Generic Event Boundary Detection via Denoising Diffusion

Anonymous ICCV submission

Paper ID 0018

Abstract

Generic event boundary detection (GEBD) aims to identify natural boundaries in a video, segmenting it into distinct and meaningful chunks. Despite the inherent subjectivity of event boundaries, previous methods have focused on deterministic predictions, overlooking the diversity of plausible solutions. In this paper, we introduce a novel diffusion-based boundary detection model, dubbed DiffGEBD, that tackles the problem of GEBD from a generative perspective. The proposed model encodes relevant changes across adjacent frames via temporal self-similarity and then iteratively decodes random noise into plausible event boundaries being conditioned on the encoded features. Classifier-free guidance allows the degree of diversity to be controlled in denoising diffusion. In addition, we introduce a new evaluation metric to assess the quality of predictions considering both diversity and fidelity. Experiments show that our method achieves strong performance on two standard benchmarks, TAPOS and Kinetics-GEBD, generating diverse and plausible event boundaries.

1. Introduction

Through the intricate workings of visual perception, humans can effortlessly detect and interpret a wide range of changes in subjects, objects, and scenes. Research in cognitive science demonstrates that the human visual system easily divides a temporal sequence of images into units of semantic significance [49]. The task of generic event boundary detection (GEBD) has recently been proposed to identify these natural event boundaries in a similar spirit [20, 22–25, 35, 39, 40]. While conventional video tasks in computer vision, such as action recognition [3, 5, 41, 42, 44, 45], temporal action detection [19, 47, 51, 52, 57], and temporal action segmentation [9, 11, 26, 48] mainly focus on identifying class labels or boundaries of predefined action classes, GEBD aims to localize more generic and class-agnostic event boundaries from a video.

Since generic event boundaries are inherently subjective and variable, the problem of GEBD needs to consider the diversity of human judgment; perception of these boundaries can differ significantly among individuals, leading to variation in how people identify boundaries. To account for the human subjectivity, Kinetics-GEBD [35], the benchmark dataset for GEBD, provides multiple annotations from human annotators for each video. However, previous methods [20, 22–25, 35, 39, 40] all focus on predicting accurate boundaries with a deterministic model for a given video, ignoring the diversity of potential solutions.

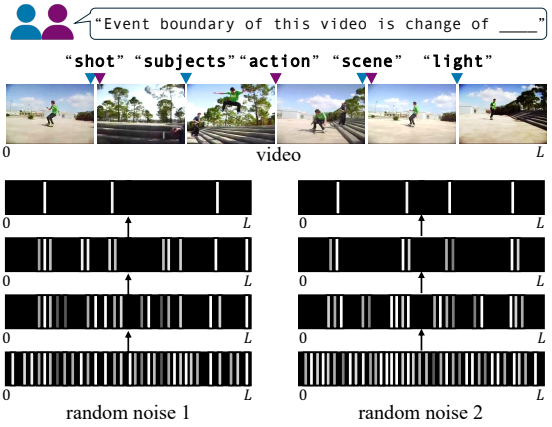

Figure 1. **Illustration of our approach.** Our method generates diverse and plausible boundary predictions for generic events via denoising diffusion.

In this paper, we introduce a diffusion-based boundary detection model, dubbed DiffGEBD, which formulates the problem of GEBD from a generative perspective (Fig. 1). The proposed method consists of a temporal self-similarity encoder and a denoising decoder, where the encoder captures dynamic visual changes across adjacent frames using temporal self-similarity [55], and then the decoder iteratively denoises random noise into plausible event boundaries being conditioned on the encoded features. Since our model outputs distinct predictions from different noises, we control the prediction diversity incorporating classifier-free guidance (CFG) [15] into our denoising process. By controlling the guidance weight, DiffGEBD effectively perform diverse yet accurate boundary detection while better reflect-

ing human judgment variability.

Given our model's ability to generate diverse predictions, a key challenge emerges in how to properly evaluate them. Conventional GEBD evaluation metric, F1 score, measures the alignment of a single prediction only against multiple GT annotations and thus fails to capture both the many-to-many alignments when a model outputs multiple predictions and the diversity among the predictions themselves. To address this limitation, we propose a novel diversity-aware evaluation protocol introducing two metrics: symmetric F1 and diversity scores. The symmetric F1 effectively captures many-to-many alignments between two sets of predictions and GT annotations, while the diversity score directly measures diversity between predictions, enabling comprehensive GEBD evaluation that reflects the inherent variability of event boundaries.

Our contributions can be summarized as follows: **1)** we introduce a novel diffusion-based event boundary detection model, dubbed DiffGEBD, formulating GEBD from a generative perspective. **2)** CFG is employed to control the degree of diversity in denoising diffusion. **3)** We propose a novel diversity-aware evaluation protocol introducing two metrics: symmetric F1 and diversity scores. **4)** DiffGEBD achieves strong performance on standard benchmark datasets, TAPOS and Kinetics-GEBD, generating diverse and plausible event boundaries.

2. Related Work

Generic event boundary detection. GEBD [35] is a video boundary detection task that segments a video into units of events, similar to how humans perceive and distinguish events in a video. Each event divides the video into shorter, taxonomy-free segments compared to traditional criteria for video segmentation. Existing approaches have primarily focused on how to effectively utilize visual information for boundary detection. UBoCo [20] propose the temporal self-similarity matrix (TSM) for capturing semantic inconsistency existing at video boundaries. Building on this insight, subsequent works have proposed various extensions. DDM-Net [40] proposes the progressive attention to fuse spatial features and temporal similarities. LCVS [53] further enhances boundary detection by incorporating motion vectors with similarity features. Recent approaches [23, 55, 56] improve the effectiveness of TSM by applying it in a sliding window manner over local temporal regions. All of the previous methods deterministically detect boundaries based on visual features. In this paper, we introduce a generative perspective to the task, which enables diverse and plausible boundary detections.

Diverse prediction. Generating diverse predictions has become important in future action anticipation [1, 50] and ambiguous segmentation tasks [21, 30], where predictions are inherently uncertain. UAAA [1] proposes a framework that models probability distributions and generates multiple samples corresponding to different possible sequences of future activities. GTDA [50] leverages diffusion models to capture the distribution of activities and propose a metric for measuring the diversity of generated samples. In medical image segmentation, multiple expert annotations often lead to ambiguity. Probabilistic U-Net [21] addresses this challenge by learning to capture the distribution of annotations. They propose using Generalized Energy Distance (GED) [4, 33, 38] to evaluate the similarity between the distribution of predicted samples and annotations. Going further, we propose a mechanism to directly control the degree of diversity, along with metrics to evaluate such controlled diverse predictions.

Diffusion model. Diffusion [36] is a generative algorithm that models data distribution, inspired by non-equilibrium statistical physics [18] and sequential Monte Carlo methods [27]. They gradually transform the data distribution into a Gaussian distribution and then reverse this process to recover the data distribution from a random Gaussian distribution. Diffusion models have shown impressive interest in image generation [16, 37]. Following the advent of class-conditional diffusion models [8, 15], high-resolution diffusion [28, 31, 32] models leveraging text conditions have been developed. Additionally, diffusion models have made significant impacts in various generative fields. Diffusion models have recently been applied to the field of human perception. They have been utilized for image segmentation [2, 46], object detection [6] and video understanding [12, 26]. These models leverage visual features as diffusion conditions to generate accurate labels. Similarly, we address boundary detection by conditioning on video features in a generative manner.

3. Preliminary

In this section, we provide the background of diffusion models [16, 36]. Diffusion models consists of two main components: the forward process, which progressively adds Gaussian noise to the data, and the reverse process, which reconstructs the original data by iteratively denoising.

Forward Process. The forward process involves adding small Gaussian noise ϵ over T timesteps to create noisy data from the given real data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$. The size of the noise added during the forward process is controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, \alpha_t\mathbf{I}), \quad (1)$$

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}). \quad (2)$$

By the Markovian chain, this can be formulated as follows:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_0}\mathbf{x}_0, (1 - \bar{\alpha}_0\boldsymbol{I})). \quad (3)$$

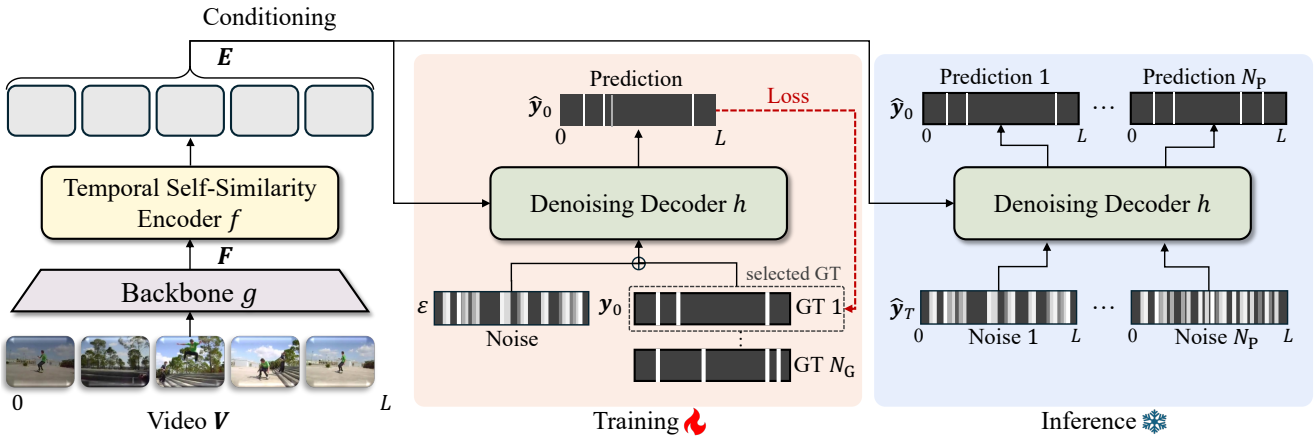

Figure 2. **Overview of DiffGEBD.** Input video $\boldsymbol{V}$ is given to the backbone network g, producing visual features $\boldsymbol{F}$ as output. Then, the extracted visual features $\boldsymbol{F}$ are produced to the encoder f, generating $\boldsymbol{E}$. During training, Gaussian noise ϵ is added to the ground-truth label $\boldsymbol{y}_0$ following the diffusion forward step. The decoder h then predicts boundaries from a noisy label $\boldsymbol{y}_t$ conditioned on $\boldsymbol{E}$. During inference, the decoder iteratively denoises starting from the random Gaussian noise $\hat{\boldsymbol{y}}_T$, generates the predictions, *i.e.*, $\hat{\boldsymbol{y}}_T \rightarrow \hat{\boldsymbol{y}}_{T-\Delta} \rightarrow \cdots \rightarrow \hat{\boldsymbol{y}}_0$, following DDIM inference step [37].

Finally, using the reparameterization trick, this can be formulated as follows:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}). \tag{4}$$

Reverse Process. The reverse process involves estimating $\mathbf{x}_0$ from $\mathbf{x}_t$, which is the opposite of the forward process. Since it is difficult to estimate the true data distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, it is defined using a model distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2 I), \tag{5}$$

where σ_t^2 is controlled by β_t. $\mu_\theta(\mathbf{x}_t, t)$ is a predicted mean parameterized by deep neural network. Instead of predicting $\mu_\theta(\mathbf{x}_t, t)$ directly, we let the model predict $\mathbf{x}_0$ by neural network $f_\theta(\mathbf{x}_t, t)$. From pure random noise $\mathbf{x}_T$, the model can reduce the noise through an update rule as follows:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\mathbf{x}_t, t) + \frac{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}}{\sqrt{1-\bar{\alpha}_t}}\mathbf{x}_t - \frac{\sqrt{\bar{\alpha}_t} f_\theta(\mathbf{x}_t, t)}{\sqrt{1-\bar{\alpha}_t}} + \sigma_t \epsilon. \tag{6}$$

By iteratively applying Eq. 6, a model can generate samples from p_θ via a trajectory from T to 0. DDIM sampling [37] strategies skip steps in the trajectory, *i.e.*, $\mathbf{x}_T \rightarrow \mathbf{x}_{T-\Delta} \rightarrow \ldots \rightarrow \mathbf{x}_0$, for better efficiency.

4. Proposed Approach

We introduce DiffGEBD, a novel diffusion-based framework for generic event boundary detection. This section provides the problem setup (Sec. 4.1), details of DiffGEBD (Sec. 4.2), the training objective (Sec. 4.3), and integration of the classifier-free guidance(Sec. 4.4).

4.1. Problem setup

Given a video $\boldsymbol{V} \in \mathbb{R}^{L\times H\times W\times 3}$ consisting of L frames, where each frame has height H, width W, and RGB channels, the goal of generic event boundary detection (GEBD) is to identify a sequence of event boundaries $\boldsymbol{y} \in \{0,1\}^L$. Each element $\boldsymbol{y}_l$ is a binary indicator that represents whether an event boundary is present, with 1 indicating presence and 0 indicating absence at frame l.

4.2. DiffGEBD

The overall architecture of DiffGEBD is illustrated in Fig. 2. The input video is fed into a backbone network g to extract visual feature representations. The encoder f captures relevant temporal changes across adjacent frames via temporal self-similarity, and the denoising decoder h refines random Gaussian noise into event boundary predictions conditioned on the visual embeddings produced by the encoder.

During training, we randomly sample a diffusion time step $t \in \{1, 2, \ldots, T\}$ and add noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ to the ground-truth boundary label $\boldsymbol{y}_0$ following Eq. 4, generating noisy boundary label $\boldsymbol{y}_t$ at time step t. The decoder takes $\boldsymbol{y}_t$ as input and is trained to reconstruct the original boundary label $\boldsymbol{y}_0$. For each video with N_G ground-truth (GT) annotations, we select one annotation per iteration to serve as the GT, ensuring that every annotation is used once per epoch.

During inference, the decoder starts with random Gaussian noise $\hat{\boldsymbol{y}}_T$ and iteratively denoises for generating final predictions, *i.e.*, $\hat{\boldsymbol{y}}_T \rightarrow \hat{\boldsymbol{y}}_{T-\Delta} \rightarrow \cdots \rightarrow \hat{\boldsymbol{y}}_0$, following DDIM inference step [37]. Here, $\hat{\boldsymbol{y}}$ denotes predicted boundaries. For diverse and plausible boundary predictions, DiffGEBD can generate N_P predictions with a single model by randomly initializing the starting Gaussian noise $\hat{\boldsymbol{y}}_T$ for each prediction.

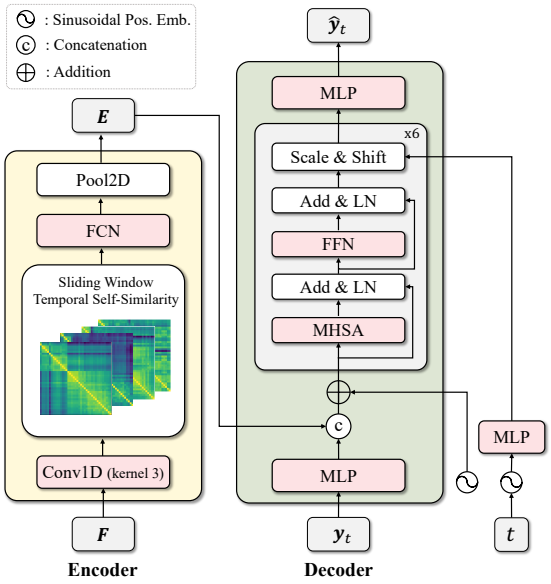

Figure 3. **Detailed architecture of encoder and decoder**

Backbone. Given an input video $\boldsymbol{V}$, We first extract video features $\boldsymbol{F} \in \mathbb{R}^{L \times D}$ through a backbone network g:

$$\boldsymbol{F} = g(\boldsymbol{V}), \tag{7}$$

where D denotes the feature dimension. We employ pre-trained ResNet-50 [14] as g.

Encoder. The encoder f is designed to capture diverse temporal variations between adjacent frames by leveraging temporal self-similarity, which helps identify subtle changes in scene dynamics that indicate event boundaries. Following [55], we adopt the temporal self-similarity encoder as f. Specifically, the encoder f comprises a 1D convolution (kernel size=3), followed by a sliding-window temporal self-similarity module, a fully convolutional network (FCN), and a 2D pooling operation, as shown in Fig. 3. The encoder f takes video features $\boldsymbol{F}$ as input and produces temporal embeddings $\boldsymbol{E} \in \mathbb{R}^{L \times C}$ as output:

$$\boldsymbol{E} = f(\boldsymbol{F}), \tag{8}$$

where C denotes the feature dimension. For further architectural details, please refer to [55].

Decoder. The decoder h is built upon the Transformer encoder layer [43], which aims to denoise input noisy boundary labels $\boldsymbol{y}_t$ at time-step t into the ground-truth boundary labels, conditioned on the temporal embeddings $\boldsymbol{E}$. As illustrated in Fig. 3, the input $\boldsymbol{y}_t$ is first processed by an `MLP` layer, then concatenated with $\boldsymbol{E}$ along the channel dimension. A sinusoidal position embedding is added to this combined feature, which is passed through self-attention layers [43]. The diffusion time step t, encoded via sinusoidal embedding and `MLP` layers, is injected into the model through a scale-and-shift operation [16, 29]. Finally, the output from the decoder h is processed by an `MLP` layer to produce the final prediction $\hat{\boldsymbol{y}}_t$:

$$\hat{\boldsymbol{y}}_t = h(\boldsymbol{y}_t, t, \boldsymbol{E}). \tag{9}$$

4.3. Training objective

The model is trained using mean squared error loss $\mathcal{L}$ between the ground-truth boundary label $\boldsymbol{y}_0$ and the prediction $\hat{\boldsymbol{y}}_t$ at time-step t:

$$\mathcal{L} = \frac{1}{L}\sum_{l=1}^{L}(\boldsymbol{y}_{0,l} - \hat{\boldsymbol{y}}_{t,l})^2. \tag{10}$$

4.4. Classifier-free guidance (CFG)

To address the inherent ambiguity in event boundary detection, we use classifier-free guidance [15]. This method balances prediction diversity and fidelity by combining conditional and unconditional diffusion models.

Training with CFG. Both conditional and unconditional diffusion models are trained for classifier-free guidance. To achieve this, We randomly drop the conditional features $\boldsymbol{E}$ with probability $p \in [0, 1]$ in Eq. 9, effectively training the model to predict with and without conditioning jointly:

$$\hat{\boldsymbol{y}}_t = \begin{cases} \hat{\boldsymbol{y}}_t^{\mathrm{c}} = h(\boldsymbol{y}_t, t, \boldsymbol{E}), & \text{with probability } 1-p, \\ \hat{\boldsymbol{y}}_t^{\mathrm{u}} = h(\boldsymbol{y}_t, t, \boldsymbol{0}_{L\times C}), & \text{with probability } p. \end{cases} \tag{11}$$

where $\boldsymbol{0}_{m\times n}$ denotes a zero matrix with size of m and n.

Inference with CFG. During inference, diversity can be adjusted by changing the value of classifier-free guidance weight w:

$$\hat{\boldsymbol{y}}_t = (1+w)\hat{\boldsymbol{y}}_t^{\mathrm{c}} - w\hat{\boldsymbol{y}}_t^{\mathrm{u}}, \tag{12}$$

where $\hat{\boldsymbol{y}}_t^{\mathrm{c}}$ and $\hat{\boldsymbol{y}}_t^{\mathrm{u}}$ denote the conditional and unconditional predictions, respectively, obtained from Eq. 11. A larger w leads to more deterministic predictions that closely follow the video content, while a smaller w allows for more diverse predictions that reflect the inherent ambiguity in boundary distribution. The overall training and inference algorithms are provided in the supplementary material.

5. Experiments

5.1. Setup

In our experiment, we evaluate our method on two standard GEBD benchmarks: Kinetics-GEBD [35] and TAPOS [34]. Each video is uniformly sampled to 100 frames. We

use ResNet-50 [14] pretrained on ImageNet-1K [7] as the backbone network g. We employ the BasicGEBD-L4 encoder [55] and a 6-layer Transformer [43] for our encoder f and decoder h, . We adopt FiLM [29] for the diffusion timestep embedding. During training, we set probability p of classifier-free guidance as 0.1. For Kinetics-GEBD [35], which provides five annotations per video, we use a maximum of four annotations, selected based on F1 consistency score [23, 35, 55]. Please refer to our supplementary materials for more details of the datasets and our implementation.

5.2. Evaluation Metrics

5.2.1. F1 score

In the conventional evaluation of GEBD, a single prediction is evaluated for each video [23, 35, 55, 56]. The F1 score based on relative distance (Rel.Dis. [35]) is the basic evaluation metric. When multiple annotations are available, the F1 score is computed by taking the maximum F1 score among all possible prediction-annotation pairs.

However, the F1 score does not account for scenarios where multiple solutions are generated, nor does it capture the inherent diversity among ground-truth annotations. In the following, we introduce new evaluation metrics, *i.e.*, *symmetric F1* and *diversity* scores, that consider both multiple predictions and the diversity of GT annotations.

5.2.2. Symmetric F1 score

When multiple predictions are generated for a video, evaluating the many-to-many alignment between predictions and GT annotations requires considering two key aspects: (1) how accurately each prediction matches one of the GT annotations (Pred-to-GT alignment) and (2) how well each GT annotation is covered by the predictions (GT-to-Pred alignment). To address these aspects, we propose the symmetric F1 score (F1_{sym}), which combines two directional F1 scores: the Pred-to-GT alignment score (F1_{p2g}) and the GT-to-Pred alignment scores (F1_{g2p}). This bi-directional metric ensures a comprehensive evaluation by jointly measuring how well predictions capture the ground truth and vice versa, reflecting both prediction accuracy and diversity.

To formally define our metrics, we first establish our notation. For each video with L frames, we denote N_{g} ground truth annotations and N_{p} model predictions by $\boldsymbol{Y} \in \mathbb{R}^{N_{\text{g}} \times L}$ and $\hat{\boldsymbol{Y}} \in \mathbb{R}^{N_{\text{p}} \times L}$, respectively.

Pred-to-GT alignment score F1_{p2g}. The Pred-to-GT alignment score, F1_{p2g}, measures how well each predicted boundary aligns with at least one ground truth annotation, similar to the conventional GEBD evaluation. It is computed by taking each prediction $\hat{Y}_i$, finding its highest F1 score across all ground truth annotations Y_j, and averaging these maximum scores across all predictions as:

$$\text{F1}_{\text{p2g}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{1}{N_{\text{p}}} \sum_{i=1}^{N_{\text{p}}} \max_{j \in \{1,\dots,N_{\text{g}}\}} \text{F1}(\hat{Y}_i, Y_j), \quad (13)$$

where $\text{F1}(X, Y)$ computes the F1 score between X and Y.

GT-to-Pred alignment score F1_{g2p}. To account for the variability and diversity in GT annotations, the GT-to-Pred score, F1_{g2p}, evaluates how well each annotation is covered by any of the predictions. This is achieved by reversing the formulation to assess each ground truth annotation against all predictions, as follows:

$$\text{F1}_{\text{g2p}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{1}{N_{\text{g}}} \sum_{j=1}^{N_{\text{g}}} \max_{i \in \{1,\dots,N_{\text{p}}\}} \text{F1}(\hat{Y}_i, Y_j). \quad (14)$$

Symmetric F1 score F1_{sym}. The symmetric F1 score finally combines the two directional F1 scores, *i.e.*, F1_{p2g} and F1_{g2p}, by taking a harmonic mean as:

$$\text{F1}_{\text{sym}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{2 \times \text{F1}_{\text{p2g}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) \times \text{F1}_{\text{g2p}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y})}{\text{F1}_{\text{p2g}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}) + \text{F1}_{\text{g2p}}(\hat{\boldsymbol{Y}}, \boldsymbol{Y})}. \quad (15)$$

The final symmetric F1 score for the entire dataset is obtained by computing the score for each video individually and then taking an average across all videos in the dataset.

5.2.3. Diversity score

Although the proposed symmetric F1 score measures a comprehensive alignment between multiple predictions and ground truth annotations, it does not directly measure the diversity among predictions. We thus introduce the diversity score that quantifies the average pairwise dissimilarity among predictions, following [54]. The diversity score among N_p predictions $\hat{\boldsymbol{Y}}$ is defined as:

$$\text{Diversity}(\hat{\boldsymbol{Y}}) = \frac{2}{N_{\text{p}}^2} \sum_{i=1}^{N_{\text{p}}} \sum_{j=1}^{N_{\text{p}}} (1 - \text{F1}(\hat{Y}_i, \hat{Y}_j)), \quad (16)$$

which computes the average dissimilarity among all predictions. Here, the F1 score serves as the similarity measure, ensuring that the diversity score reflects how different the generated predictions are from each other. Note that higher values indicate higher diversity. Similar to the symmetric F1 score, the diversity score is averaged across all videos in the dataset.

5.3. Effect of the CFG Weight w

The CFG weight w is a key factor in balancing the conditional and unconditional diffusion models. A larger w increases the influence of the conditional model, strengthening the impact of the temporal self-similarity feature in

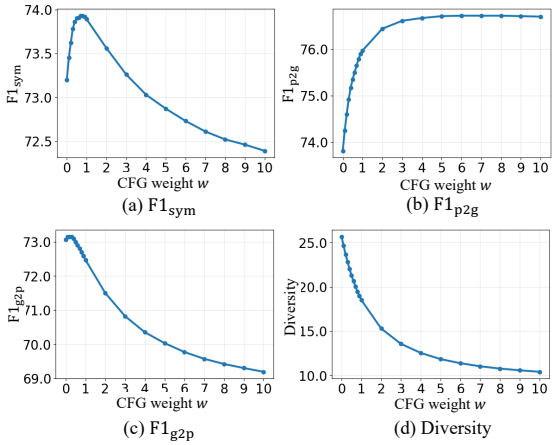

Figure 4. **Effect of CFG weight w.** The x-axis represents the CFG weight w, while the y-axis shows (a) F1_{sym}, (b) F1_{p2g}, (c)F1_{g2p}, and (d) diversity, respectively.

the diffusion process. In contrast, a smaller w increases the influence of the unconditional model, enabling the generation of more diverse predictions by relying less on the conditioning signal. To evaluate the effect of w, we conduct experiments by varying its value in Eq. 12 from 0.0 to 10.0 during inference, following [15]. Figure 4 presents the results, where the x-axis represents w, while the y-axis shows (a) F1_{sym}, (b) F1_{p2g}, (c) F1_{g2p}, and (d) diversity score.

The Pred-to-GT alignment score F1_{p2g} increases (Fig. 4b) as w increases, indicating that the model places greater emphasis on visual features when generating predictions. In contrast, both the GT-to-Pred alignment score F1_{g2p} (Fig. 4c) and the diversity score (Fig. 4d) decreases with higher w, as stronger conditioning reduces variability in predictions. Conversely, a smaller w increases diversity by enhancing the influence of the unconditional model while relatively reducing dependence on the conditional model. Interestingly, F1_{p2g} exhibits an opposite pattern to both F1_{g2p} and diversity score, while F1_{g2p} and diversity score follow a similar tendency. This also suggests that a higher F1_{p2g} score does not always guarantee diverse predictions, and excessive diversity may negatively impact Pred-to-GT alignment. Furthermore, diversity appears to be closely related to GT-to-Pred alignment, as a higher F1_{g2p} indicates better coverage of ground truth annotations, which inherently requires greater diversity in predictions.

The symmetric F1 score F1_{sym}, defined as the harmonic mean of F1_{p2g} and F1_{g2p}, exhibits a non-monotonic relationship with the guidance weight, reaching its peak at $w = 0.7$. This result highlights the trade-off between Pred-to-GT alignment and GT-to-Pred alignment. A moderate guidance weight effectively balances these trade-offs, maximizing the symmetric F1 score by preserving alignment with the ground truth while ensuring sufficient diversity in predictions. The complete numerical results are provided in the supplementary material.

Method	F1_{sym}	F1_{p2g}	F1_{g2p}	Diversity
Temporal Perceiver† [39]	69.4	72.2	67.4	14.6
SC-Transformer† [23]	<u>72.9</u>	74.9	<u>71.6</u>	<u>18.9</u>
BasicGEBD† [55]	72.2	74.5	70.6	18.6
EfficientGEBD† [55]	72.6	**76.0**	70.2	14.9
DiffGEBD(Ours)	**73.9**	<u>75.7</u>	**72.8**	**20.0**

Table 1. **Diversity-aware evaluation of GEBD on Kinetics-GEBD.** † reproduced by out setup. **Boldface** and <u>underline</u> indicate the best and the second-best scores, respectively.

5.4. Diversity-aware Evaluation of GEBD

In Table 1, we compare DiffGEBD with previous methods [23, 39, 55] on the Kinetics-GEBD dataset using the diversity-aware evaluation protocol in Sec. 5.2.2. For multiple prediction generations, we set the number of predictions N_{p} to 5, as the average number of annotations per video in the dataset is 4.93 [35]. Since all previous methods produce deterministic outputs, we reproduce and evaluate each model by training it five times with random initialization to obtain multiple predictions. Please note that our experiments are conducted on models with publicly available code[1]. The reproduced models are marked with † in Table 1, and their performance is presented in the supplementary material. In contrast to these deterministic models, DiffGEBD generates diverse predictions from a single trained model by varying the initial Gaussian noise $\hat{\boldsymbol{y}}$, eliminating the need for multiple training runs to achieve diversity. In this experiment, we set the CFG weight w to 0.7 and the relative distance threshold for the F1 score to 0.05.

Table 1 presents the overall results, where DiffGEBD achieves the state-of-the-art performance on F1_{sym}, F1_{g2p}, and the diversity score, while showing comparable results on F1_{p2g} compared to the previous methods. This result indicates that DiffGEBD not only generates diverse predictions but also maintains solid alignment with ground truth annotations, achieving an optimal balance between diversity and plausibility. EfficientGEBD [55] achieves the highest score in F1_{p2g}; however, its lower F1_{g2p} leads to lower score on F1_{sym}, and its diversity score is also low. This suggests that its predictions cover fewer ground truth annotations with lower diversity, prioritizing precision over diversity. By comparing the results of EfficentGEBD to BasicGEBD [55], we observe that a significant increase in diversity does not necessarily lead to a proportional improvement in F1_{g2p}. This finding implies that higher diversity alone does not guarantee better GT-to-Pred alignment, emphasizing the importance of plausibility in predictions. Full results with varying relative distance values are presented in

[1]We utilize the official Github repositories for Temporal Perceiver [39]: https://github.com/MCG-NJU/TemporalPerceiver, SC-Transformer [23]: https://github.com/lufficc/SC-Transformer, and BasicGEBD/EfficientGEBD [55]: https://github.com/Ziwei-Zheng/EfficientGEBD.

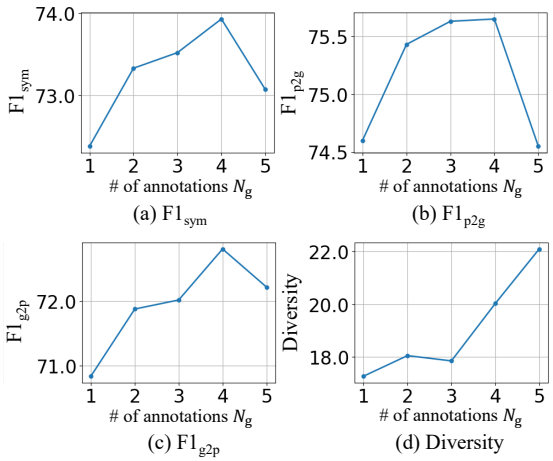

Figure 5. **Effect of the number of annotations.** Model performance with varying numbers of annotations (1-5).

Conditioning	$F1_{sym}$	$F1_{p2g}$	$F1_{g2p}$	Diversity
$\boldsymbol{F}$	66.9	69.2	65.3	15.8
$\boldsymbol{E}$	**73.9**	**75.7**	**72.3**	**20.0**

Table 2. **Effect of conditioning in diffusion.** Using temporal self-similarity feature $\boldsymbol{E}$ as a diffusion condition is effective.

the supplementary material.

5.5. Analysis

Effect of the number of annotations N_g. Since each annotation represents an individual's subjective interpretation of event boundaries, we experiment by adjusting the number of annotations used during training. Rather than selecting annotations randomly, we rank them based on reliability using the F1 consistency scores introduced in [35]. We incrementally increase N_g from 1 to 5, prioritizing annotations from annotators with the highest consistency scores.

Figure 5 presents the results, where the x-axis represents N_g, while the y-axis shows (a) $F1_{sym}$, (b) $F1_{p2g}$, (c) $F1_{g2p}$, and (d) diversity score. We observe a consistent improvement in overall performance as N_g increases from 1 to 4, indicating that incorporating multiple reliable annotators helps the model better capture variations in boundary annotations while improving fidelity. However, when all five annotators are included, we observe a decline in $F1_{sym}$, $F1_{p2g}$, and $F1_{g2p}$, while the diversity score continues to increase. This suggests that although using more annotations enhances diversity, incorporating low-consistency annotations can negatively impact performance.

Effect of conditioning in diffusion. To examine the impact of the conditioning feature in denoising diffusion, we conduct experiments by varying the conditioning feature in the diffusion process. Specifically, we replace the temporal self-similarity feature $\boldsymbol{E}$ with visual features $\boldsymbol{F}$ extracted directly from the backbone network g. Table 2 presents the results. We observe a significant performance drop when using $\boldsymbol{F}$, demonstrating the importance of temporal self-similarity features as a conditioning input for the diffusion model. Since self-similarity captures subtle changes across frames, using $\boldsymbol{E}$ is more effective.

Steps	$F1_{sym}$	$F1_{p2g}$	$F1_{g2p}$	Diversity
1	54.0	60.5	50.2	20.7
2	72.6	75.9	71.1	19.2
4	73.4	75.7	71.8	18.0
8	73.8	**75.8**	72.4	18.8
16	**73.9**	**75.8**	72.6	19.5
32	**73.9**	75.7	**72.8**	20.0
50	**73.9**	75.6	**72.8**	**20.4**

Table 3. **Effect of inference step.** Following the DDIM sampling strategy, the model can skip the timestep T.

Method	F1@0.05	
	Kinetics-GEBD	TAPOS
BMN [25]	18.6	-
BMN-StartEnd [25]	49.1	-
ISBA [10]	-	10.6
TCN [22]	58.8	23.7
CTM [17]	-	24.4
TransParser [34]	-	23.9
PC [35]	62.5	52.2
SBoCo [20]	73.2	-
Temporal Perceiver [39]	74.8	55.2
DDM-Net [40]	76.4	60.4
CVRL [24]	74.3	-
LCVS [53]	76.8	-
SC-Transformer [23]	77.7	61.8
BasicGEBD [55]	76.8	60.0
EfficientGEBD [55]	78.3	<u>63.1</u>
DyBDet [56]	**79.6**	62.5
DiffGEBD (ours)	<u>78.7</u>	**66.0**

Table 4. **Conventional evaluation of GEBD**. We report F1 scores based on the conventional GEBD evaluation protocol. **Boldface** and <u>underline</u> indicate the best and the second-best, respectively.

Effect of DDIM inference steps. We investigate the impact of diffusion steps by varying T from 1 to 50. As shown in Table 3, while diversity remains consistently high regardless of the number of steps, other metrics show improvement with increasing steps with diminishing marginal benefits. Considering the linear increase in computational cost with the number of steps, we choose $T = 32$ as a balanced choice between performance and efficiency.

5.6. Conventional Evaluation of GEBD

Table 4 compares the performance of the proposed method on two standard GEBD benchmark datasets, Kinetics-GEBD and TAPOS, following the conventional evaluation protocol [35] as described in Sec. 5.2.1. Note that all methods use ResNet-50 [13] trained on ImageNet [7] as the

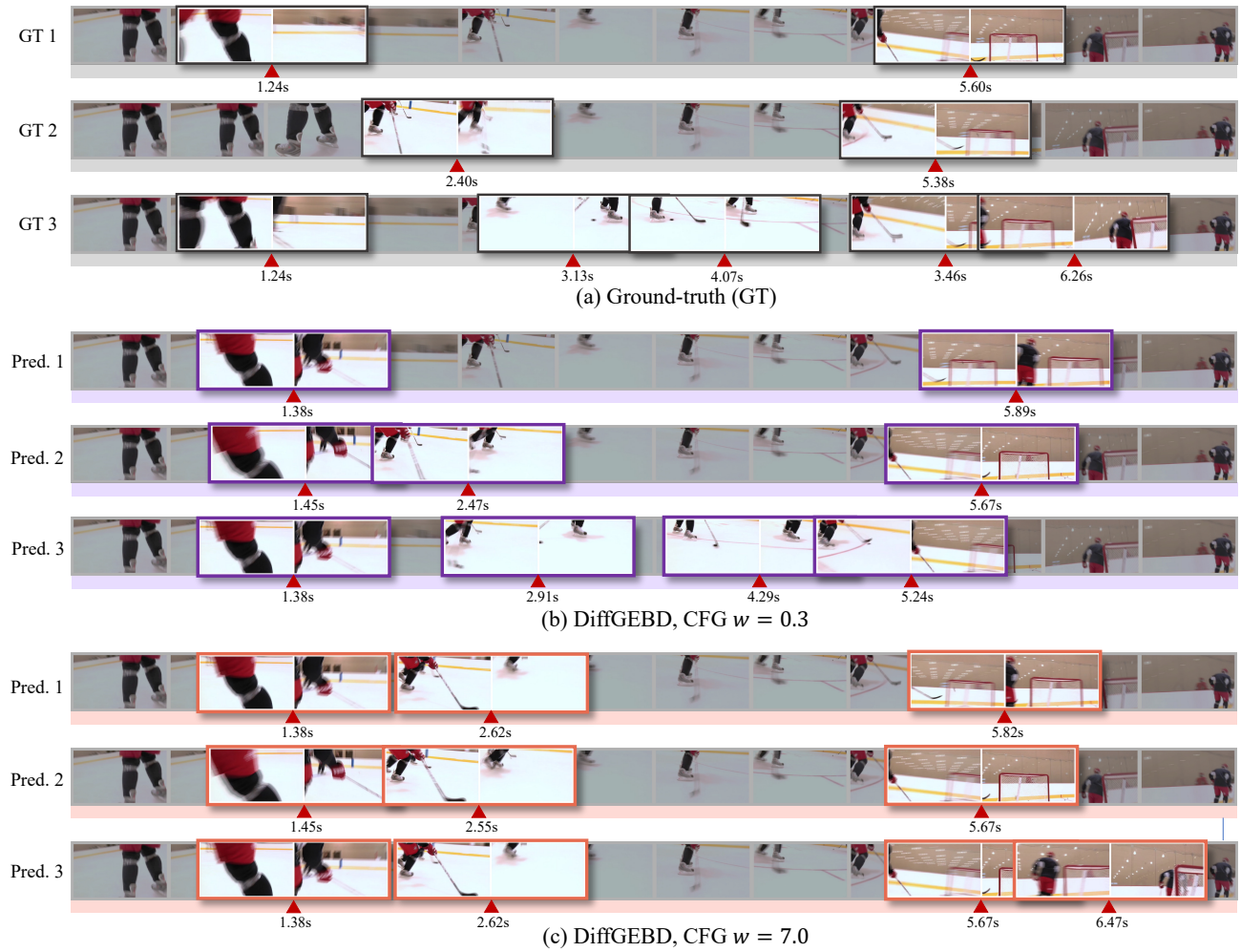

Figure 6. **Example results on Kinetics-GEBD.** The figure illustrates (a) Ground truth annotations, (b) predictions with $w = 0.3$, and (c) predictions with $w = 7.0$.

backbone network for a fair comparison. In this experiment, we set the CFG weight w to 7.0 to enhance the influence of temporal self-similarity features in the video when generating predictions. DiffGEBD achieves comparable results on Kinetics-GEBD and outperforms on TAPOS. This result demonstrates that DiffGEBD can effectively generate highly feasible predictions with a high guidance weight, ensuring stronger adherence to the conditioning features.

5.7. Example Results

Figure 6 illustrates example results of DiffGEBD on the Kinetics-GEBD dataset, showing (a) ground-truth annotations, (b) predictions with $w = 0.3$, and (c) predictions with $w = 7.0$. All outputs were generated using the same model with different initial noise. We observe that clear boundaries (e.g., subject's movements between 1.24s to 1.45s) are consistently detected across the predictions, regardless of the guidance weight. However, boundaries that exhibit human ambiguity, such as subtle action changes (e.g., hockey stick movements at 2.91s and 4.29s in Pred. 3 of (b)), vary across different generations. Notably, we observe that lower weight guidance allow for diverse predictions, while higher guidance weights lead to more consistent predictions.

6. Conclusion

We have presented DiffGEBD, a novel diffusion-based boundary detection model from a generative perspective. The proposed method encodes temporal changes between adjacent frames using self-similarity, then iteratively refines random noise into plausible boundaries via denoising diffusion. By integrating classifier-free guidance, it enables explicit control over the degree of diversity. Furthermore, we present a diversity-aware evaluation protocol, introducing the symmetric F1 and diversity scores, which jointly capture many-to-many alignments and the variability in model predictions. We believe that our model offers a novel perspective on producing diverse yet plausible generic event boundaries, paving the way for a richer and more nuanced understanding of event boundaries.

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[2] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 2

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1

[4] 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

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[6] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *CVPR*, 2009. 5, 7

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[9] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[10] 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. 7

[11] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 1

[12] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, and Jun Liu. Action detection via an image diffusion process. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18351–18361, 2024. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1, 2, 4, 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[17] 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. 7

[18] Christopher Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018, 1997. 2

[19] Kumara Kahatapitiya and Michael S Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8385–8394, 2021. 1

[20] 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. 1, 2, 7

[21] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018. 2

[22] 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. 1, 7

[23] 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, 5, 6, 7

[24] 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. 7

[25] 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. 1, 7

[26] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10139–10149, 2023. 1, 2

[27] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001. 2

[28] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion

models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[29] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 4, 5

[30] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11536–11546, 2023. 2

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

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

[34] 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. 4, 7

[35] 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, 4, 5, 6, 7

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 3

[38] 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

[39] 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. 1, 6, 7

[40] 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. 1, 2, 7

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[42] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4, 5

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[45] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 1

[46] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. 2

[47] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165, 2020. 1

[48] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021. 1

[49] J. M. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127:3–21, 2001. 1

[50] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Gated temporal diffusion for stochastic long-term dense anticipation. In *European Conference on Computer Vision*, pages 454–472. Springer, 2024. 2

[51] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7094–7103, 2019. 1

[52] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1

[53] 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. 2, 7

[54] Wei Zhang, Xiaohong Zhang, Sheng Huang, Yuting Lu, and Kun Wang. A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. page 4751–4759, New York, NY, USA, 2022. Association for Computing Machinery. 5

[55] 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. 1, 2, 4, 5, 6, 7

[56] 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. 2, 5, 7

[57] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525, 2021. 1