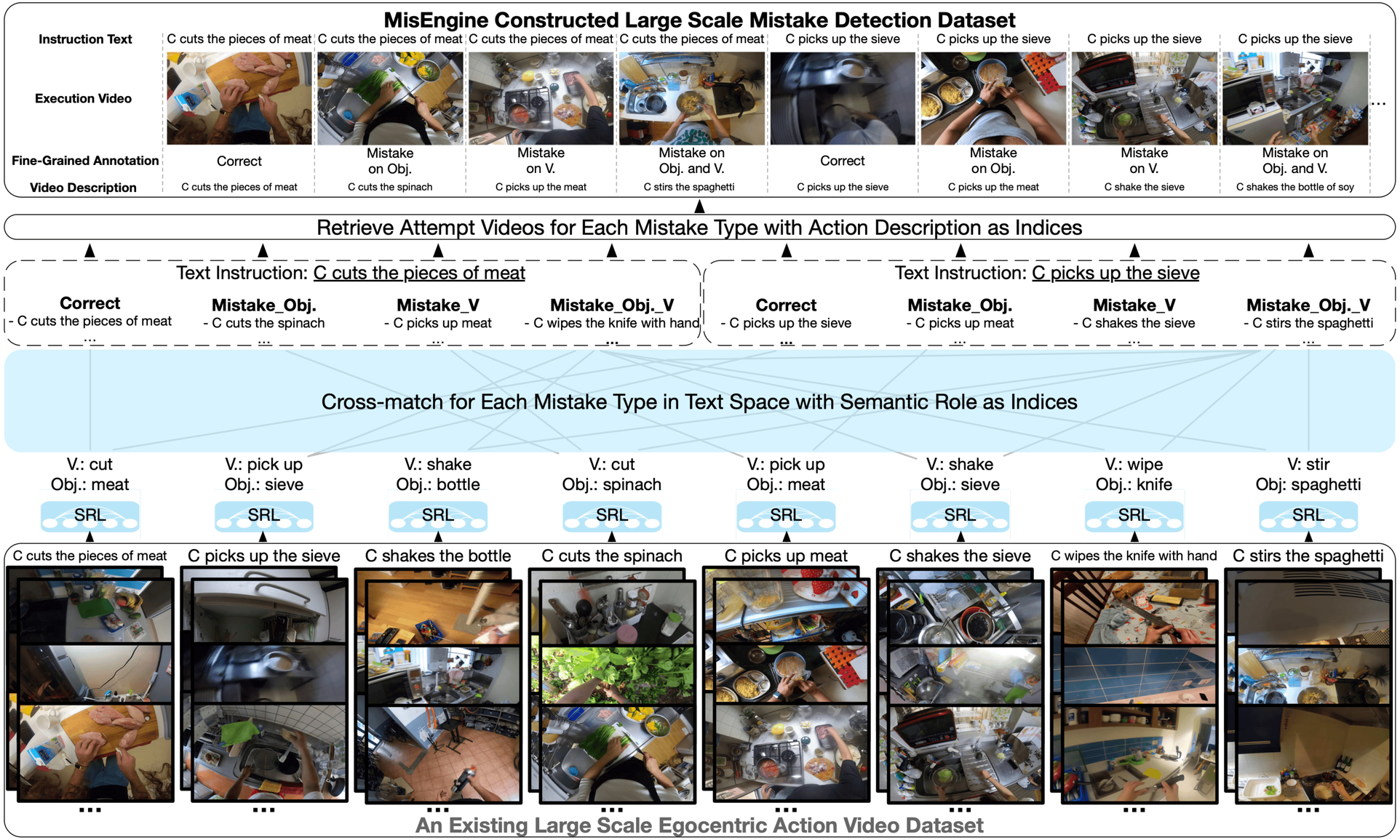
MisEngine

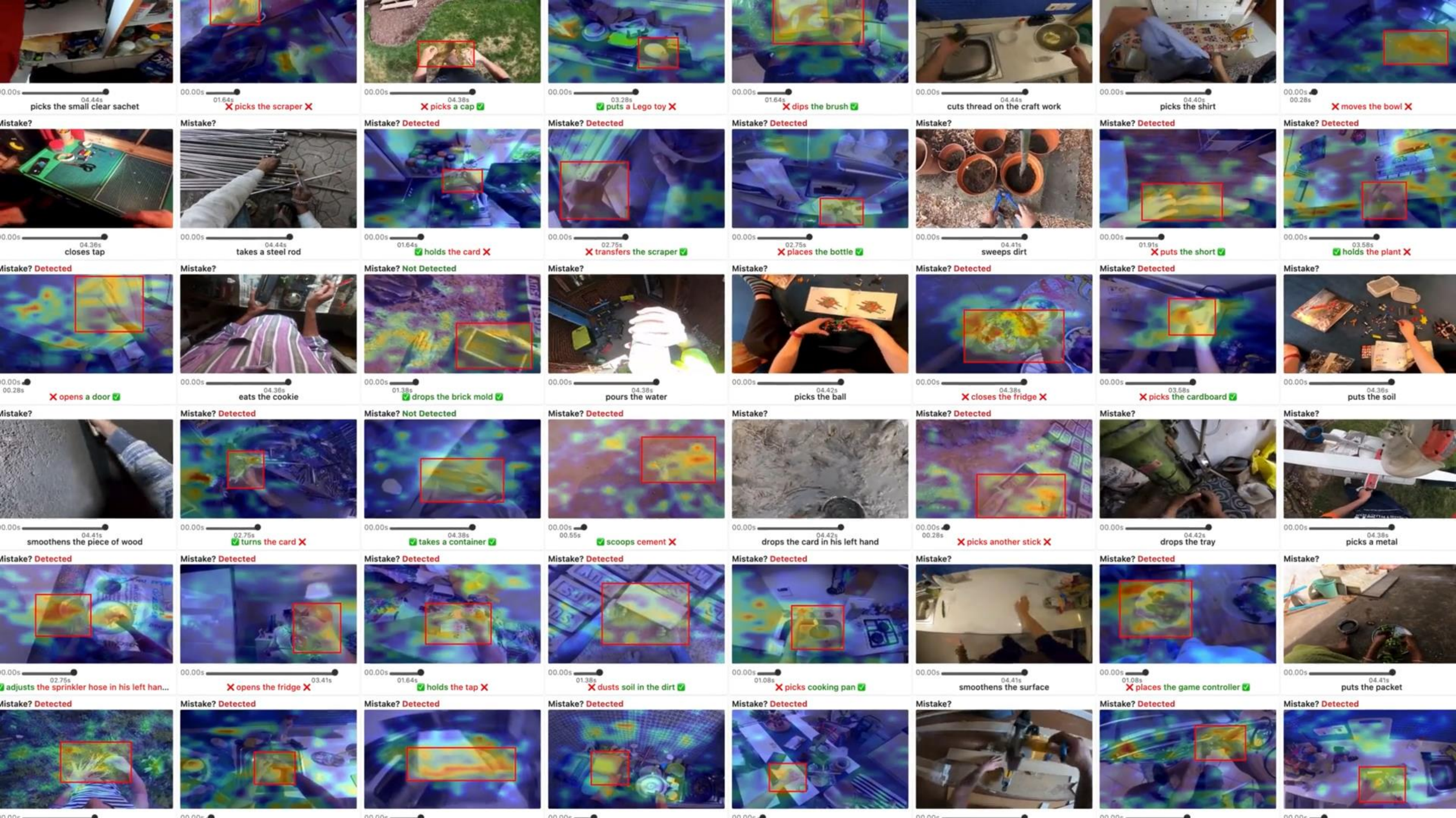

MisEngine Constructed Large Scale Mistake Detection Dataset
Instruction Text
C cuts the pieces of meat
Execution Video
Fine-Grained Annotation
Mistake
on Obj.
Video Description
C cuts the spinach
Retrieve Attempt Videos for Each Mistake Type with Action Description as Indices
Text Instruction: C cuts the pieces of meat
Correct
Mistake_Obj.
- C cuts the spinach
Mistake_V
Mistake_Obj._V
Text Instruction: C picks up the sieve
Correct
Mistake_Obj.
Mistake_V
Mistake_Obj._V
Cross-match for Each Mistake Type in Text Space with Semantic Role as Indices
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

MisEngine


MisEngine Constructed Large Scale Mistake Detection Dataset
Instruction Text
C cuts the pieces of meat
Execution Video
Fine-Grained Annotation
Mistake on Obj.
Video Description
C cuts the spinach
Retrieve Attempt Videos for Each Mistake Type with Action Description as Indices
Text Instruction: C cuts the pieces of meat
Correct
- C cuts the pieces of meat
Mistake_Obj.
- C cuts the spinach
Mistake_V
- C picks up meat
Mistake_Obj._V
- C wipes the knife with hand
Text Instruction: C picks up the sieve
Correct
- C picks up the sieve
Mistake_Obj.
- C picks up meat
Mistake_V
- C shakes the sieve
Mistake_Obj._V
- C stirs the spaghetti
Cross-match for Each Mistake Type in Text Space with Semantic Role as Indices
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset

MisEngine


MisEngine Constructed Large Scale Mistake Detection Dataset
Instruction Text
Execution Video
Fine-Grained Annotation
Video Description
C cuts the pieces of meat
C cuts the pieces of meat
C cuts the pieces of meat
C cuts the pieces of meat
C picks up the sieve
C picks up the sieve
C picks up the sieve
C picks up the sieve
Correct
C cuts the pieces of meat
Mistake on Obj.
C cuts the spinach
Mistake on V.
C picks up the meat
Mistake on Obj. and V.
C stirs the spaghetti
Correct
C picks up the sieve
Mistake on Obj.
C picks up the meat
Mistake on V.
C shake the sieve
Mistake on Obj. and V.
C shakes the bottle of soy
Retrieve Attempt Videos for Each Mistake Type with Action Description as Indices
Text Instruction: C cuts the pieces of meat
Correct
- C cuts the pieces of meat
Mistake_Obj.
- C cuts the spinach
Mistake_V
- C picks up meat
Mistake_Obj._V
- C wipes the knife with hand
Text Instruction: C picks up the sieve
Correct
- C picks up the sieve
Mistake_Obj.
- C picks up meat
Mistake_V
- C shakes the sieve
Mistake_Obj._V
- C stirs the spaghetti
Cross-match for Each Mistake Type in Text Space with Semantic Role as Indices
V.: cut
Obj.: meat
V.: pick up
Obj.: sieve
V.: shake
Obj.: bottle
V.: cut
Obj.: spinach
V.: pick up
Obj.: meat
V.: shake
Obj.: sieve
V.: wipe
Obj.: knife
V: stir
Obj: spaghetti
SRL
C cuts the pieces of meat
C picks up the sieve
C shakes the bottle
C cuts the spinach
C picks up meat
C shakes the sieve
C wipes the knife with hand
C stirs the spaghetti
An Existing Large Scale Egocentric Action Video Dataset


00.00s 04.44s picks the small clear sachet
00.00s 01.64s ✗ picks the scraper ✗
00.00s 04.38s ✗ picks a cap ✓
00.00s 03.28s ✓ puts a Lego toy ✗
00.00s 01.64s ✗ dips the brush ✓
00.00s 04.44s cuts thread on the craft work
00.00s 04.40s picks the shirt
00.00s 00.28s ✗ moves the bowl ✗
Mistake?
00.00s 04.36s closes tap
Mistake?
00.00s 04.44s takes a steel rod
Mistake? Detected
00.00s 01.64s ✓ holds the card ✗
Mistake? Detected
00.00s 02.75s ✗ transfers the scraper ✓
Mistake? Detected
00.00s 02.75s ✗ places the bottle ✓
Mistake?
00.00s 04.41s sweeps dirt
Mistake? Detected
00.00s 01.91s ✗ puts the short ✓
Mistake? Detected
00.00s 03.58s ✓ holds the plant ✗
Mistake? Detected
00.00s 00.28s ✗ opens a door ✓
Mistake?
00.00s 04.36s eats the cookie
Mistake? Not Detected
00.00s 01.38s ✓ drops the brick mold ✓
Mistake?
00.00s 04.38s pours the water
Mistake?
00.00s 04.42s picks the ball
Mistake? Detected
00.00s 04.38s ✗ closes the fridge ✗
Mistake? Detected
00.00s 03.58s ✗ picks the cardboard ✓
Mistake?
00.00s 04.36s puts the soil
Mistake?
00.00s 04.41s smoothens the piece of wood
Mistake? Detected
00.00s 02.75s ✓ turns the card ✗
Mistake? Not Detected
00.00s 04.38s ✓ takes a container ✓
Mistake? Detected
00.00s 00.55s ✓ scoops cement ✗
Mistake?
00.00s 04.42s drops the card in his left hand
Mistake? Detected
00.00s 00.28s ✗ picks another stick ✗
Mistake?
00.00s 04.42s drops the tray
Mistake?
00.00s 04.38s picks a metal
Mistake? Detected
00.00s 02.75s ✓ adjusts the sprinkler hose in his left han...
Mistake? Detected
00.00s 03.41s ✗ opens the fridge ✗
Mistake? Detected
00.00s 01.64s ✓ holds the tap ✗
Mistake? Detected
00.00s 01.38s ✗ dusts soil in the dirt ✓
Mistake? Detected
00.00s 01.08s ✗ picks cooking pan ✓
Mistake?
00.00s 04.41s smoothens the surface
Mistake? Detected
00.00s 01.08s ✗ places the game controller ✓
Mistake?
00.00s 04.41s puts the packet
Mistake? Detected
Mistake? Detected
Mistake? Detected
Mistake? Detected
Mistake? Detected
Mistake?
Mistake? Detected
Mistake? Detected

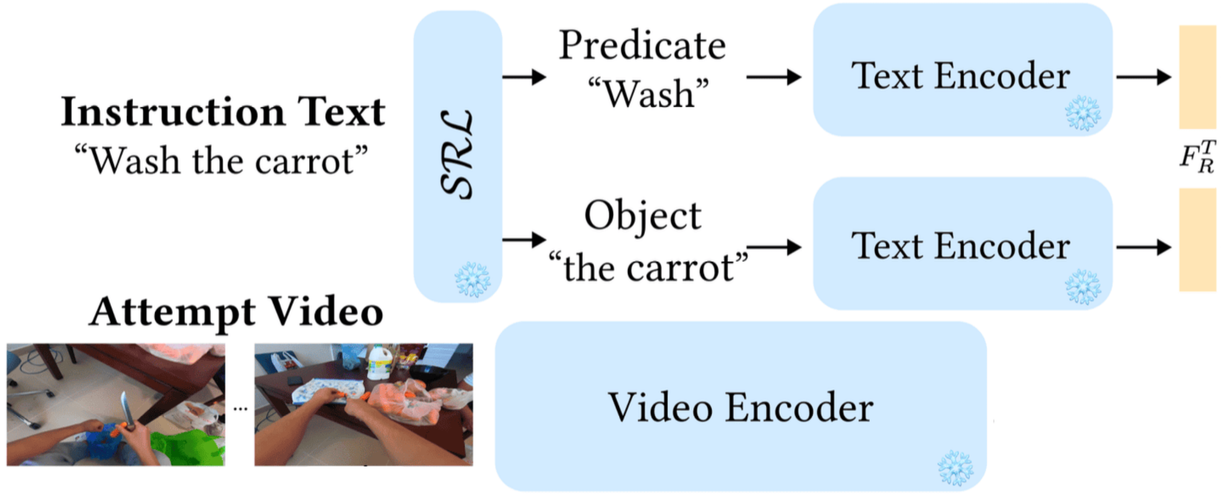
MisEngine

Inherit scale and diversity from existing action datasets
The composition implies mistake detection & semantic attribution annotation
Inherit temporal and spatial attribution from original datasets

	# Samples		Semantic Var.		Visual Var.		Annotation			
Dataset	**Total**	**By activity**	**Activ.**	**Dm.**	**Env.**	**Part.**	**Det.**	**Sem.**	**Temp.**	**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	✓	▲	▲	✗
CaptionCook4D [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	✓	✓	✓	✓

MisFormer

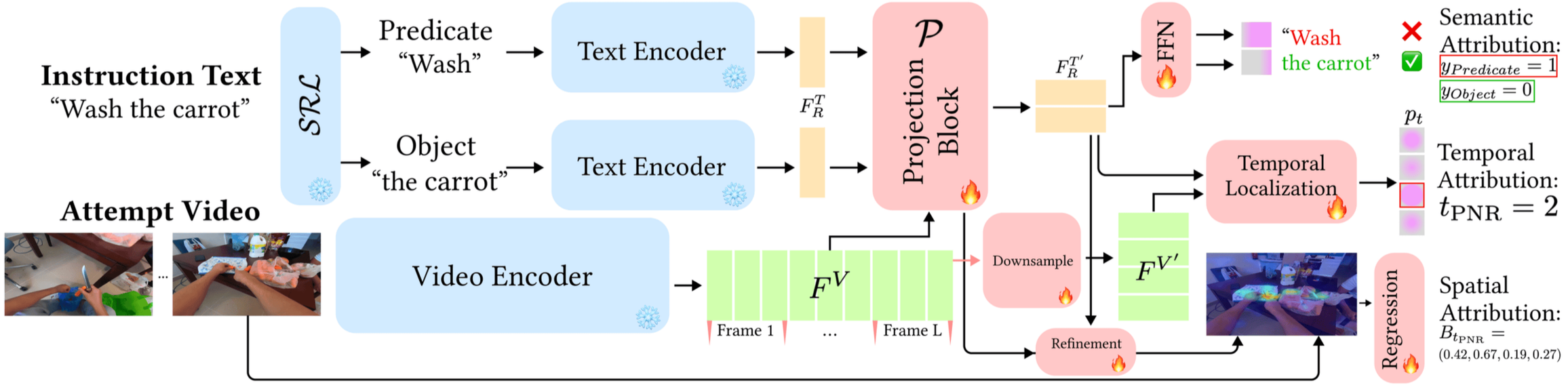
Pretrained VLM for feature extraction



MisFormer

Pretrained VLM for feature extraction

Tailored attribution heads



Quantitative Results

MisFormer, an unified model, outperforms baselines for individual task

Dataset	Method	Average		Predicate		Object	
		Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	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**	**55.44**	**53.03**	**63.33**	**57.79**
	MisFormer (Ours)	**62.03**	**56.24**	**58.40**	**55.22**	**65.66**	**57.25**

Semantic Attribution

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

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

Spatial Attribution

Results: Semantic Attribution

Prune chili peppers

Instruction Text Attempt Video

Prune chili peppers

Mistaken Predicate Correct Object

GroundTruth

Prune chili peppers

Mistaken Predicate Mistaken Object

✅ ❌

ChatGPT-4o[1]

Prune chili peppers

Correct Predicate Mistaken Object

❌ ❌

LLaVa-Vid[2]

Prune chili peppers

Mistaken Predicate Correct Object

✅ ✅

MisFormer (Ours)[3]

1. OpenAI. (2025). GPT-4o System Card. arXiv:2410.21276.
2. Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. EMNLP 2024.
3. Yayuan Li, Aditya Jain, Foteini Bellos, Jason Corso. Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos. CVPR 2026.

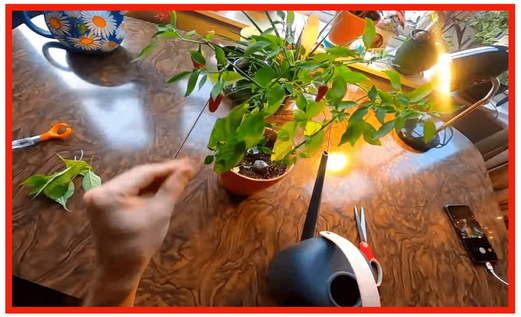
Results: Temporal Attribution

Prune chili peppers

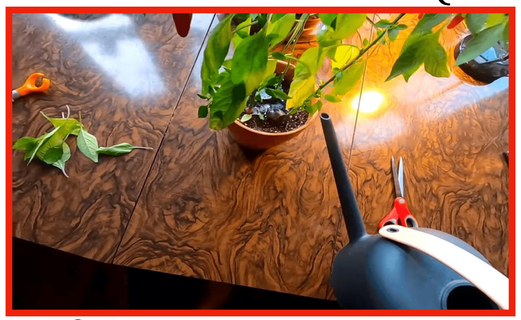
Instruction Text

Attempt Video

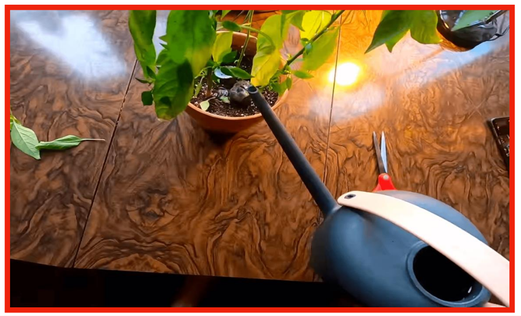
Ground Truth Point of No Return: t=2.1s



EgoMotion-COMPASS[1]: t=1.2s (Error=17.3%)



EgoT2[2]: t=3.4s (Error=24.5%)



MisFormer (Ours)[3]: t=2.5s (Error=7.7%)

1. Feichtenhofer et al. Masked Autoencoders for Egocentric Video Understanding. Ego4D Challenge Workshop, CVPR 2022.
2. Zihui Xue, Yale Song, Kristen Grauman, Lorenzo Torresani. Egocentric Video Task Translation. CVPR 2023.
3. Yayuan Li, Aditya Jain, Foteini Bellos, Jason Corso. Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos. CVPR 2026.

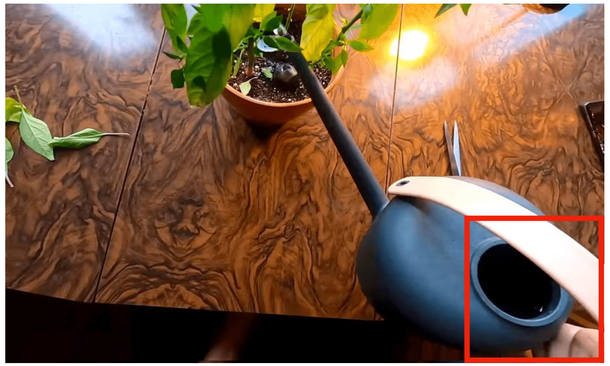
Results: Spatial Attribution

Prune chili peppers

Instruction Text

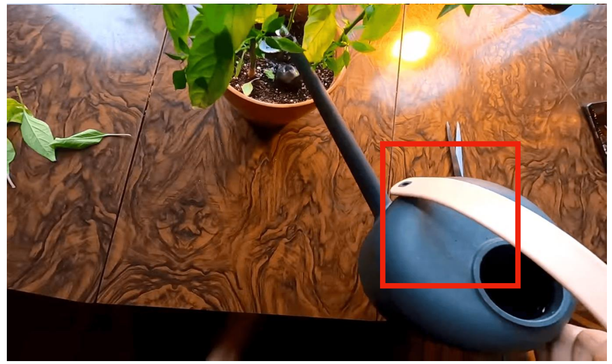
Attempt Video

Point of No Return: t=2.1s



Ground Truth

MediaPipe-U[1]: Not detected



SSDA[2]: IoU=12.1%

MisFormer(Ours): IoU=48.5%

1. Camillo Lugaresi, et al.. MediaPipe: A Framework for Perceiving and Processing Reality. CVPR Workshop, 2019.

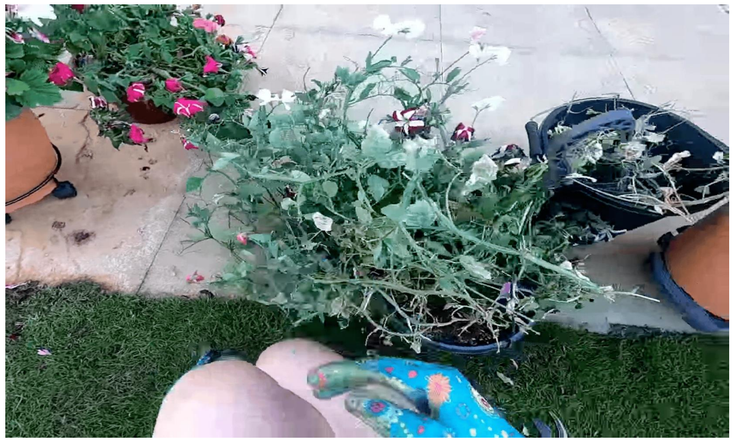
2. Rosario Leonardi, Antonino Furnari, Francesco Ragusa, Giovanni Maria Farinella. Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection? ECCV 2024.

3. Yayuan Li, Aditya Jain, Foteini Bellos, Jason Corso. Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos. CVPR 2026.

More Results

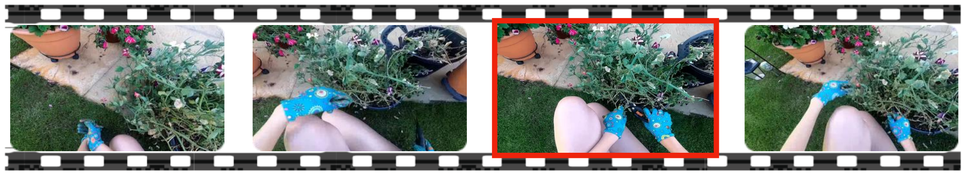
Semantic

Tie stem

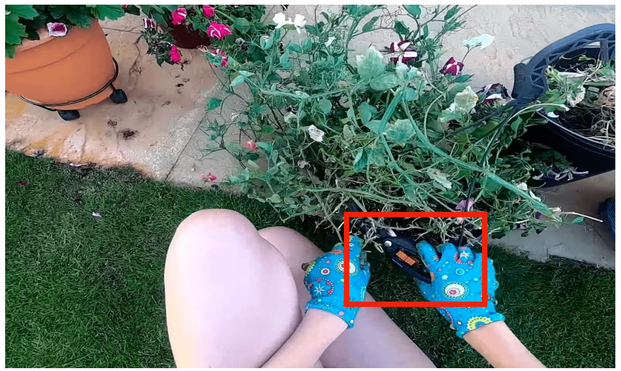


Take plate

Temporal



Spatial



Mistake?

00.00s
00.04s

picks the small clear sachet

Mistake?

00.00s
00.01s

picks the scraper

Mistake?

00.00s
00.00s

dips the brush

Mistake?

00.00s
00.00s

picks the shirt

Mistake?

00.00s
00.00s

moves the bowl

Mistake?

00.00s
00.00s

places the bottle

Mistake?

00.00s
00.00s

opens a door

Mistake?

00.00s
00.00s

smoothens the piece of wood

Mistake?

00.00s
00.00s

takes a container

Mistake?

00.00s
00.00s

scoops cement

Mistake?

00.00s
00.00s

picks another stick

Mistake?

00.00s
00.00s

dusts soil in the dirt

Follow-up

CVPR JUNE 3-7, 2026 DENVER COLORADO

Paper

Contact

Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos

Yayuan Li[1], Aadit Jain[1], Filippos Bellos[1], Jason Corso[1,2]
[1]University of Michigan, [2]Voxel51