Mistake Attribution: Fine-Grained Mistake Understanding in Egocentric Videos

Supplementary Material

A. MisEngine Details

Role-matching function $\mathcal{J}$. $\mathcal{J}(g_i^r, g_j^r)$ determines whether role r matches between two action descriptions, and thus whether a cross-matched video contains a mistake on that role. The matching rule is granularity-dependent: e.g., "cut" vs. "dice" are distinct in a professional kitchen but interchangeable in casual cooking. We implement two variants(Sec. 3.1): (1) *taxonomy-based* $\mathcal{J}$ for Ego4D, grouping synonymous verbs and nouns from the original dataset; and (2) *literal string matching* for EPIC-KITCHENS. SRL on 1K Ego4D descriptions yields 97.97% predicate and 94.13% object F1, reflecting the short imperative phrasing of the source annotations.

B. Dataset Ecological Validation

We conduct a user study (16 participants, 200 ⟨text, video⟩ pairs per dataset, no prior exposure) on 4 datasets to evaluate how realistic our constructed mistake samples and inherited annotations are. We report Real-%: the % of participants who judge the sample pair or inherited annotation as realistic, averaged over videos.

EK-M (ours)	Ego4D-M (ours)	EgoPER [23]	CC4D [37]
92.32±6.04	92.42±7.01	94.17±4.02	95.75±4.03

(a) % of samples human-rated as realistic (Real-% ↑) on mistake videos.

Tmp. (%)	Spa. (%)
96.29 ± 3.99	95.81 ± 2.16

(b) Real-% on inherited annotations.

Mis_V	Mis_{Obj}	Mis_{Both}
60.73	57.99	58.91

(c) Spatial-Attribution mIoU by mistake type.

Tab. (a): All four datasets receive comparably high realism ratings (within std). This is expected since cross-matched samples are drawn from the same domain, grounded in original metadata. Participants occasionally flag samples in prior datasets as "staged," consistent with the manual collection challenges discussed in Sec. 1.

Tab. (b): Participants judge whether the inherited PNR frame marks the consolidation of a mistake (Tmp.) and whether the bounding box captures mistake-relevant detail (Spa.). Both achieve >95% Real-%, validating the quality of inherited annotations. The temporal attribution frame occurs at 79.48% of video duration on average. Samples are constructed in equal numbers across mistake types for balanced training; this count is configurable (Sec. 3.1).

C. Additional Experiments

Zero-shot transfer to EgoPER. *MisFormer* generalizes zero-shot to EgoPER, achieving 34.46 F1 / 35.52 Acc and surpassing both baselines. Tab. 5 results show that manually collected datasets provide insufficient training signal for *MisFormer* (Tab. 1), underscoring the necessity of *MisEngine*.

Spatial attribution mIoU by mistake type. Tab. (c) reports spatial attribution mIoU broken down by mistake type. Mis_{Both} (which uses a union bounding box over hand and object) performs comparably to single-role types. Union boxes remain compact because, at the PNR frame, hand and object regions substantially overlap during active interaction.

Ablation: semantic head and projection block. Tab. 6 in the main paper ablates the projection block $\mathcal{P}$ and task-specific heads. Removing $\mathcal{P}$ degrades all tasks (row c). Removing temporal or spatial supervision each hurts the complementary task (rows d, e), confirming their mutual benefit. We additionally ablate the semantic head: without it, semantic attribution and mistake detection are disabled, while temporal and spatial attribution improve to 0.358 s MAE and 65.33% mIoU. This suggests the current $\mathcal{P}$ faces tension between representing semantic-level and grounded-level features simultaneously, which we identify as a direction for future work. The unified structure nonetheless avoids training separate models, runs 2.42× faster (Sec. 4.2), and outperforms 6 of 7 baselines (Tabs. 2 to 4).

D. Qualitative Results and Failure Analysis

Qualitative examples. Fig. 5 shows representative outputs. Row 1 correctly produces semantic attribution, localizes the PNR frame, and generates a reasonable spatial box (IoU=0.56). Row 2 illustrates a failure case: semantic attribution is correct, but temporal attribution selects an earlier frame before the bag is set down. Judging the spatial completion state of an object from 2D video alone is inherently difficult, suggesting that depth or 3D cues are a promising direction for future work. Notably, the spatial bounding box still correctly highlights the hand performing the unintended action, which is consistent with the correct semantic prediction.

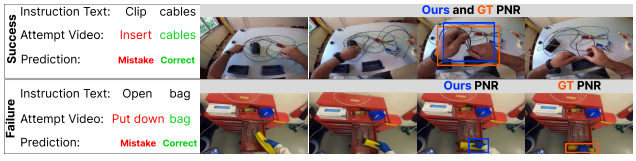

Figure 5. Top: correct predictions (IoU=0.56). Bottom: failure case with correct semantic but incorrect temporal attribution.

E. Implementation Details

Training configuration. We train with AdamW (lr 10^{-4}) for 20 epochs on 4×H100 GPUs with batch size 64.