Generic Event Boundary Detection via Denoising Diffusion
- Supplementary Material -

Algorithm 1 DiffGEBD training algorithm

```
def train_loss(V, T, y_0, p):
   """
   V: video [B, T, H, W, 3]
   T: diffusion time-step
   y_0: ground-truth boundary labels [B, L, 1]
   p: CFG probability
   """

   # Extract features from backbone network g
   F = g(V)

   # Extract visual embeddings from the encoder f
   E = f(F)

   # Random sample for time-step
   t = uniform(0, T)

   eps = normal(mean=0, std=1)

   # Corrupt data
   y_crpt = sqrt(   alpha_cumprod(t)) * y_0 +
            sqrt(1 - alpha_cumprod(t)) * eps

   # Classifier-free Guidance by probability p
   if uniform(0, 1) < p:
      E = zeros_like(E)

   # Predict with the decoder h
   y_hat = h(y_0, E, t)

   # Mean squared loss
   loss = (y_0 - y_hat)**2
   loss = mean(loss)

   return loss
```

`alpha_cumprod(t)`: cumulative product of α_i, *i.e.*, $\prod_{i=1}^{t} \alpha_i$

Algorithm 2 DiffGEBD inference algorithm

```
def inference(V, T, steps, w):
   """
   V: video [B, T, H, W, 3]
   T: diffusion time step
   steps: the number of inference steps
   w: classifier-free guidance weight
   """

   # Extract features from backbone network g
   F = g(V)

   # Extract visual embeddings from the encoder f
   E = f(F)

   y_t = normal(mean=0, std=1)

   # Uniform sample step size
   times = reversed(linespace(-1, T, steps))
   time_pairs = list(zip(times[:-1], times[1:]))

   for t_now, t_next in zip(time_pairs):
      # conditional prediction
      y_hat_c = h(y_t, E, t_now)

      # unconditional prediction
      y_hat_u = h(y_t, zeros_like(E), t_now)

      # Form the classifier-free guided prediction
      y_hat = (1 + w) * y_hat_c - w * y_hat_u

      # Estimate x at t_next
      y_t = ddim_step(y_t, y_hat, t_now, t_next)

   return y_t
```

In this supplementary material, we present detailed explanations and additional experimental results. Specifically, we include the training and inference algorithms in Sec. 1, experimental details in Sec. 2, additional experimental results in Sec. 3, more example results in Sec. 4, and a discussion in Sec. 5.

1. Algorithms

We present the training and inference algorithms in Alg. 1 and Alg. 2, respectively. During training, both conditional and unconditional models are jointly trained with probability p, enabling classifier-free guidance. During inference, we iteratively refine the output by balancing the unconditional and conditional outputs according to the guidance weight w, obtaining the final output.

2. Experimental Details

2.1. Datasets

Kinetics-GEBD. Kinetics-GEBD [14] is the largest GEBD dataset, encompassing a wide spectrum of videos. Each boundary is composed of various taxonomy-free boundaries, including action and object changes. The dataset includes multiple annotators, with each annotation providing subjective event boundaries. Each of the training and validation set contains 20K videos from Kinetics-400 [5]. In our experiments, we report the results on the validation set.

TAPOS. The TAPOS dataset [13] comprises 21 distinct action categories derived from Olympic sports videos. It consists of 13,094 action instances in the training set and 1,790 instances in the validation set. Each video is annotated with a single annotator, which divides a single action into multiple sub-actions. Following [14], we adapt TAPOS for our GEBD task by trimming each action instance with its action label hidden and conducting experiments on them.

2.2. Implementation details

We train our model using AdamW with a batch size of 2 and a learning rate of 2e-5 in all experiments. In determining the final boundary predictions, we identify consecutive predictions that exceed a predefined threshold δ as bound-

Model	Diffusion	Diversity-aware GEBD				Conventional GEBD
		$F1_{sym}$	$F1_{p2g}$	$F1_{g2p}$	Diversity	F1@0.05
cVAE	-	62.7	66.8	59.9	15.2	70.0
DiffGEBD	-	73.4	75.2	72.3	20.2	77.5
DiffGEBD	✓	**74.0**	**75.6**	**72.9**	**20.4**	**78.4**

Table 1. **Effects of the diffusion process.** The diffusion-based approach consistently outperforms baselines across all metrics.

Model	Sampler	Diversity-aware GEBD				Conventional GEBD
		$F1_{sym}$	$F1_{p2g}$	$F1_{g2p}$	Diversity	F1@0.05
DiffGEBD	DPM-Solver++	73.8	**76.0**	72.2	18.0	78.2
DiffGEBD	UniPC	73.8	76.1	72.2	17.7	78.4
DiffGEBD	DDIM	**74.0**	75.6	**72.9**	**20.4**	**78.4**

Table 2. **Effect of the diffusion sampler.**

Model	$F1_{sym}$	Div.	Train time	Inf. time	Mem.	#param
Temporal Perciever [18]	69.4	14.6	8.7h	**0.03s**	**0.1G**	52.2M
SC-Transformer [7]	72.9	18.9	44.7h	0.15s	9.4G	71.6M
BasicGEBD [22]	72.2	18.6	19.9h	0.15s	7.2G	**32.2M**
EfficientGEBD [22]	72.6	14.9	16.4h	0.15s	6.2G	33.2M
DiffGEBD (4 steps)	73.4	18.5	6.6h	0.16s	7.1G	68.0M
DiffGEBD (8 steps)	73.7	19.4	6.6h	0.19s	7.1G	68.0M
DiffGEBD (32 steps)	**74.0**	**20.4**	**6.6h**	0.36s	7.1G	68.0M

Table 3. **Computational cost on Diversity-aware evaluation.** We report the training time and parameters per whole video and the inference time and memory per frame.

ary candidates. The midpoint of each boundary candidate sequence is then designated as the final boundary prediction [7, 22]. We set δ to 0.5 for both Kinetics-GEBD [14] and 0.3 for TAPOS [13] in our experiments.

3. Additional Experimental Results

We present additional experimental results following the same settings as in the main paper. All experiments were conducted on the Kinetics-GEBD dataset.

Effect of diffusion process. We evaluate our diffusion-based approach against two primary baselines at Table 1: a non-diffusion method training multiple times, and a CVAE-based [15] model. Our method outperforms these alternatives, as diffusion models are inherently capable of accurately approximating complex data distributions, which led to our superior performance.

Effect of the different samplers. We adopt DPM-Solver++ [11] and UniPC [21] samplers for diffusion inference. As in Table 2, the performance remains consistent across different samplers, showing its generalizability.

Computational cost on the diversity-aware evaluation. Table 3 compares the computational efficiency of our method with others reported in Table 1. All results are obtained using RTX 6000 Ada GPU under the same settings. For deterministic methods, we report the total training time across five runs. Temporal Perciever [18] uses pre-extracted features without end-to-end backbone fine-tuning, which explains its lower Inf. time and memory usage. Compared to EfficientGEBD, our method achieves substantially lower training time and comparable inference time with 4 sampling steps, while still outperforming it. Although inference time increases with more steps, it can be reduced via recent advances, *e.g.*, Flow Matching [10], which we leave for future work. In terms of memory footprint and model size, our method maintains similar memory usage to other baselines using a moderate number of parameters, offering a good balance between efficiency and capacity.

Method	D^2_{GED} ($\downarrow$)
Temporal Perceiver† [18]	45.5
SC-Transformer† [7]	36.7
BasicGEBD† [22]	37.0
EfficientGEBD† [22]	38.6
DiffGEBD (ours)	**34.8**

Table 4. **Generalized energy distance on Kinetics-GEBD.** † Models with dagger marks are reproduced.

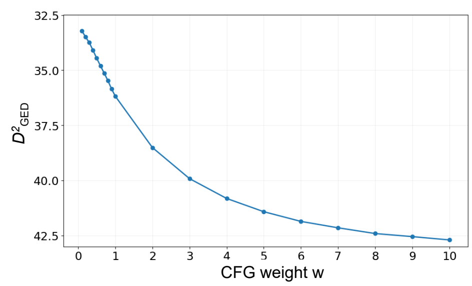

Figure 1. **Effects of CFG weight w on GED.**

Generalized energy distance (GED). Additionally, we employ the Generalized Energy Distance (D^2_{GED}) [1, 12, 17] on the diversity-aware evaluation. D^2_{GED} measures the discrepancy between the predicted distributions $\hat{\boldsymbol{Y}}$ and the ground truth boundary distributions $\boldsymbol{Y}$:

$$D^2_{\text{GED}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = 2\mathbb{E}[d(\hat{Y}, Y)] - \mathbb{E}[d(\hat{Y}, \hat{Y}')] - \mathbb{E}[d(Y, Y')], \tag{1}$$

where d is a distance metric, $\hat{Y}, \hat{Y}'$ are independent samples drawn from the predicted distribution $\hat{\boldsymbol{Y}}$, and Y, Y' are independent samples drawn from the ground truth distribution $\boldsymbol{Y}$. We adopt $d(i,j) = 1 - \text{F1@}\tau(i,j)$ to evaluate boundary matching score, for arbitrary i and j. Here, we set τ to 0.05. A lower GED score indicates better alignment between predicted and ground truth distributions. The detailed computation is provided in Eq. 2.

Method	Threshold δ								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DDM-Net [19]	74.9	75.3	75.7	76.3	76.8	76.0	67.1	43.0	16.7
SC-Transformer† [7]	70.9	76.3	77.6	77.3	76.2	74.1	69.6	60.3	37.0
EfficientGEBD [22]	51.2	70.5	78.3	75.5	65.2	50.4	34.7	20.2	7.9
DiffGEBD(Ours)	**78.3**	**78.4**	**78.4**	**78.4**	**78.4**	**78.4**	**78.4**	**78.4**	**78.4**

Table 5. **Robustness on threshold δ.** We report F1@0.05 with different thresholds on Kinetics-GEBD. †: reproduced from the official code.

Method	Reproduced	Paper
Temporal Perceiver† [18]	74.9	74.8
SC-Transformer† [7]	77.4	77.7
BasicGEBD† [22]	76.9	76.8
EfficientGEBD† [22]	78.3	78.3

Table 6. **F1@0.05 of conventional evaluation protocol on Kinetics-GEBD.** † Models with dagger marks are reproduced using official implementations.

$$
\begin{aligned}
D^2_{GED}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = & \frac{2}{N_p N_g} \sum_{i=1}^{N_p} \sum_{j=1}^{N_g} d(\hat{Y}_i, Y_j) \\
& - \frac{1}{{N_p}^2} \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} d(\hat{Y}_i, \hat{Y}_j) \\
& - \frac{1}{{N_g}^2} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} d(Y_i, Y'_j)
\end{aligned} \tag{2}
$$

Following the diversity-aware evaluation protocol, we evaluate our model using D^2_{GED}. As presented in Table 4, the results demonstrate that our predicted distributions closely align with the ground truth boundary distributions. Figure 1 shows that the CFG weight w increases, the D^2_{GED} increases, indicating that stronger guidance reduces the diversity of predictions and leads to larger discrepancy between predicted and ground truth distributions.

Robustness on threshold δ. In boundary detection, the final boundary prediction $\hat{\boldsymbol{y}}_0$ for each frame is thresholded by δ to determine whether it is classified as a boundary. To assess the robustness of the predictions, we vary δ from 0.1 to 0.9. As shown in Table 5, DiffGEBD maintains consistently strong performance across different threshold values, demonstrating the robustness of the predicted boundaries.

Reproduced results of previous methods. For diversity-aware evaluation protocol, we conduct 5 independent runs for each model. As shown in Table 6, the average performances of the conventional protocol closely match the reported performance in previous methods, validating the reproducibility.

CFG weight w	F1_{sym}	F1_{p2g}	F1_{g2p}	Diversity
0.1	73.45	74.24	73.14	**24.64**
0.2	73.63	74.60	73.16	23.64
0.3	73.79	74.94	73.17	22.75
0.4	73.87	75.19	73.11	21.88
0.5	73.93	75.42	73.03	21.09
0.6	**73.96**	75.60	72.92	20.38
0.7	**73.96**	**75.76**	72.79	19.73
0.8	73.93	75.87	72.65	19.24
0.9	73.91	75.97	72.53	18.57
1.0	73.85	76.04	72.37	18.07
2.0	73.42	76.42	71.25	14.91
3.0	73.02	76.48	70.49	13.28
4.0	72.73	**76.49**	69.97	12.31
5.0	72.52	76.48	69.58	11.70
6.0	72.35	76.44	69.32	11.29
7.0	72.21	76.39	69.12	11.02
8.0	72.08	76.34	68.93	10.82
9.0	72.00	76.31	68.80	10.69
10.0	71.91	76.26	68.68	10.60

Table 7. **Numerical results of the effect of CFG weight w (Fig. 4).**

N_{G}	F1_{sym}	F1_{p2g}	F1_{g2p}	Diversity
1	70.9	73.9	68.8	15.1
2	72.1	74.3	70.6	17.6
3	73.5	75.5	72.1	18.4
4	**74.0**	**75.6**	**72.9**	20.4
5	73.0	74.4	72.4	**22.9**

Table 8. **Numerical results of the effect of number of annotations (Fig. 5).**

Numerical results of CFG weight w. Table 7 presents the complete numerical results in Fig. 4 in the main paper. While the main paper visualizes these results as plots for better trend analysis, we provide the exact values here for reference.

Numerical results of number of annotations. The numerical results of Fig. 5 are presented in Table 8.

Full results on the diversity-aware evaluation. We provide full results with Rel. Dis. threshold ranging from 0.05 to 0.5 on the diversity-aware evaluation protocol. Table 9 presents the F1_{sym} performance across all thresholds. DiffGEBD outperforms previous methods across all Rel. Dis. thresholds.

Full results on the conventional evaluation. We provide full results with Rel. Dis. threshold ranging from 0.05 to 0.5 on the conventional evaluation protocol. Table 10 and

Table 11 show the results of Kinetics-GEBD and TAPOS, respectively.

4. More Example Results

We provide additional qualitative results in Fig 2. The model demonstrates robust detection of boundaries with significant scene changes across all guidance weights. However, for subtle transitions, such as minor object movements observed at 1.70s (2b), the model becomes less sensitive to these boundaries at higher weights (2c). This suggests that lower guidance weights enable the model to capture ambiguous boundaries through its stochastic generation process.

5. Discussion

Limitations and future work. While our diffusion-based method effectively generates multiple predictions, its iterative process significantly slows down inference. Future work will address this limitation by adapting methods like Flow Matching [10] and Consistency Models [16]. These approaches can achieve high-quality results with a single sampling step, directly addressing the speed limitations.

Broader impact. To the best of our knowledge, this work presents the first generative formulation of generic event boundary detection, along with a novel evaluation framework for multiple predictions scenario. We believe that our approach opens up new possibilities for for addressing inherent human ambiguity in event boundaries and provides a new paradigm for future research in this direction.

Method	F1$_{sym}$ @ Rel. Dis.										
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg.
Temporal Perceiver† [18]	69.4	76.9	79.3	80.7	81.6	82.2	82.6	83.0	83.3	83.5	80.2
SC-Transformer† [7]	<u>72.9</u>	<u>80.7</u>	<u>83.1</u>	<u>84.5</u>	<u>85.3</u>	85.9	86.4	86.7	87.0	87.2	<u>84.0</u>
BasicGEBD† [22]	72.2	79.7	82.2	83.6	84.6	85.2	85.6	86.0	86.2	86.5	83.2
EfficientGEBD† [22]	72.6	80.3	82.8	84.3	<u>85.3</u>	<u>86.0</u>	<u>86.5</u>	<u>86.9</u>	<u>87.2</u>	<u>87.5</u>	83.9
DiffGEBD (ours)	**74.0**	**81.8**	**84.2**	**85.5**	**86.4**	**87.0**	**87.4**	**87.8**	**88.1**	**88.4**	**85.1**

Table 9. **Diversity-aware evaluation on Kinetics-GEBD with Rel.Dis. threshold from 0.05 to 0.5.** We report F1$_{\mathbf{sym}}$ score varying different relative distance thresholds. Bold numbers indicate the best score, while underlined numbers represent the second-best performance.

Method	F1 @ Rel. Dis.										
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg.
BMN [9]	18.6	20.4	21.3	22.0	22.6	23.0	23.3	23.7	23.9	24.1	22.3
BMN-StartEnd [9]	49.1	58.9	62.7	64.8	66.0	66.8	67.4	67.8	68.1	68.3	64.0
TCN [6]	58.8	65.7	67.9	69.1	69.8	70.3	70.6	70.8	71.0	71.2	68.5
PC [14]	62.5	75.8	80.4	82.9	84.4	85.3	85.9	86.4	86.7	87.0	81.7
SBoCo [4]	73.2	82.7	85.3	87.7	88.2	89.1	89.4	89.9	89.9	90.7	86.6
Temporal Perceiver [18]	74.8	82.8	85.2	86.6	87.4	87.9	88.3	88.7	89.0	89.2	86.0
DDM-Net [19]	76.4	84.3	86.6	88.0	88.7	89.2	89.5	89.8	90.0	90.2	87.3
CVRL [8]	74.3	83.0	85.7	87.2	88.0	88.6	89.0	89.3	89.6	89.8	86.5
LCVS [20]	76.8	84.8	87.2	88.5	89.2	89.6	89.9	90.1	90.3	90.6	87.7
SC-Transformer [7]	77.7	84.9	87.3	88.6	89.5	90.0	90.4	90.7	90.9	91.1	88.1
BasicGEBD [22]	76.8	83.4	85.7	87.1	87.9	88.5	88.8	89.1	89.4	89.6	86.6
EfficientGEBD [22]	78.3	<u>85.1</u>	<u>87.4</u>	<u>88.7</u>	<u>89.6</u>	<u>90.1</u>	<u>90.5</u>	<u>90.8</u>	<u>91.1</u>	<u>91.3</u>	<u>88.3</u>
DyBDet [23]	**79.6**	**85.8**	**88.0**	**89.3**	**90.1**	**90.7**	**91.1**	**91.5**	**91.7**	**91.9**	**89.0**
DiffGEBD (ours)	<u>78.4</u>	84.8	86.8	87.9	88.6	89.1	89.4	89.7	89.9	90.1	87.5

Table 10. **Comparison with the state of the art on Kinetics-GEBD**. We report F1 score varying different relative distance thresholds. The numbers in boldface indicate the highest score. DiffGEBD shows the competitive performance on overall metrics.

Method	F1 @ Rel. Dis.										
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	avg.
ISBA [2]	10.6	17.0	22.7	26.5	29.8	32.6	34.8	36.9	38.2	39.6	30.2
TCN [6]	23.7	31.2	33.1	33.9	34.2	34.4	34.7	34.8	34.8	34.8	33.0
CTM [3]	24.4	31.2	33.6	35.1	36.1	36.9	37.4	38.1	38.3	38.5	35.0
TransParser [13]	23.9	38.1	43.5	47.5	50.0	51.4	52.7	53.4	54.0	54.5	47.4
PC [14]	52.2	59.5	62.8	64.7	66.0	66.6	67.2	67.6	68.0	68.4	64.3
Temporal Perceiver [18]	55.2	66.3	71.3	73.8	75.7	76.5	77.4	77.9	**78.4**	**78.8**	73.2
DDM-Net [19]	60.4	68.1	71.5	73.5	74.7	75.3	75.7	76.0	76.3	76.7	72.8
SC-Transformer [7]	61.8	69.4	72.8	74.9	76.1	76.7	77.1	77.4	77.7	78.0	74.2
BasicGEBD [22]	60.0	66.6	-	-	-	73.1	-	-	-	74.8	71.0
EfficientGEBD [22]	63.1	70.5	-	-	-	**77.4**	-	-	-	78.6	74.8
DyBDet [23]	62.5	70.1	73.4	75.6	**76.7**	77.2	**77.5**	**77.9**	78.1	78.4	74.7
DiffGEBD (ours)	**65.8**	**71.8**	**74.1**	**75.7**	<u>76.4</u>	77.0	<u>77.4</u>	77.7	78.0	78.1	**75.2**

Table 11. **Comparison with the state of the art on TAPOS.** We report F1 score varying different relative distance thresholds. The numbers in boldface indicate the highest score. DiffGEBD shows the state-of-the-art performance on overall metrics.

Figure 2. **Example results on Kinetics-GEBD.** The figure illustrates (a) Ground truth annotations, (b) predictions with $w = 0.3$, and (c) predictions with $w = 7.0$.

References

[1] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. 2

[2] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6508–6516, 2018. 5

[3] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 137–153. Springer, 2016. 5

[4] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20073–20082, 2022. 5

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[6] Colin Lea, Austin Reiter, René Vidal, and Gregory D Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 36–52. Springer, 2016. 5

[7] Congcong Li, Xinyao Wang, Dexiang Hong, Yufei Wang, Libo Zhang, Tiejian Luo, and Longyin Wen. Structured context transformer for generic event boundary detection. *arXiv preprint arXiv:2206.02985*, 2022. 2, 3, 5

[8] Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, and Libo Zhang. End-to-end compressed video representation learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13967–13976, 2022. 5

[9] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5

[10] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*. 2, 4

[11] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025. 2

[12] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018. 2

[13] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Intra-and inter-action understanding via temporal action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 730–739, 2020. 1, 2, 5

[14] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8075–8084, 2021. 1, 2, 5

[15] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2

[16] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023. 4

[17] Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. 2

[18] Jing Tan, Yuhong Wang, Gangshan Wu, and Limin Wang. Temporal perceiver: A general architecture for arbitrary boundary detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 5

[19] Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, and Limin Wang. Progressive attention on multi-level dense difference maps for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3355–3364, 2022. 3, 5

[20] Libo Zhang, Xin Gu, Congcong Li, Tiejian Luo, and Heng Fan. Local compressed video stream learning for generic event boundary detection. *International Journal of Computer Vision*, 132(4):1187–1204, 2024. 5

[21] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023. 2

[22] Ziwei Zheng, Zechuan Zhang, Yulin Wang, Shiji Song, Gao Huang, and Le Yang. Rethinking the architecture design for efficient generic event boundary detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1215–1224, 2024. 2, 3, 5

[23] Ziwei Zheng, Lijun He, Le Yang, and Fan Li. Fine-grained dynamic network for generic event boundary detection. In *European Conference on Computer Vision*, pages 107–123. Springer, 2025. 5