SAVA-X: Ego-to-Exo Imitation Error Detection via Scene-Adaptive View Alignment and Cross-Fusion

Supplementary Material

A. Implementation Details

A.1. Metric Computation

AUPRC computation. For a fixed temporal IoU threshold, we collect all predicted segments and their confidence scores $\{s_i\}_{i=1}^{N}$ together with binary labels $\{y_i\}_{i=1}^{N}$ indicating whether each prediction is a true positive ($y_i = 1$) or a false positive ($y_i = 0$). Let P denote the total number of positive ground-truth instances. Following the COCO protocol, we first construct a precision–recall curve from these predictions and then approximate the area under the curve (AUPRC) by averaging the interpolated precision values at 101 uniformly sampled recall points in $[0, 1]$. This yields an interpolated average precision, which we report as AUPRC. In our experiments, we apply this procedure separately for different "positive" definitions (e.g., correct vs. error events) by treating each target category as positive and all others as negative.

A.2. Base Encoder

Our base encoder follows the multi-scale temporal convolutional encoder in Wang et al. [5]. Given frame visual features $\mathbf{v} \in \mathbb{R}^{T\times d}$ from the backbone, we first reshape them to a 1D feature map of size $T \times d$ and feed them into a lightweight temporal pyramid. The first level uses a 1×1 Conv1D and GroupNorm to project the input feature dimension to the transformer hidden dimension. Subsequent levels apply 3×3 Conv1D with stride 2 and GroupNorm to progressively downsample the sequence, producing multi-scale features with decreasing temporal resolution and shared hidden dimension. For each level, we add a sine-based temporal positional encoding and pass the resulting feature maps and masks to the Deformable Transformer encoder. This design keeps the hidden sizes and number of feature levels consistent with [5], so that performance gains mainly stem from our cross-view modules rather than changes in the base encoder.

A.3. Deformable DETR

Our event detection head is a 1D multi-scale Deformable DETR that follows the design of Wang et al. [5]. Given the multi-scale temporal features and masks from the base encoder, we first flatten all levels into a single sequence and add level-specific embeddings and temporal positional encodings. The encoder then applies L_{enc} layers of multi-scale deformable self-attention and feed-forward networks to produce a unified feature memory. The decoder takes a fixed set of learnable query embeddings and performs, at each of its L_{dec} layers, (i) self-attention among queries and (ii) multi-scale deformable cross-attention to the encoded memory using normalized reference points, followed by a feed-forward network. As in the original Deformable DETR, we use iterative reference-point refinement via a small MLP head attached to each decoder layer, and share the hidden dimension and attention configuration with [5]. This keeps the detection head identical to the PDVC Deformable DETR implementation, so that the performance gains in our experiments mainly arise from the proposed cross-view modules rather than changes in the underlying transformer architecture.

Algorithm 1 COCO-style AUPRC computation for a fixed tIoU

Require: Scores s_i, labels $y_i \in \{0, 1\}$ for $i = 1, \ldots, N$; number of positives $P = \sum_i y_i$
Ensure: AUPRC value AP

Sort indices π such that $s_{\pi_1} \geq s_{\pi_2} \geq \cdots \geq s_{\pi_N}$
Initialize cumulative true/false positives: $\mathrm{TP}[k] = \sum_{i=1}^{k} \mathbb{1}[y_{\pi_i} = 1]$, $\mathrm{FP}[k] = \sum_{i=1}^{k} \mathbb{1}[y_{\pi_i} = 0]$
for $k = 1$ to N **do**
  $\mathrm{precision}[k] \leftarrow \frac{\mathrm{TP}[k]}{\max(1, \mathrm{TP}[k]+\mathrm{FP}[k])}$
  $\mathrm{recall}[k] \leftarrow \frac{\mathrm{TP}[k]}{\max(1, P)}$
end for
For $k = N - 1$ down to 1:
  $\mathrm{precision}[k] \leftarrow \max\big(\mathrm{precision}[k], \mathrm{precision}[k + 1]\big)$
Initialize $\mathrm{AP} \leftarrow 0$
for $j = 0$ to 100 **do**
  $r_j \leftarrow j/100$ {101 uniformly sampled recall points}
  Find the smallest index k such that $\mathrm{recall}[k] \geq r_j$
  if such k exists **then**
    $p(r_j) \leftarrow \mathrm{precision}[k]$
  else
    $p(r_j) \leftarrow 0$
  end if
  $\mathrm{AP} \leftarrow \mathrm{AP} + p(r_j)$
end for
$\mathrm{AP} \leftarrow \mathrm{AP}/101$
return AP

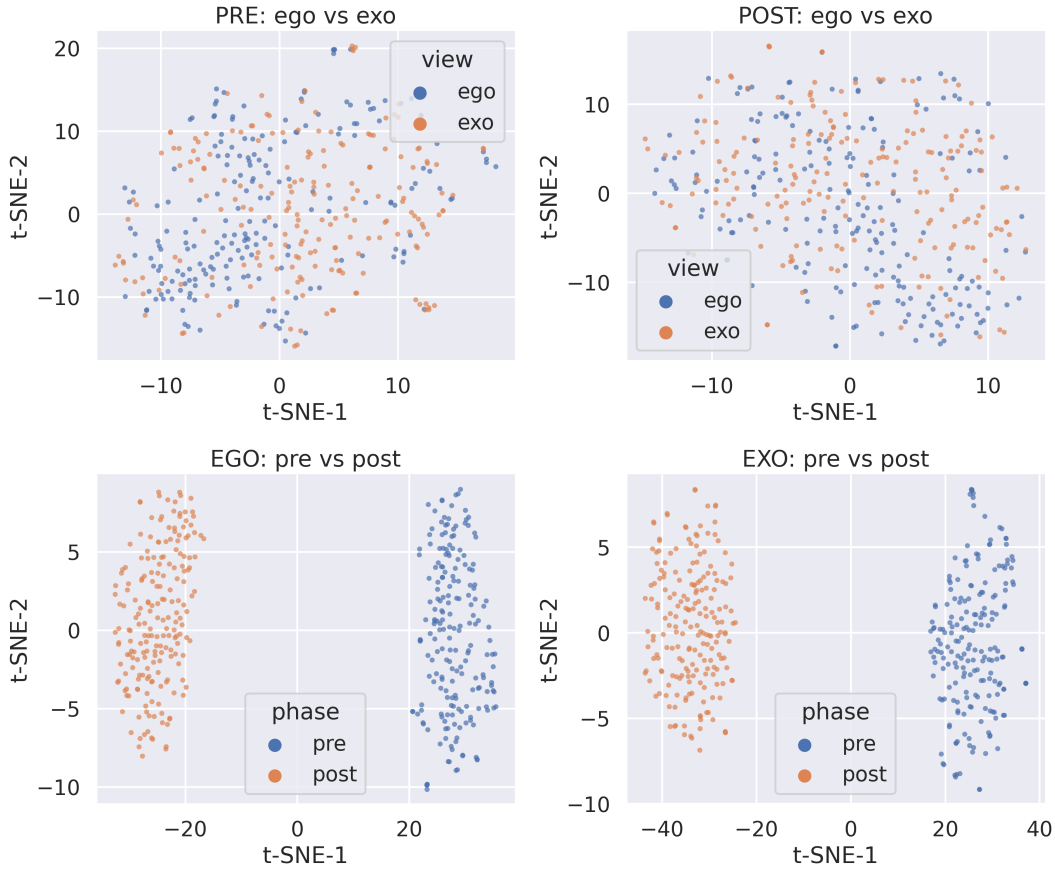

Figure A.1. t-SNE visualization of video-level features before and after SVE. **Top**: ego and exo features colored by view; SVE reduces the cross-view gap and yields more overlapped distributions. **Bottom**: ego (left) and exo (right) features colored by phase (pre vs. post), showing that SVE applies a structured, non-trivial transformation while avoiding representational collapse.

A.4. Task-specific heads

Except for the imitation-error prediction heads introduced in our work, all task heads follow Wang et al. [5]. Given the decoder features for each query, we apply a linear classification head to predict the event category and a regression MLP to predict the normalized start/end coordinates of the temporal segment.

Following PDVC, we additionally use a count head to estimate the number of events and a captioning head to generate a natural-language description for each detected segment, sharing the same hidden dimension and decoder outputs as in [5]. On top of these standard heads, we attach two lightweight binary imitation-error heads: (i) a fine-grained query-level error head that predicts whether each event segment corresponds to a correct or erroneous execution step, and (ii) a global video-level head that aggregates query features to predict the overall imitation quality of the sequence. These additions leave the original PDVC heads unchanged, ensuring that performance gains mainly stem from our cross-view alignment and error modeling rather than modifications to the baseline detection and captioning heads.

A.5. Losses

Dense video captioning loss. Following PDVC [5], we treat dense video captioning as a set prediction problem and supervise all decoder layers with a Hungarian-matching loss. Let $\{\hat{\mathbf{s}}_i, \hat{\mathbf{c}}_i, \hat{\mathbf{y}}_i\}_{i=1}^{N}$ denote the predicted temporal segments (center–length parameterization), foreground scores, and caption word distributions for N event queries in one decoder layer, and let $\{\mathbf{s}_j, \mathbf{y}_j\}_{j=1}^{N_{\text{gt}}}$ be the ground-truth segments and captions. We first solve a bipartite matching between predictions and ground truths with cost

$$\mathcal{C}_{ij} = \alpha_{\text{giou}}\, \mathcal{L}_{\text{giou}}(\hat{\mathbf{s}}_i, \mathbf{s}_j) + \alpha_{\text{cls}}\, \mathcal{L}_{\text{cls}}(\hat{\mathbf{c}}_i, \mathbb{1}[j \le N_{\text{gt}}]) \tag{A.1}$$

where $\mathcal{L}_{\text{giou}}$ is the temporal generalized IoU loss and $\mathcal{L}_{\text{cls}}$ is the focal classification loss between foreground/background. Given the optimal assignment σ,

the DVC loss for one decoder layer is

$$\begin{aligned}
\mathcal{L}_{\text{DVC}} &= \beta_{\text{giou}}\,\mathcal{L}_{\text{giou}} + \beta_{\text{cls}}\,\mathcal{L}_{\text{cls}} + \beta_{\text{ec}}\,\mathcal{L}_{\text{ec}} + \beta_{\text{cap}}\,\mathcal{L}_{\text{cap}} \\
\mathcal{L}_{\text{giou}} &= \frac{1}{N_{\text{gt}}}\sum_{j=1}^{N_{\text{gt}}}\big(1 - \text{GIoU}(\hat{\mathbf{s}}_{\sigma(j)}, \mathbf{s}_j)\big) \\
\mathcal{L}_{\text{cls}} &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\text{focal}}\big(\hat{\mathbf{c}}_i,\, y_i^{\text{fg/bg}}\big) \\
\mathcal{L}_{\text{ec}} &= -\log p_{\text{ec}}\big(N_{\text{gt}}\big) \\
\mathcal{L}_{\text{cap}} &= \frac{1}{N_{\text{gt}}}\sum_{j=1}^{N_{\text{gt}}}\frac{1}{T_j}\sum_{t=1}^{T_j} -\log p\big(w_{j,t} \mid w_{j,<t}, \hat{\mathbf{z}}_{\sigma(j)}\big)
\end{aligned} \tag{A.2}$$

where p_{ec} is the event-counter distribution over event numbers, $\hat{\mathbf{z}}_{\sigma(j)}$ is the matched query feature, and $w_{j,t}$ is the t-th word of the j-th ground-truth caption of length T_j. As in [5], we attach prediction heads to all decoder layers and sum the layer-wise losses.

Imitation-error classification losses. On top of the standard DVC loss, we introduce two lightweight error-prediction heads: a fine-grained query-level head and a global video-level head (Sec. A.4). For each matched query j, let $\hat{z}^{\text{fine}}_{\sigma(j)}$ be the scalar logit from the fine-grained error head and $e_j^{\text{fine}} \in \{0, 1\}$ be the corresponding binary error label (1 for erroneous execution, 0 for correct). We supervise this head with a binary cross-entropy (BCE) loss over the matched events:

$$\begin{aligned}
\mathcal{L}^{\text{fine}}_{\text{err}} = \frac{1}{N_{\text{gt}}}\sum_{j=1}^{N_{\text{gt}}}\Big[&- e_j^{\text{fine}} \log \sigma\big(\hat{z}^{\text{fine}}_{\sigma(j)}\big) \\
&-(1 - e_j^{\text{fine}}) \log\big(1 - \sigma\big(\hat{z}^{\text{fine}}_{\sigma(j)}\big)\big)\Big]
\end{aligned} \tag{A.3}$$

where $\sigma(\cdot)$ is the sigmoid function. In addition, the global head aggregates all valid query features of a video into a single logit $\hat{z}^{\text{overall}}$ that predicts the overall imitation quality, with binary label $e^{\text{overall}} \in \{0, 1\}$ (error-free vs. containing errors). We apply another BCE loss:

$$\begin{aligned}
\mathcal{L}^{\text{overall}}_{\text{err}} = &-e^{\text{overall}} \log \sigma\big(\hat{z}^{\text{overall}}\big) \\
&-(1 - e^{\text{overall}}) \log\big(1 - \sigma\big(\hat{z}^{\text{overall}}\big)\big)
\end{aligned} \tag{A.4}$$

Adaptive sampling regularization. We further regularize the adaptive sampler (AS) to avoid collapsed selections and redundant representations. Denote by $\{s_{x,t}\}_{t=1}^{T_x}$ and $\{s_{y,t}\}_{t=1}^{T_y}$ the normalized selection probabilities over Ego/Exo tokens. We add a selection-entropy regularizer [3] that encourages coverage instead of concentrating mass on a few positions:

$$\begin{aligned}
\mathcal{L}_{\text{sel}} = &\frac{1}{\log T_x}\sum_{t=1}^{T_x} s_{x,t}\,\log\big(s_{x,t} + \varepsilon\big) \\
&+\frac{1}{\log T_y}\sum_{t=1}^{T_y} s_{y,t}\,\log\big(s_{y,t} + \varepsilon\big)
\end{aligned} \tag{A.5}$$

where $\varepsilon > 0$ is a small constant for numerical stability. Let $u \in \{\text{exo}, \text{ego}\}$ index the view and $\hat{\mathbf{Z}}^u \in \mathbb{R}^{K_u \times d}$ be the matrix of gated active tokens (after selection and gating), with K_u tokens and feature dimension d. We further attach VICReg-style [1] variance and covariance penalties to suppress collapse and dimensional collinearity:

$$\begin{aligned}
\boldsymbol{\mu}^u &= \frac{1}{K_u}\sum_{i=1}^{K_u}\hat{\mathbf{Z}}^u_i, \quad \hat{\mathbf{Z}}^u_{\text{c}} = \hat{\mathbf{Z}}^u - \mathbf{1}\,\boldsymbol{\mu}^{u\top} \\
\mathcal{L}^u_{\text{var}} &= \frac{1}{d}\sum_{j=1}^{d}\Big[\max\big(0,\, \gamma - \sqrt{\text{Var}(\hat{\mathbf{Z}}^u_{\text{c},j}) + \varepsilon}\big)\Big]^2 \\
\mathbf{C}^u &= \frac{1}{K_u - 1}\,\hat{\mathbf{Z}}^{u\top}_{\text{c}}\hat{\mathbf{Z}}^u_{\text{c}}, \quad \mathcal{L}^u_{\text{cov}} = \frac{1}{d}\sum_{\substack{i=1\\ i\neq j}}^{d}\sum_{j=1}^{d}\big(\mathbf{C}^u_{ij}\big)^2 \\
\mathcal{L}_{\text{vic}} &= \mathcal{L}^{\text{exo}}_{\text{var}} + \mathcal{L}^{\text{ego}}_{\text{var}} + \mathcal{L}^{\text{exo}}_{\text{cov}} + \mathcal{L}^{\text{ego}}_{\text{cov}}
\end{aligned} \tag{A.6}$$

where $\gamma > 0$ is the variance lower bound and $\varepsilon > 0$ again ensures numerical stability.

B. Results

B.1. Correct Class Results

Table A.1 summarizes the AUPRC performance for the *correct* (non-error) class on EgoMe validation and test splits.

Among DVC-style baselines, PDVC [5] remains strong, while Exo2EgoDVC [2] does not consistently improve over PDVC despite explicitly transferring exocentric knowledge. TAL-only models (ActionFormer [6], TriDet [4]) perform clearly worse on AUPRC and tIoU, showing that off-the-shelf localization architectures are not sufficient for our fine-grained imitation setting.

Using only egocentric input further degrades PDVC, highlighting the benefit of multi-view information even when evaluating the correct class.

Our SAVA-X achieves the best validation performance across all tIoU thresholds and mean AUPRC, and attains comparable or better test performance than PDVC, with particularly noticeable gains at high tIoU (e.g., +1.1 AUPRC@0.7 on validation). These results indicate that SAVA-X not only improves error detection, but also preserves or slightly enhances recognition and

Method	AUPRC on Validation					AUPRC on Test				
	0.3	0.5	0.7	Mean	tIoU	0.3	0.5	0.7	Mean	tIoU
Dense Video Captioning (DVC) baselines										
PDVC [5]	68.41	46.21	12.72	42.45	58.58	**66.53**	43.88	12.30	40.90	57.98
Exo2EgoDVC [2]	67.52	44.52	12.43	41.49	59.06	65.37	42.27	11.40	39.68	58.15
Temporal Action Localization (TAL) baselines										
ActionFormer [6]	65.87	31.68	4.54	34.03	48.89	63.25	29.20	4.17	32.20	48.25
TriDet [4]	65.05	32.25	4.35	33.88	49.05	62.45	30.29	4.30	32.35	49.02
Only Egocentric Input										
PDVC [5]	64.77	42.11	10.94	39.27	57.63	63.94	40.96	12.03	38.98	57.19
Ours										
SAVA-X	**69.02**	**46.58**	**13.85**	**43.15**	**59.31**	66.32	**43.98**	**12.41**	**40.90**	**58.32**

Table A.1. Comparison on EgoMe validation and test split. Left: results on *validation set*. Right: results on the *test set*. We report AUPRC for the correct class at multiple tIoU thresholds (0.3, 0.5, 0.7), their mean, and standalone temporal IoU (tIoU) for localization quality.

localization of correct executions, rather than trading off one class for the other.

B.2. TSNE

To better understand how SVE reshapes the feature space, we visualize pre- and post-SVE video-level features using t-SNE (Fig. A.1).

In the top row, we color points by view (ego vs. exo). Before SVE (top-left), ego and exo features form two partially separated clouds with a clear domain shift.

After SVE (top-right), the two distributions become much more interleaved, indicating that SVE effectively reduces the cross-view gap and encourages a more view-invariant embedding. In the bottom row, we fix the view and color points by phase (pre vs. post).

For both ego (bottom-left) and exo (bottom-right), pre- and post-SVE features form two well-separated clusters along the first t-SNE dimension, showing that SVE applies a non-trivial, structured transformation to the representations rather than a small perturbation.

Together, these plots suggest that SVE consistently aligns ego and exo distributions while preserving meaningful intra-view variability and avoiding collapse.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 3

[2] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8324–8335. IEEE, 2025. 3, 4

[3] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations*, 2017. 3

[4] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 3, 4

[5] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-End Dense Video Captioning with Parallel Decoding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6827–6837, Montreal, QC, Canada, 2021. IEEE. 1, 2, 3, 4

[6] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–510. Springer, 2022. 3, 4