

Learning Action Hierarchies via Hybrid Geometric Diffusion

Arjun Ramesh Kaushik Nalini K. Ratha Venu Govindaraju
 University at Buffalo, SUNY

{kaushik3, nratha, govind}@buffalo.edu

Abstract

Temporal action segmentation is a critical task in video understanding, where the goal is to assign action labels to each frame in a video. While recent advances leverage iterative refinement-based strategies, they fail to explicitly utilize the hierarchical nature of human actions. In this work, we propose HybridTAS - a novel framework that incorporates a hybrid of Euclidean and hyperbolic geometries into the denoising process of diffusion models to exploit the hierarchical structure of actions. Hyperbolic geometry naturally provides tree-like relationships between embeddings, enabling us to guide the action label denoising process in a coarse-to-fine manner: higher diffusion timesteps are influenced by abstract, high-level action categories (root nodes), while lower timesteps are refined using fine-grained action classes (leaf nodes). Extensive experiments on three benchmark datasets, GTEA, 50Salads, and Breakfast, demonstrate that our method achieves state-of-the-art performance, validating the effectiveness of hyperbolic-guided denoising for the temporal action segmentation task.

1. Introduction

Temporal Action Segmentation (TAS) aims to assign an action label to every frame in an untrimmed video, enabling fine-grained understanding of complex human activities. This task is crucial for applications such as human-computer interaction, video surveillance, and robotic perception. Despite recent progress, achieving accurate segmentation remains challenging due to variability in action durations, temporal dependencies, and ambiguous transitions between actions.

Most existing approaches follow a refinement-based strategy, where a sequence of models iteratively improves frame-level predictions by leveraging contextual and temporal cues [1, 15, 28, 33, 36, 64, 72]. Recently, diffusion models have emerged as a promising direction in this space, demonstrating strong performance due to their ability to model complex temporal dynamics in a generative manner

[29, 48]. However, current diffusion-based methods treat action labels as flat categories, ignoring the rich hierarchical structure often present in human activities. For example, high-level actions like “prepare meal” can be decomposed into sub-actions such as “cut vegetables”, “boil water”, and “stir ingredients”. To address this limitation, we propose a novel approach that incorporates a hybrid of Euclidean and hyperbolic geometry into the denoising process of diffusion models to model hierarchical relationships between actions.

Hyperbolic geometry is inherently suited for representing tree-like structures due to its exponential growth property, making it a natural choice for representing hierarchy in datasets. Additionally, hyperbolic distances are also a natural measure of uncertainty and class boundaries [4]. Our core insight is to guide the denoising trajectory in a coarse-to-fine manner, aligning the generative process with the action hierarchy. At higher diffusion timesteps (early in the generation process), the model is guided by coarse, high-level action categories (root nodes), and as the noise is reduced (towards lower timesteps), the model is increasingly influenced by fine-grained, low-level actions (leaf nodes).

We instantiate this idea through Euclidean and hyperbolic losses in the DiffAct framework [48]. These hybrid losses are applied in two phases: *Stabilization Phase* and *Guidance Phase*. In the *Stabilization Phase*, the model learns global representational embeddings of actions (referred to as action prototypes) that are optimized in hyperbolic space. Next, the *Guidance Phase* controls and enforces radial outward movement of denoised embeddings towards their action prototypes. This structured guidance enables the model to align its diffusion trajectory along the geodesic path connecting the action prototype and the origin. We evaluate our approach on three widely used benchmarks for TAS: GTEA [23], 50Salads [65], and Breakfast [40]. Our method consistently outperforms the SOTA baselines, demonstrating the benefit of our hybrid geometry optimization.

To summarize, our contributions are as follows: (1) We propose a novel hierarchical diffusion model for temporal action segmentation, which integrates a hybrid of Euclidean and hyperbolic geometry to capture action hierarchies. (2)

We introduce a two-step optimization strategy using hyperbolic loss functions. The first step refines the action labels from coarse to fine levels of abstraction. (3) The second step enforces the model to align its denoising trajectory with the semantic hierarchy of actions. Additionally, our model surpasses SOTA works with fewer inference steps.

2. Related Work

Temporal Action Segmentation. TAS aims to assign frame-wise action labels to video sequences [1, 6, 8, 15, 28, 36, 47, 48, 72, 74]. Early methods addressed this task using temporal sliding windows [37, 54] or grammar-based approaches [40, 41] to incorporate hierarchical structure in actions. With the rise of deep learning, temporal convolutional networks [44, 47] and transformer-based models [8, 74] have been introduced to model temporal dependencies. However, capturing long-range temporal relations in videos remains challenging. To address this, several works [1, 15, 28, 33, 36, 64, 72] have proposed iterative refinement strategies that operate on top of TAS predictions to improve performance. More recently, DiffAct [48] and ActFusion [29] leverage a diffusion-based framework to iteratively denoise action label predictions conditioned on video features. ActFusion [29] adopts the diffusion process with a novel anticipative masking strategy, aiming to jointly unify TAS and Long-Term Action Anticipation.

Diffusion Models. Diffusion-based generative models [17, 32, 59, 60], which have been theoretically unified with score-based approaches [61–63], are widely recognized for their stable training dynamics and the absence of adversarial mechanisms typically required in generative learning. These models have demonstrated remarkable success across various domains, including image generation [19, 42, 55, 70], natural language generation [75], text-to-image synthesis [30, 39], and audio generation [43, 45]. Recent advancements have proposed gradient-based guidance techniques to further improve sampling efficiency [20]. While diffusion models have been repurposed for several image understanding tasks, such as object detection [16] and semantic segmentation [3, 7], their application to video-related problems remains relatively limited. Notable exceptions include works on video forecasting and infilling [34, 69, 73], as well as recent efforts in video memorability prediction [67] and frequency-aware video captioning [76].

Hyperbolic Geometry. Hyperbolic geometry has become a powerful tool for embedding hierarchical and tree-like structures with minimal distortion [10, 50, 56, 57]. Since the introduction of Hyperbolic Neural Networks (HNN) [27], hyperbolic space has been successfully integrated into diverse neural architectures, including convolutional net-

works [58], attention-based models [31], graph neural networks [11, 49], and more recently, vision transformers [22]. Recent studies have also used the hyperbolic radius to capture uncertainty [12, 22, 24, 26] and to model hierarchical relationships such as parent-child structures [4, 22, 50, 66, 68]. In image segmentation, hyperbolic geometry has gained traction due to its strengths in uncertainty modeling and hierarchical representations [4, 12, 25].

Despite significant progress in hyperbolic deep learning, its application to Temporal Action Segmentation (TAS) remains unexplored. Notably, existing approaches overlook the hierarchical structure of actions in the latent space. To the best of our knowledge, this is the first work to leverage hyperbolic geometry to guide the denoising trajectory of diffusion models for addressing the TAS task.

3. Background

In this section, we provide a summary of DiffAct [48] to contextualize our contributions and introduce Hyperbolic Geometry [10, 50]. For a better understanding of Diffusion Models [17, 32, 59, 60], we refer the reader to Sec. A.2 (Appendix).

3.1. Diffusion Action Segmentation (DiffAct)

Diffusion models approximate a data distribution by corrupting samples with Gaussian noise and learning to reverse this process through iterative denoising. At each timestep t , clean data \mathbf{x}_0 is transformed into a noisy version \mathbf{x}_t via a variance schedule $\gamma(t)$, while a neural network $f(\mathbf{x}_t, t)$ is trained to predict the original signal using an L2 loss. During inference, the model begins from pure noise and progressively refines it into a coherent sample. In our setting, this corresponds to generating frame-wise action label sequences from Gaussian noise, conditioned on video features for temporal action segmentation.

$$\mathbf{x}_t = \sqrt{\gamma(t)} \mathbf{x}_0 + \sqrt{1 - \gamma(t)} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \quad (1)$$

DiffAct [48] introduces a generative formulation for temporal action segmentation by leveraging denoising diffusion probabilistic models (DDPMs). Instead of directly predicting labels, it treats segmentation as iterative denoising, where action sequences are generated from pure noise conditioned on video features. The encoder–decoder framework integrates masking strategies inspired by human priors (position, boundary, and relation), which guide the model to better localize and infer actions. During inference, the decoder refines noisy label sequences step-by-step, optionally skipping intermediate steps for efficiency, until producing the final action segmentation. For more details, refer to Sec. A.1.

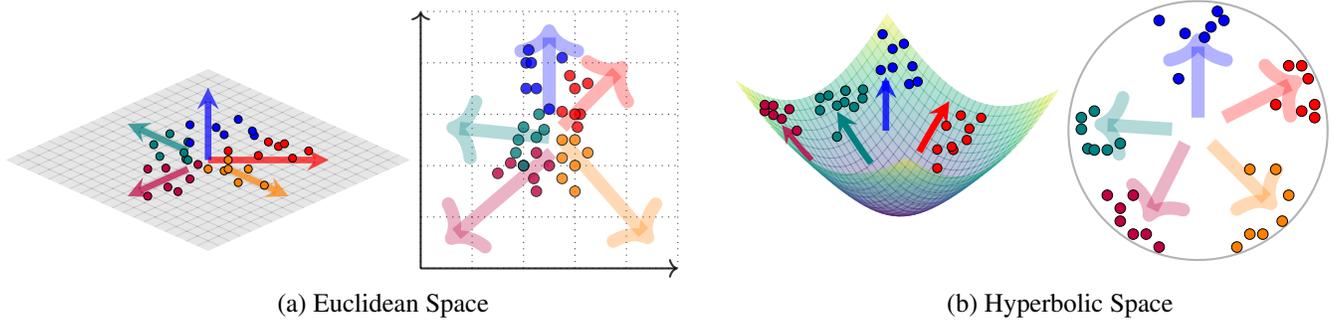


Figure 1. **Embeddings in Euclidean and Hyperbolic Space.** (a) In the Euclidean space, embeddings tend to cluster towards the origin, making it difficult to distinguish classes. (b) On the other hand, the hyperbolic space allows embeddings to spread due to its exponential distance growth. Additionally, hyperbolic geometry naturally provides information on hierarchy in data, class boundaries, and uncertainty in predictions [4].

3.2. Hyperbolic Geometry

Hyperbolic Space. Hyperbolic space is a Riemannian manifold characterized by constant negative curvature [50]. It admits several isometric models, among which the n -dimensional Hyperboloid model, defined on the hypersurface \mathbb{H}_c^n , is the most fundamental. The Poincaré Ball model \mathbb{B}_c^n can be obtained by projecting this hyperboloid onto a space-like hyperplane.

Poincaré Ball Model. An n -dimensional Poincaré Ball model with constant sectional curvature $-c$ is defined as the Riemannian manifold (\mathbb{B}_c^n, g_c) , where

$$\mathbb{B}_c^n = \{x \in \mathbb{R}^n \mid c\|x\|^2 < 1\},$$

and the Riemannian metric is given by

$$g_c(x) = \lambda_c^2(x) I_n,$$

where $\lambda_c(x) = \frac{2}{1-c\|x\|^2}$ is the conformal factor and I_n is the Euclidean metric tensor, where $\|x\|$ denotes the Euclidean norm of x .

Exponential Map. Let $x \in \mathbb{B}_c^n$ and $v \in T_x\mathbb{B}_c^n$, the exponential map $\exp_x^c : T_x\mathbb{B}_c^n \rightarrow \mathbb{B}_c^n$ is given by, where \oplus_c denotes the Möbius addition operator:

$$\exp_x^c(v) = x \oplus_c \frac{1}{\sqrt{c}} \tanh\left(\sqrt{c}\lambda_x^c \frac{\|v\|}{2}\right) [v] \quad (2)$$

Distance Function. For $x, y \in \mathbb{B}_c^n$, the hyperbolic distance is defined as:

$$d_{\mathbb{B}}(x, y) = \frac{2}{\sqrt{c}} \tanh^{-1}(\sqrt{c}\| -x \oplus_c y \|) \quad (3)$$

The distance from any point to the origin in \mathbb{B}_c^n reflects its uncertainty [4]. Specifically, the closer an embedding is to the center of the ball (origin), the higher the uncertainty.

Exterior Angle. We define the exterior angle between two points x and y in the Poincaré ball as the minimum angle between the axis of the tangent cone at x and the vector pointing toward y [18]. Formally, the angle $\theta(x, y)$ is as follows, where $\langle x, y \rangle$ denotes the standard Euclidean inner product:

$$\theta(x, y) = \cos^{-1}\left(\frac{\langle x, y \rangle(1 + \|x\|^2) - \|x\|^2(1 + \|y\|^2)}{\|x\| \cdot \|x - y\| \sqrt{1 + \|x\|^2} (\|y\|^2 - 2\langle x, y \rangle)}\right) \quad (4)$$

Aperture. The aperture of a cone centered at point $x \in \mathbb{B}^d$ is given by:

$$\alpha(x) = \arcsin\left(\frac{K(1 - \|x\|^2)}{\|x\|}\right) \quad (5)$$

where K is a scalar hyperparameter (usually set to 0.1), and $\|x\|$ is the Euclidean norm of x . This aperture decreases as $\|x\| \rightarrow 1$, i.e., as x approaches the boundary of the Poincaré ball [18].

4. Problem Formulation

Temporal Action Segmentation (TAS) involves assigning a sequence of action labels to each frame in a video, effectively classifying every input frame into one of several predefined action classes. Formally, let a video be represented as a sequence of frames $\mathbf{F} = [F_1, F_2, \dots, F_L]$ of length L . TAS aims to predict a sequence of frame-wise action labels $\mathbf{A} = [A_1, A_2, \dots, A_L]$, where each A_i is a one-hot vector corresponding to one of the C action classes.

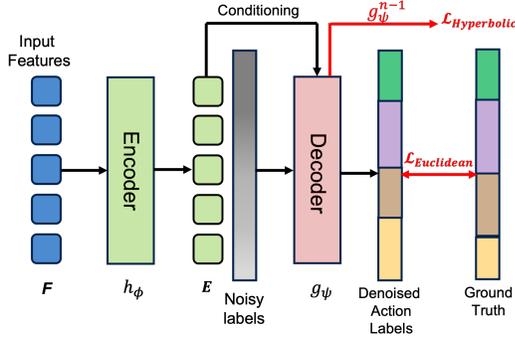


Figure 2. **Model Architecture.** Our architecture builds upon DiffAct [48], but introduces a hybrid design that operates jointly in Euclidean and hyperbolic spaces to utilize hierarchical action relationships. The Euclidean losses are applied in the label space, acting on the predicted action probabilities, while the hyperbolic losses are computed in the embedding space, directly supervising the outputs from the decoder’s final layer.

5. Method

We present **HybridTAS**, a diffusion-based framework for temporal action segmentation that operates in both Euclidean and hyperbolic spaces. Our model builds upon DiffAct’s [48] diffusion architecture (see Sec. A.1), but with a critical shift: while DiffAct optimizes solely in label space, we optimize in both hyperbolic latent space and Euclidean label space (Fig. 2). Our key innovation lies in structuring this trajectory to reflect the semantic hierarchy of action classes.

To this end, we define a unified training objective composed of four hyperbolic losses operating in the latent space: (1) **Temporal Entailment Loss** to enforce temporal consistency across frames, (2) **Prototype Margin Loss** for inter-class separation, (3) **Hyperbolic Push-Pull Loss** to reduce prediction uncertainty, and (4) **Geodesic Guidance Loss** to align denoising with hierarchical paths. It is important to note that these losses operate on the decoder’s final embeddings before outputting the action labels. On the other hand, we optimize in the label space using the standard **Cross-entropy Loss**.

Let \mathbb{B}_c^d denote the d -dimensional Poincaré ball model with curvature parameter $c > 0$. For all hyperbolic losses, the embeddings are first projected into \mathbb{B}_c^d using the manifold exponential map (Eq. 2) prior to computing distances via $d_{\mathbb{B}}$ (Eq. 3).

5.1. Training Overview

Training proceeds in two phases: (1) *Stabilization Phase*, where action prototypes are learned dynamically, and (2) *Guidance Phase*, where prototypes are fixed and used to direct the denoising path. However, two losses optimize throughout both phases: **Cross-entropy Loss** and **Tempo-**

ral Entailment Loss. The losses, in this section, are defined as the optimization objective of the diffusion model at timestep t .

Cross-entropy Loss. This is the standard cross-entropy for classification, minimizing the negative log-likelihood of the ground truth action labels for each frame. It operates directly in the Euclidean label space and is defined as:

$$\mathcal{L}_{ce} = \frac{1}{LC} \sum_{i=1}^L \sum_{c=1}^C -Y_{i,c} \log P_{i,c} \quad (6)$$

where i is the frame index and c is the label index.

Temporal Entailment Loss. To ensure temporal consistency between adjacent denoised label embeddings, we impose a structural constraint based on angular entailment. Specifically, we interpret each frame embedding as semantically dependent on its predecessor. In hyperbolic space, this relation is captured by requiring the exterior angle (Eq. 4) between consecutive embeddings to lie within the aperture (Eq. 5) of the predecessor:

$$\mathcal{L}_{entail} = \frac{1}{L-1} \sum_{l=1}^{L-1} \max(0, \theta(x_l, x_{l+1}) - \alpha(x_l)) \quad (7)$$

where $\theta(x_l, x_{l+1})$ is the exterior angle between x_l and x_{l+1} . \mathcal{L}_{entail} is not phase dependent and enforces temporal smoothness throughout training.

Training consists of two phases: *stabilization phase* and *guidance phase*. The *stabilization phase* is the early phase when the model learns the action prototype embeddings. Once the prototypes are stable, we begin the *guidance phase* to refine predictions and provide the diffusion model with a hierarchical trajectory using the prototypes as targets. Recall that these losses operate on the final embeddings of the decoder.

5.2. Step 1: Stabilization Phase

We begin by initializing C learnable embeddings for each action (action prototypes). These prototypes will serve as dynamic anchors in the *stabilization phase* when the model will learn global action representations.

Prototype Margin Loss. To avoid prototype collapse and promote discriminative representations, we introduce a margin-based repulsion loss:

$$\mathcal{L}_{margin} = \frac{1}{C(C-1)} \sum_{i < j} \max(0, m - d_{\mathbb{B}}(z_i, z_j)) \quad (8)$$

where $z_i, z_j \in \mathbb{B}^d$ are action prototypes, m is a predefined margin, and $d_{\mathbb{B}}$ is the Poincaré distance. This loss

Method	50 Salads [65]						Breakfast [40]						GTEA [23]					
	F1@10	F1@25	F1@50	Edit	Acc	Avg	F1@10	F1@25	F1@50	Edit	Acc	Avg	F1@10	F1@25	F1@50	Edit	Acc	Avg
MS-TCN++ [47]	80.7	78.5	70.1	74.3	83.7	77.5	64.1	58.6	45.9	65.6	67.6	60.4	88.8	85.7	76.0	83.5	80.1	82.8
SSTDA [15]	83.0	81.5	73.8	75.8	83.2	79.5	75.0	69.1	55.2	73.7	70.2	68.6	90.0	89.1	78.0	86.2	79.8	84.6
GTRM [33]	75.4	72.8	63.9	67.5	82.6	72.4	57.5	54.0	43.3	58.7	65.0	55.7	-	-	-	-	-	-
BCN [71]	82.3	81.3	74.0	74.3	84.4	79.3	68.7	65.5	55.0	66.2	70.4	65.2	88.5	87.1	77.3	84.4	79.8	83.4
MTDA [14]	82.0	80.1	72.5	75.2	83.2	78.6	74.2	68.6	56.5	73.6	71.0	68.8	90.5	88.4	76.2	85.8	80.0	84.2
Global2Local [28]	80.3	78.0	69.8	73.4	82.2	76.7	74.9	69.0	55.2	73.3	70.7	68.6	89.9	87.3	75.8	84.6	78.5	83.2
HASR [1]	86.6	85.7	78.5	81.0	83.9	83.1	74.7	69.5	57.0	71.9	69.4	68.5	90.9	88.6	76.4	87.5	78.7	84.4
ASRF [36]	84.9	83.5	77.3	79.3	84.5	81.9	74.3	68.9	56.1	72.4	67.6	67.9	89.4	87.8	79.8	83.7	77.3	83.6
ASFormer [74]	85.1	83.4	76.0	79.6	85.6	81.9	76.0	70.6	57.4	75.0	73.5	70.5	90.1	88.8	79.2	84.6	79.7	84.5
UARL [13]	85.3	83.5	77.8	78.2	84.1	81.8	65.2	59.4	47.4	66.2	67.8	61.2	92.7	91.5	82.8	88.1	79.6	86.9
DPRN [51]	87.8	86.3	79.4	82.0	87.2	84.5	75.6	70.5	57.6	75.1	71.7	70.1	92.9	92.0	82.9	90.9	82.0	88.1
SEDT [38]	89.9	88.7	81.1	84.7	86.5	86.2	-	-	-	-	-	-	93.7	92.4	84.0	91.3	81.3	88.5
TCTr [5]	87.5	86.1	80.2	83.4	86.6	84.8	76.6	71.1	58.5	76.1	77.5	72.0	91.3	90.1	80.0	87.9	81.1	86.1
FAMMSDTN [21]	86.2	84.4	77.9	79.9	86.4	83.0	78.5	72.9	60.2	77.5	74.8	72.8	91.6	90.9	80.9	88.3	80.7	86.5
DTL [72]	87.1	85.7	78.5	80.5	86.9	83.7	78.8	74.5	62.9	77.7	75.8	73.9	-	-	-	-	-	-
UVAST [8]	89.1	87.6	81.7	83.9	87.4	85.9	76.9	71.5	58.0	77.1	69.7	70.6	92.7	91.3	81.0	92.1	80.2	87.5
BrPrompt [46]	89.2	87.8	81.3	83.8	88.1	86.0	-	-	-	-	-	-	94.1	92.0	83.0	91.6	81.2	88.4
MCFM [35]	90.6	89.5	84.2	84.6	90.3	87.8	-	-	-	-	-	-	91.8	91.2	80.8	88.0	80.5	86.5
LTContext [6]	89.4	87.7	82.0	83.2	87.7	86.0	77.6	72.6	60.1	77.0	74.2	72.3	-	-	-	-	-	-
DiffAct [48]	90.1	89.2	83.7	85.0	88.9	87.4	80.3	75.9	64.6	78.4	76.4	75.1	92.5	91.5	84.7	89.6	82.2	88.1
ActFusion [29]	91.6	90.7	84.8	86.0	89.3	88.5	81.0	76.2	64.7	79.3	76.4	75.5	94.1	93.3	86.9	91.6	81.9	89.6
HybridTAS (Ours)	92.8	91.8	88.4	89.4	90.6	90.6	82.8	77.9	68.1	81.1	80.2	78.0	97.0	97.0	90.8	95.2	83.5	92.7

Table 1. **Quantitative Results.** Comparison of temporal action segmentation performance on GTEA [23], 50Salads [65], and Breakfast [40] datasets. Best results are denoted in bold.

enforces a minimum separation between all prototype pairs, effectively distributing them across the manifold to preserve

class-wise semantic boundaries. By maximizing inter-class distances in hyperbolic space, it enhances both alignment and separability, as illustrated in Fig. 1.

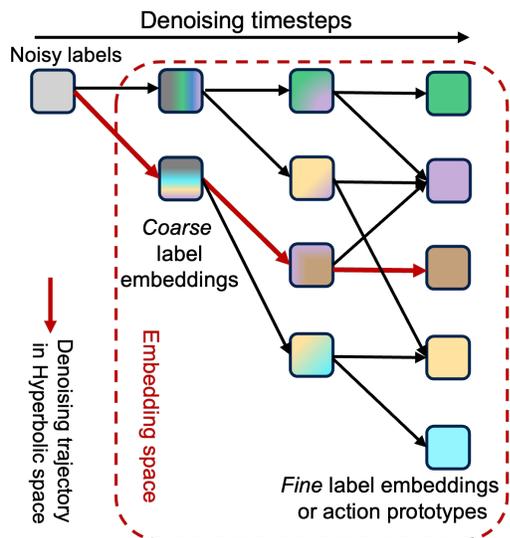


Figure 3. **Denoising trajectory in hyperbolic space.** Our hyperbolic loss functions guide the diffusion model to align its denoising trajectory along the geodesic between the origin and the target action prototype. This enforces the model to follow a hierarchical, coarse-to-fine progression in the label embedding space.

Hyperbolic Push-Pull Loss. In hyperbolic space, an embedding’s distance from the origin naturally encodes prediction confidence, with points farther from the origin indicating lower uncertainty [4]. We leverage this property by encouraging denoised embeddings to move outward (away from the origin) while simultaneously pulling them toward their corresponding prototypes. However, given that diffusion models denoise data in a coarse-to-fine manner [52,53], applying a static distance-based loss is suboptimal. To address this, we introduce a timestep-aware objective that modulates the outward push using an exponential decay, allowing aggressive updates in early steps (lower values of t) and fine-grained alignment near convergence:

$$\mathcal{L}_{pp} = \frac{1}{N} \sum_{i=1}^N \left[\frac{d_{\mathbb{B}}(x_i, z_i)}{d_{\mathbb{B}}(O, x_i)} - d_{\mathbb{B}}(O, x_i) \cdot \exp\left(-\frac{t}{T}\right) \right] \quad (9)$$

Here, x_i is the denoised embedding, z_i its corresponding prototype, O the origin of the Poincaré ball, and t/T represents the normalized timestep. The ratio term prevents prototypes from collapsing toward the origin, while the decay balances directional pull based on denoising progress.

The overall loss in the *Stabilization Phase* can be sum-

marized as:

$$\mathcal{L}_{\text{stable}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{entail}}\mathcal{L}_{\text{entail}} + \lambda_{\text{margin}}\mathcal{L}_{\text{margin}} + \lambda_{\text{pp}}\mathcal{L}_{\text{pp}} \quad (10)$$

5.3. Step 2: Guidance Phase

Following the *Stabilization Phase*, the action prototypes are fixed and serve as semantic anchors in the latent space. In the *Guidance Phase*, the model learns to steer denoised embeddings toward their corresponding prototypes along geometrically meaningful paths (Fig. 3).

Geodesic Guidance Loss. To align the diffusion trajectory with the shortest semantic path on the hyperbolic manifold, we propose a geodesic guidance loss:

$$\mathcal{L}_{\text{gg}} = \frac{1}{N} \sum_{i=1}^N [d_{\mathbb{B}}(O, z_i) - (d_{\mathbb{B}}(O, x_i) + d_{\mathbb{B}}(x_i, z_i))]^2 \quad (11)$$

Here, O denotes the origin, z_i the target prototype, and x_i the denoised embedding. The loss penalizes deviation from the geodesic triangle inequality, encouraging embeddings to evolve along the hyperbolic geodesic connecting the origin to the prototype via the intermediate point x_i . This enforces minimal deviation and nudges the trajectory toward semantically optimal paths.

The overall loss in the *Guidance Phase* can be summarized as:

$$\mathcal{L}_{\text{guidance}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{entail}}\mathcal{L}_{\text{entail}} + \lambda_{\text{gg}}\mathcal{L}_{\text{gg}} \quad (12)$$

5.4. Total Loss

The two-step training objective, where *Step 1* utilizes E_1 epochs and e be the current epoch, can be summarized as:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{stable}}, & \text{if } e < E_1 \\ \mathcal{L}_{\text{guidance}}, & \text{if } e \geq E_1 \end{cases} \quad (13)$$

The integration of Euclidean and hyperbolic losses is crucial: Euclidean loss (\mathcal{L}_{ce}) supervises the model’s output predictions for local accuracy, while hyperbolic losses organize the embedding space for global hierarchy and separation. Euclidean losses alone cannot guarantee that the learned representations are hierarchically meaningful or robust to ambiguous cases, while hyperbolic losses alone cannot ensure frame-level accuracy. By jointly optimizing both sets of objectives, our model achieves accurate, temporally coherent, and boundary-aware action segmentation, underpinned by a geometry-aware, hierarchically structured representation space.

6. Ablation Studies

Extensive ablation studies are performed to validate the design choices in our method on the GTEA dataset [23].

Decaying function. We study the impact of different decaying functions in \mathcal{L}_{pp} , which controls the radial outward movement of embeddings across timesteps in our model. As shown in Table 2, the exponential decay function e^{-x} yields the best performance across all metrics, achieving an average score of 92.7. This function sharply penalizes early errors while allowing more flexibility at later stages of the denoising process. In comparison, both the linear decay and cosine decay underperform slightly, with average scores of 91.2 and 92.0, respectively. These results demonstrate that sharper decay functions better align with the progressive nature of the diffusion trajectory and provide lower uncertainty in predictions.

Decaying function	F1@10	F1@25	F1@50	Edit	Acc	Avg
$1 - x$	95.2	94.4	90.6	92.8	82.9	91.2
$\frac{1}{2}(1 + \cos(\pi x))$	96.0	95.8	90.5	93.6	83.8	92.0
e^{-x}	97.0	97.0	90.8	95.2	83.5	92.7

Table 2. **Decaying function.** We experiment with different decaying functions in \mathcal{L}_{pp} and observe that exponential decay works best, as denoted in **bold**.

Effect of curvature. We empirically evaluate the effect of curvature c in HybridTAS. As summarized in Table 3, the model performance is maximized at $c = 1$. Except for $c = 0.5$, we observe a drop in “Avg” scores on either side of $c = 1$. Low curvatures fail to capture the hierarchy in data, whereas high curvatures can cause numerical instability and distort distances, impacting optimization.

Curvature	F1@10	F1@25	F1@50	Edit	Acc	Avg
0.1	95.8	95.0	89.7	92.7	82.8	91.2
0.3	93.6	93.6	89.1	92.8	81.7	90.2
0.5	95.8	95.8	91.9	95.6	82.8	92.4
0.7	93.1	93.1	87.8	90.0	81.5	89.1
0.9	95.1	95.1	90.5	92.1	83.7	91.3
1.0	97.0	97.0	90.8	95.2	83.5	92.7
2.0	95.1	95.1	91.2	93.1	82.0	91.3

Table 3. **Effect of curvature.** HybridTAS exhibits maximal performance for $c = 1$ on the GTEA dataset [23]. Best results have been **bolded**.

Need for two-phase optimization. During the *Guidance Phase*, the diffusion model corrects its trajectory by aligning the denoised embeddings with the geodesic between the origin and the action prototypes. This requires the action prototypes to be static, as in a two-step optimization strategy. In single-step optimization, the action prototypes be-

have as dynamic targets, thereby making training unstable. Our experiments further validate this, as shown in Table 4.

Method	F1@10	F1@25	F1@50	Edit	Acc	Avg
One-step optimization	93.2	93.0	90.7	89.1	83.3	89.9
Two-step optimization	97.0	97.0	90.8	95.2	83.5	92.7

Table 4. **Optimization strategies.** The two-step optimization strategy enables HybridTAS to iteratively refine its trajectory toward static targets (i.e., action prototypes), leading to improved performance over the one-step approach on the GTEA dataset [23]. Best results have been **bolded**.

Effect of training losses. We conduct ablations exclusively on the hyperbolic loss components, as the cross-entropy loss (\mathcal{L}_{ce}) is fundamental to our framework. Our findings indicate that HybridTAS attains peak performance when all the proposed hyperbolic losses are jointly employed, highlighting their complementary contributions as shown in Table 5.

\mathcal{L}_{tail}	\mathcal{L}_{margin}	\mathcal{L}_{pp}	\mathcal{L}_{gg}	F1@10	F1@25	F1@50	Edit	Acc	Avg
✓		✓	✓	95.8	95.8	91.2	94.6	83.6	92.2
	✓	✓	✓	95.0	94.2	88.8	93.2	82.3	90.7
✓	✓	✓		95.4	95.4	88.6	92.6	82.6	90.9
✓	✓	✓	✓	97.0	97.0	90.8	95.2	83.5	92.7

Table 5. **Effect of training losses.** HybridTAS performs best with all hyperbolic losses, on the GTEA dataset [23]. Best results have been **bolded**.

Effect of inference steps. Based on our experiments with different numbers of inference steps, as reported in Table 6, we observe a steady and marginal improvement in performance as step number increases. Interestingly, we also observe that HybridTAS outperforms ActFusion [29] (25 inference steps) with 66% fewer step numbers. This can be attributed to the improved semantic and hierarchical understanding of the model in the latent space.

7. Quantitative Analysis

Table 1 presents the performance of **HybridTAS** compared to SOTA methods across three benchmark datasets. Our method consistently outperforms prior approaches across all metrics, with particularly substantial improvements on GTEA [23]. On 50Salads [65], HybridTAS achieves a remarkable improvement of **+3.6** in F1@50 and **+3.4** in Edit score over ActFusion [29], highlighting its effectiveness in enhancing temporal consistency and boundary precision. Our average score of **90.6**, a **+2.1** gain over baselines, validates the complementary benefit of jointly optimizing

Inference Steps	F1@10	F1@25	F1@50	Edit	Acc	Avg
1	83.6	83.6	78.4	75.3	83.0	80.8
2	87.6	87.6	82.1	81.1	84.1	84.5
4	93.0	93.0	87.1	88.7	84.2	89.2
8	95.5	95.5	89.4	93.8	83.0	91.4
12	95.6	95.6	90.3	93.9	83.4	91.8
16	96.1	96.1	90.1	94.4	83.8	92.1
20	94.5	94.5	90.4	95.3	84.0	92.6
25	97.0	97.0	90.8	95.2	83.5	92.7
50	96.2	96.0	92.2	96.3	83.7	92.9
100	96.2	96.1	92.4	96.3	83.9	93.0

Table 6. **Inference steps.** We ablate on the number of inference steps and observe that HybridTAS surpasses ActFusion [29] with 66% fewer steps (8 inference steps) on the GTEA dataset [23]. Best results have been **bolded**.

Euclidean and hyperbolic objectives. On Breakfast [40], we observe notable gains in F1@50 (**+3.4**) and Accuracy (**+3.8**), indicating that HybridTAS significantly reduces prediction uncertainty. Improvements in Edit score (**+1.8**) and the overall average score (**+2.5**) further demonstrate our model’s ability to capture fine-grained actions. The most pronounced improvements are observed on GTEA [23], with HybridTAS surpassing ActFusion by **+3.9**, **+3.7**, and **+3.9** in F1@10, F1@25, and F1@50 respectively, and by **+3.6** in Edit. These gains, leading to an average improvement of **+3.1**, underscore the efficacy of our hierarchical supervision in modeling atomic human actions with high fidelity.

Computational cost. Hyperbolic projections introduce a marginal increase in training time compared to DiffAct [48]. On the other hand, inference time remains unchanged as we have leveraged 25 inference steps, similar to previous works.

8. Qualitative Analysis

Fig. 4 presents 2D UMAP projections of action embeddings learned by DiffAct [48] and HybridTAS (Ours). For each of the videos, we compare the Euclidean projection (left) with the hyperbolic projection (right). In both tasks (*Cheese* and *CofHoney*), the hyperbolic embedding produces clearer inter-class boundaries compared to the Euclidean counterpart. For instance, in *S3-Cheese-C1*, the actions *open*, *pour*, and *take* appear as compact, radially separated clusters in the hyperbolic space, whereas their Euclidean projections exhibit elongated and partially overlapping trajectories. **This observation suggests that hyperbolic geometry better preserves the hierarchical separability of action classes and improves boundary definition. The Poincaré embeddings emphasize cluster com-**

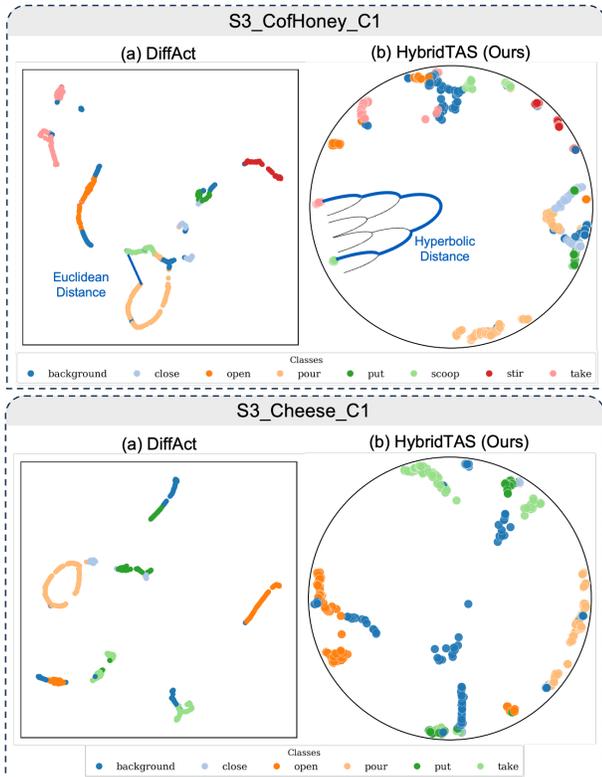


Figure 4. **Euclidean and Hyperbolic UMAP on the GTEA dataset [23].** UMAP projections of action embeddings from DiffAct [48] in Euclidean space (left) and HybridTAS (Ours) in hyperbolic space (right). Hyperbolic embeddings yield well-separated clusters, with background states centralized and specific actions arranged radially. Further, it better preserves inter-class boundaries and highlights hierarchical structure across tasks. Refer to Fig. 6 (Appendix) for cluster centroid specific distances.

pactness, with same-class samples pulled toward localized regions near the boundary. This effect is particularly visible for the *take* and *open* classes, which are tightly grouped in hyperbolic space but more diffusely scattered in Euclidean space. Moreover, hyperbolic spaces implicitly encode a hierarchy: high-frequency background states are centralized, while task-specific actions (e.g., *open*, *pour*, *scoop*) occupy peripheral zones, reflecting their relative semantic specificity. When comparing Cheese and CofHoney tasks, we observe consistent structuring of common classes (*open*, *take*, *pour*). **Despite contextual differences in the tasks, the hyperbolic model projects these actions into analogous radial sectors, suggesting that the representation generalizes across recipes while retaining class separability.** In contrast, Euclidean projections show more task-specific drifts, particularly for the *background* and *put* actions. More samples have been provided in the appendix.

Figure 5 compares the segmentation outputs of DiffAct [48] with HybridTAS (Ours), alongside ground truth anno-

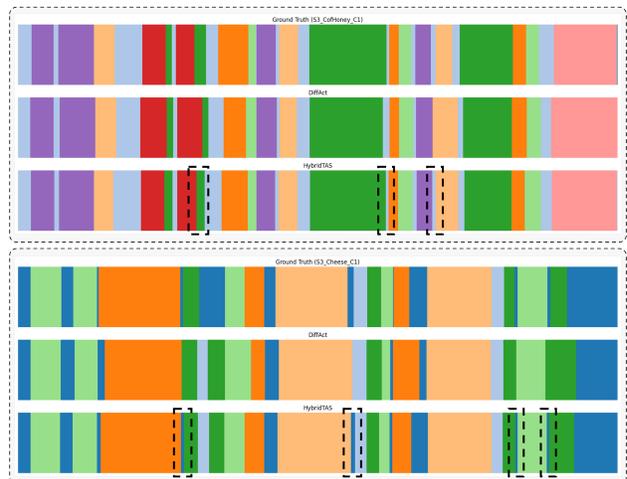


Figure 5. **Qualitative results on the GTEA dataset [23].** We present a comparison of segmentation outputs of DiffAct [48] and HybridTAS (Ours) on *S3_CofHoney_C1* (top) and *S3_Cheese_C1* (bottom) with dashed boxes representing areas of improvement. HybridTAS yields clearer temporal boundaries, better preserves short actions, and maintains semantic consistency across transitions.

tations. In both sequences, HybridTAS demonstrates tighter alignment with ground truth action boundaries compared to DiffAct. Dashed regions highlight instances where DiffAct fails to capture short actions or incorrectly models action boundaries, whereas HybridTAS more faithfully captures transition boundaries between segments. For example, in Cheese, HybridTAS captures short *background* (blue) transitions unlike DiffAct.

9. Conclusion

In this paper, we introduce **HybridTAS**, a novel diffusion-based framework for temporal action segmentation that integrates both Euclidean and hyperbolic geometries to capture the hierarchical structure of actions. By modeling the denoising process along semantically meaningful trajectories in hyperbolic space, HybridTAS enables coarse-to-fine action latent generation. Our two-phase training strategy ensures temporally coherent predictions and provides the model with targets to correct its denoising trajectory. Extensive experiments on GTEA, 50Salads, and Breakfast datasets confirm that HybridTAS outperforms existing diffusion-based models across all standard metrics. Beyond improved segmentation, our approach enables faster convergence in fewer inference steps due to its geometry-aware denoising path.

References

- [1] Hyemin Ahn and Dongheui Lee. Refining action segmentation with hierarchical video representations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16282–16290, 2021. 1, 2, 5
- [2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4575–4583, 2016. 12
- [3] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models, 2022. 2
- [4] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4443–4452, 2022. 1, 2, 3, 5
- [5] Nicolas Aziere and Sinisa Todorovic. Multistage temporal convolution transformer for action segmentation. *Image and Vision Computing*, 128:104567, 2022. 5
- [6] Emad Bahrami, Gianpiero Francesca, and Juergen Gall. How much temporal long-term context is needed for action segmentation? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10317–10327, 2023. 2, 5
- [7] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2022. 2
- [8] Nadine Behrmann, S. Alireza Golestaneh, Zico Kolter, Jürgen Gall, and Mehdi Noroozi. Unified fully andnbsp;timestamp supervised temporal action segmentation vianbsp;sequence tonbsp;sequence translation. In *17th European Conference on Computer Vision Proceedings, Part XXXV*, page 52–68, Berlin, Heidelberg, 2022. Springer-Verlag. 2, 5
- [9] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 14
- [10] Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. From trees to continuous embeddings and back: hyperbolic hierarchical clustering. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2020. Curran Associates Inc. 2
- [11] Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks, 2019. 2
- [12] Bike Chen, Wei Peng, Xiaofeng Cao, and Juha Röning. Hyperbolic uncertainty aware semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):1275–1290, 2024. 2
- [13] Lei Chen, Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Uncertainty-aware representation learning for action segmentation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 820–826, 7 2022. 5
- [14] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan Al-Regib. Action segmentation with mixed temporal domain adaptation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 594–603, 2020. 5
- [15] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zolt Kira. Action Segmentation With Joint Self-Supervised Temporal Domain Adaptation . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9451–9460, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. 1, 2, 5
- [16] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusionet: Diffusion model for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19773–19786, 2023. 2
- [17] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, Sept. 2023. 2
- [18] Ankit Dhall, Anastasia Makarova, Octavian Ganea, Dario Pavllo, Michael Greeff, and Andreas Krause. Hierarchical image classification using entailment cone embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3649–3658, 2020. 3
- [19] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2021. Curran Associates Inc. 2
- [20] Anh-Dung Dinh, Daochang Liu, and Chang Xu. PixelAs-Param: A gradient view on diffusion sampling with guidance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8120–8137. PMLR, 23–29 Jul 2023. 2
- [21] Zexing Du and Qing Wang. Dilated transformer with feature aggregation module for action segmentation. *Neural Process. Lett.*, 55(5):6181–6197, Dec. 2022. 5
- [22] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7399–7409, 2022. 2
- [23] Alireza Fathi, Xiaofeng Ren, and James M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, pages 3281–3288, 2011. 1, 5, 6, 7, 8, 12, 13, 14
- [24] Alessandro Flaborea, Bardh Prenkaj, Bharti Munjal, Marco Aurelio Sterpa, Dario Aragona, Luca Podo, and Fabio Galasso. Are we certain it’s anomalous? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2897–2907, 2023. 2
- [25] Luca Franco, Paolo Mandica, Konstantinos Kallidromitis, Devin Guillory, Yu-Teng Li, Trevor Darrell, and Fabio Galasso. Hyperbolic active learning for semantic segmentation under domain shift. In *Proceedings of the 41st Interna-*

- tional Conference on Machine Learning, ICML. JMLR.org, 2024.* 2
- [26] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations, 2023. 2
- [27] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS*, page 5350–5360, Red Hook, NY, USA, 2018. Curran Associates Inc. 2
- [28] Shang-Hua Gao, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, and Ming-Ming Cheng. Global2local: Efficient structure search for video action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16800–16809, 2021. 1, 2, 5
- [29] Dayoung Gong, Suha Kwak, and Minsu Cho. Actfusion: a unified diffusion model for action segmentation and anticipation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS*, Red Hook, NY, USA, 2024. Curran Associates Inc. 1, 2, 5, 7
- [30] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10686–10696, 2022. 2
- [31] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks, 2018. 2
- [32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS*, Red Hook, NY, USA, 2020. Curran Associates Inc. 2, 14
- [33] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5
- [34] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling, 2022. 2
- [35] Kenta Ishihara, Gaku Nakano, and Tetsuo Inoshita. Mefm: Mutual cross fusion module for intermediate fusion-based action segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 1701–1705, 2022. 5
- [36] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2321–2330, 2021. 1, 2, 5
- [37] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV Thumos Workshop*, volume 1, page 5, 2014. 2
- [38] Gyeong-hyeon Kim and Eunwoo Kim. Stacked encoder–decoder transformer with boundary smoothing for action segmentation. *Electronics Letters*, 58, 11 2022. 5
- [39] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2416–2425, 2022. 2
- [40] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014. 1, 2, 5, 7, 13, 14
- [41] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163:78–89, 2017. Language in Vision. 2
- [42] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person Image Synthesis via Denoising Diffusion Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5976, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. 2
- [43] Max W. Y. Lam, Jun Wang, Dan Su, and Dong Yu. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis, 2022. 2
- [44] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012, 2017. 2
- [45] Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiang-Yang Li, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis, 2022. 2
- [46] Muheng Li, Lei Chen, Yueqi Duarr, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-Prompt: Towards Ordinal Action Understanding in Instructional Videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19848–19857, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. 5
- [47] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6647–6658, 2023. 2, 5, 14
- [48] Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10105–10115, 2023. 1, 2, 4, 5, 7, 8, 12, 13, 14
- [49] Qi Liu, Maximilian Nickel, and Douwe Kiela. *Hyperbolic graph neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019. 2
- [50] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS*, page 6341–6350, Red Hook, NY, USA, 2017. Curran Associates Inc. 2, 3

- [51] Junyong Park, Daekyum Kim, Sejoon Huh, and Sungho Jo. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. *Pattern Recognition*, 129:108764, Sept. 2022. [5](#)
- [52] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2023. Curran Associates Inc. [5](#)
- [53] Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models, 2023. [5](#)
- [54] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201, 2012. [2](#)
- [55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, Los Alamitos, CA, USA, June 2022. [2](#)
- [56] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4460–4469. PMLR, 10–15 Jul 2018. [2](#)
- [57] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Conference on Graph Drawing*, GD, page 355–366, Berlin, Heidelberg, 2011. Springer-Verlag. [2](#)
- [58] Ryohei Shimizu, YUSUKE Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *International Conference on Learning Representations*, 2021. [2](#)
- [59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2256–2265. PMLR, 2015. [2](#)
- [60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. [2](#)
- [61] Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019. [2](#)
- [62] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2020. Curran Associates Inc. [2](#)
- [63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. [2](#)
- [64] Yaser Souri, Yazan Abu Farha, Fabien Despinoy, Gianpiero Francesca, and Juergen Gall. Fifa: Fast inference approximation for action segmentation, 2021. [1](#), [2](#)
- [65] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, page 729–738, New York, NY, USA, 2013. Association for Computing Machinery. [1](#), [5](#), [7](#), [13](#), [14](#)
- [66] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12602–12612, 2021. [2](#)
- [67] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. Diffusing surrogate dreams of video scenes to predict video memorability, 2022. [2](#)
- [68] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019. [2](#)
- [69] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: masked conditional video diffusion for prediction, generation, and interpolation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS, 2022. [2](#)
- [70] Yunke Wang, Xiyu Wang, Anh-Dung Dinh, Bo Du, and Charles Xu. Learning to schedule in diffusion probabilistic models. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD, page 2478–2488, New York, NY, USA, 2023. Association for Computing Machinery. [2](#)
- [71] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *16th European Conference on Computer Vision, Proceedings, Part XXV*, page 34–51, Berlin, Heidelberg, 2020. Springer-Verlag. [5](#)
- [72] Ziwei Xu, Yogesh S Rawat, Yongkang Wong, Mohan S. Kankanhalli, and Mubarak Shah. Don’t pour cereal into coffee: differentiable temporal logic for temporal action segmentation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2022. Curran Associates Inc. [1](#), [2](#), [5](#)
- [73] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation, 2022. [2](#)
- [74] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation, 2021. [2](#), [5](#), [14](#)
- [75] Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, and Ying Nian Wu. Latent diffusion energy-based model for interpretable text modeling. In *Proceedings of International Conference on Machine Learning (ICML)*, July 2022. [2](#)
- [76] Xian Zhong, Zipeng Li, Shuqin Chen, Kui Jiang, Chen Chen, and Mang Ye. Refined semantic enhancement towards frequency diffusion for video captioning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence and 35th Conference on Innovative Applications of Artificial Intelligence and 13th Symposium on Educational Advances in Artificial Intelligence*, AAAI/IAAI/EAAI. AAAI Press, 2023. [2](#)