

Video Summarization Using Denoising Diffusion Probabilistic Model

Zirui Shang¹, Yubo Zhu¹, Hongxi Li¹, Shuo Yang², Xinxiao Wu^{1, 2*}

¹Beijing Key Laboratory of Intelligent Information Technology, School of Computer Science & Technology, Beijing Institute of Technology, China

²Guangdong Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, China
{shangzirui, 3120211052, lihongxi, wuxinxiao}@bit.edu.cn, yangshuo@smbu.edu.cn

Abstract

Video summarization aims to eliminate visual redundancy while retaining key parts of video to construct concise and comprehensive synopses. Most existing methods use discriminative models to predict the importance scores of video frames. However, these methods are susceptible to annotation inconsistency caused by the inherent subjectivity of different annotators when annotating the same video. In this paper, we introduce a generative framework for video summarization that learns how to generate summaries from a probability distribution perspective, effectively reducing the interference of subjective annotation noise. Specifically, we propose a novel diffusion summarization method based on the Denoising Diffusion Probabilistic Model (DDPM), which learns the probability distribution of training data through noise prediction, and generates summaries by iterative denoising. Our method is more resistant to subjective annotation noise, and is less prone to overfitting the training data than discriminative methods, with strong generalization ability. Moreover, to facilitate training DDPM with limited data, we employ an unsupervised video summarization model to implement the earlier denoising process. Extensive experiments on various datasets (TVSum, SumMe, and FPVSum) demonstrate the effectiveness of our method.

Introduction

With the popularity of video-sharing platforms and social media, video data is experiencing explosive growth, and increasing attentions are paid to automatically identifying and extracting representative segments from a video. Video summarization emerges as a critical technique that condenses video content into a concise summary while preserving essential information and key moments. These summaries enable users to quickly grasp the video content, making it more easily accessible and manageable in various scenarios, such as online platforms, surveillance systems, and multimedia archives.

Considerable progress has been made in video summarization recently, and many deep models have been applied, such as Long Short Term Memory (LSTM) (Zhang et al. 2016), Fully Convolutional Network (FCN) (Liang et al. 2022), Transformer (Li et al. 2023; Hsu, Liao, and Huang

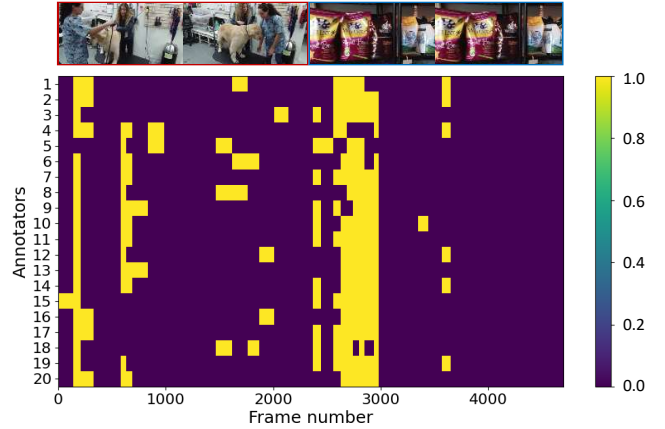


Figure 1: An example of subjective annotation noise in the TVSum dataset, where yellow blocks represent the annotated video frames of summaries by different annotators.

2023; Terbouche et al. 2023), etc. However, these existing methods use discriminative models to predict the importance scores of video frames, which are susceptible to annotation inconsistency. The inconsistency is inevitable because of the inherent subjectivity of different annotators when annotating the same video. For example, as shown in Figure 1, when facing the video of a pet store advertisement, some annotators may focus on the cute pets in the video (shown in the red box), and others may show more interest in the products in the store (shown in the blue box), resulting in different annotations of importance scores.

To address this challenge, we introduce a generative framework for video summarization that learns how to generate summaries from a probability distribution perspective, effectively reducing the interference of subjective annotation noise. In this paper, we propose a novel diffusion summarization method based on the Denoising Diffusion Probabilistic Model (DDPM), which learns the probability distribution of training data through noise prediction, and generates summaries by iterative denoising. Specifically, we use video frame features as guidance, and take the importance scores after adding noise as input to train DDPM for generating clear importance scores. We directly train DDPM using raw annotations rather than averaging them, enabling

*Xinxiao Wu is the corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

it to learn the distribution of training data. Therefore, our method is more resistant to subjective annotation noise and simultaneously less prone to overfitting the training set, with stronger generalization ability than existing discriminative methods.

Moreover, we observe that DDPM faces performance degradation with limited training data, and combine an unsupervised video summarization model with it to address this issue. During training, we reduce the number of maximum noise addition steps, allowing DDPM to start denoising from a relatively clear intermediate result instead of a Gaussian noise. During testing, we use the output of the unsupervised video summarization model as the intermediate result mentioned above, providing an alternative to early denoising and achieving superior performance. This strategy provides a feasible solution for DDPM training under data scarcity.

The main contributions of our work are summarized in three folds: (1) We introduce a generative framework for video summarization, which learns to generate summaries by modeling the probability distribution of training data, effectively reducing the interference of subjective annotation noise. (2) We propose a novel diffusion summarization method based on DDPM, which generates summaries by iterative denoising, and combines it with an unsupervised video summarization model to facilitate training with limited data. (3) The effectiveness of our method is demonstrated through comprehensive experiments on three benchmark datasets: TVSum, SumMe, and FPVSum.

Related Work

Video Summarization

Early traditional methods of video summarization rely on low-level visual features to extract key frames, such as color histogram (De Avila et al. 2011), spatiotemporal feature (Laganière et al. 2008), and motion cues (Ren et al. 2017). With the considerable progress of deep learning in video processing and understanding, video summarization is formulated as an importance score prediction problem. Zhang et al. (Zhang et al. 2016) use LSTM to predict importance scores by modeling the variable-range temporal dependency among video frames. Zhao et al. (Zhao, Li, and Lu 2017) propose a hierarchical LSTM tailored to long video scenarios. More recently, the Graph Neural Networks (GNNs), Convolutional Neural Networks (CNNs), and the attention mechanism have been employed in video summarization. Zhong et al. (Zhong et al. 2023) propose a Contextual Feature Transformation (CFT) mechanism that builds upon Graph Information Bottle (GIB) to enhance the temporal correlation among images. Hsu et al. (Hsu, Liao, and Huang 2023) propose a novel transformer-based method named spatiotemporal vision transformer (STVT) for video summarization, which considers both of inter-frame correlations among non-adjacent frames and intra-frame attention which attracts humans. MPFN (Khan et al. 2024) proposes a deep pyramidal refinement network to extract and refine multi-scale progressive features. LDH-based Deep CNN (Singh and Kumar 2024) selects representative frames

using Bayesian fuzzy clustering (BFC) and refines those frames using deep CNN.

All these methods rely on predicting the importance score of each frame from a discriminative perspective to generate summaries, which may be affected by subjective annotation noise. To reduce this interference, we attempt to generate summaries from a probability distribution perspective and introduce a generative framework for video summarization. Unlike existing generative models (He et al. 2019; Mahaseni, Lam, and Todorovic 2017; Apostolidis et al. 2020b) commonly used for unsupervised video summarization, our method introduces DDPM into supervised video summarization, which progressively pinpoints important content in the video and can predict more precise importance scores (Li et al. 2023).

Denoising Diffusion Probabilistic Model

DDPM belongs to a category of generative models, which has emerged as super performers in producing high-quality samples. It is initially used in the field of image generation. Inspired by considerations from nonequilibrium thermodynamics, Ho et al. (Ho, Jain, and Abbeel 2020) present high-quality image synthesis results by using a diffusion probabilistic model. Subsequently, Nichol et al. (Nichol and Dhariwal 2021) show that with a few simple modifications, DDPM can sample much faster and achieve better log-likelihoods with little impact on sample quality. Recently, DDPM has also been widely used in various fields, including image colorization (Carrillo et al. 2023), super-resolution (Li et al. 2022; Gao et al. 2023), image editing (Wu and De la Torre 2023), and semantic segmentation (Wu and De la Torre 2023).

We make the first attempt to introduce DDPM into the field of video summarization and use video frame features as guidance to recover the corresponding frame-level importance scores from noise. Extensive experiments on various datasets demonstrate that DDPM can be successfully used in video summarization.

Our Method

Background

In this section, we provide a brief overview of diffusion models. Diffusion models aim to convert the noise x_t sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ into the desired output x_0 , under the assuming that x_t is equivalent to adding t steps Gaussian noise to x_0 , with fixed variance schedule $\beta_1, \dots, \beta_t$. Formally, this is defined as a forward diffusion process:

$$\begin{aligned} x_t &= \sqrt{1 - \beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), \end{aligned} \quad (1)$$

where $q(x_t|x_{t-1})$ is the distribution of x_t under the condition of given x_{t-1} . Importantly, a noisy sample x_t can be obtained directly from the x_0 :

$$\begin{aligned} x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ q(x_t|x_0) &= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \end{aligned} \quad (2)$$

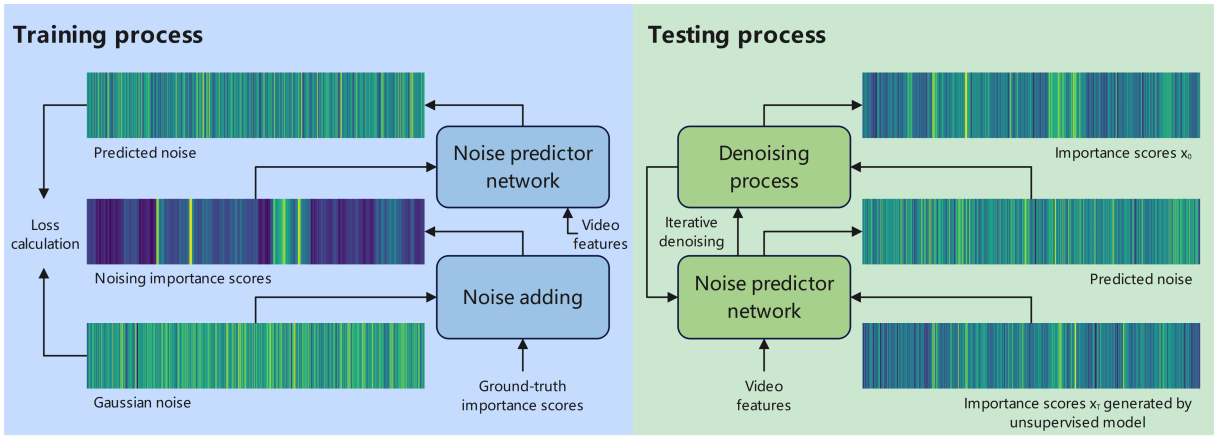


Figure 2: The framework of our method, where the training process shows how the noise predictor network learns to predict noise components, and the testing process shows how to generate accurate importance scores through denoising.

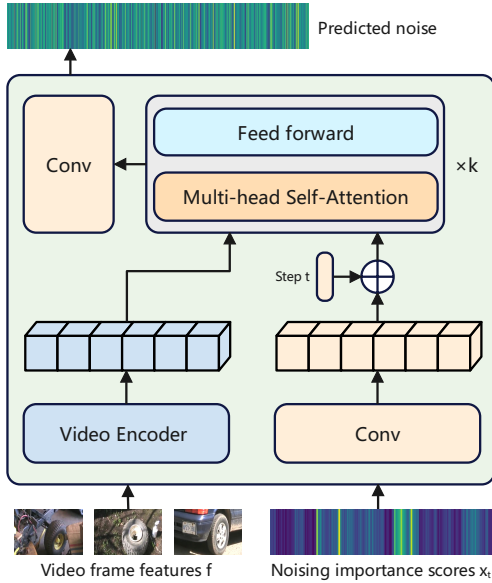


Figure 3: The structure of noise predictor network, which uses video frame features f as guidance and noising importance scores x_t as input to predict the noise component at step t .

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, and $q(x_t|x_0)$ is the distribution of x_t under the condition of given x_0 .

In particular, the diffusion model reverses the noising process, sampling from a distribution by gradually denoising, which starts with noise x_t and learns to produce a slightly more “denoised” x_{t-1} from x_t , until reaching the final sample x_0 . Using Bayes theorem, Ho et al. (Ho, Jain, and Abbeel 2020) define a forward process posteriors, which are tractable when conditioned on x_0 :

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t \mathbf{I}), \quad (3)$$

$$\text{where } \tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \tilde{\beta}_t =$$

$\frac{1-\alpha_{t-1}}{1-\bar{\alpha}_t}\beta_t$, and $q(x_{t-1}|x_t, x_0)$ is the distribution of x_{t-1} under the condition of given x_t and x_0 . Unlike the x_t sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, x_0 is unknown and cannot be obtained through sampling. Thus, to obtain the exact distribution $q(x_{t-1}|x_t)$, and run the forward diffusion process in reverse to obtain a distribution of x_0 , a neural network is used to approximate $q(x_{t-1}|x_t)$ as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)), \quad (4)$$

where $\epsilon_\theta(x_t, t)$ is a noise predictor network to predict the noise component at the step t , $\Sigma_\theta(x_t, t)$ is a covariance predictor can be either a fixed set of scalar covariances or learned as well, and $p_\theta(x_{t-1}|x_t)$ is the approximate distribution of x_{t-1} under the condition of given x_t . The noise predictor network $\epsilon_\theta(x_t, t)$ is usually selected from different variants of the UNet (Ronneberger, Fischer, and Brox 2015) architecture, and with a simple mean-squared error loss.

DDPM-based Video Summarization

In the following sections, we describe our diffusion summarization method from two stages: the training process, which involves how the noise predictor network learns to predict noise components, and the testing process, which involves how to generate accurate importance scores through denoising. Figure 2 illustrates the framework of our method.

Training Process In the training process, we consider how the noise predictor network learns to predict noise components. The noise predictor network is defined as $\epsilon_\theta(x_t, f, t)$, which uses video frame features f as guidance and noising importance scores x_t as input to predict the noise component at step t . The structure of noise predictor network is shown in Figure 3. We use Transformer layers to build a bridge between importance scores and video frame features through attention mechanism. The importance scores are used as query vectors and the video frame features are used as key and value vectors to guide the model in predicting

Algorithm 1: Training Process

Input: Video features with annotated importance scores.
Output: Parameters θ of the noise predictor network.

```
for epoch  $\in \{1, 2, \dots, M\}$  do
  for  $n \in \{1, 2, \dots, N\}$  do
    Extract video features  $f$  of the  $n$ -th video.
    for every importance scores annotation  $x_g$  do
       $t \sim \text{Uniform}(\{1, \dots, T\})$ ,  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .
      Predict noise  $\hat{\epsilon}$  using Eq. 5.
      Calculate the loss  $L$  using Eq. 6.
      Optimize the parameters  $\theta$ :  $\theta \leftarrow \theta - \eta \partial L / \partial \theta$ .
    end for
  end for
end for
```

noise components. In addition, we linearly scale the ground-truth importance scores to $[-1, 1]$ before adding noise. This ensures that the neural network reverse process operates on consistently scaled inputs starting the noise predictor network standard normal prior (Ho, Jain, and Abbeel 2020). Then a complete noise prediction in the training process can be defined as follows:

$$\begin{aligned} x_0 &= \text{Scale}(x_g), \\ x_t &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ \hat{\epsilon} &= \epsilon_\theta(x_t, f, t), \end{aligned} \quad (5)$$

where x_g represents the ground-truth importance scores, $\text{Scale}(\cdot)$ represents the linearly scaling to $[-1, 1]$, x_0 represents the importance scores after scaling, x_t represents the noising importance scores and $\hat{\epsilon}$ represents the predicted noise component.

The training objective is defined as a simple mean-squared error loss between the true noise and the predicted noise as follows:

$$L = \|\epsilon - \hat{\epsilon}\|^2, \quad (6)$$

where ϵ represents the true noise and L represents the loss. Algorithm 1 displays the complete training process, with M epochs, N videos, and maximum noise addition steps T . For each video, we traverse the importance scores annotated by different annotators, input it into the noise predictor network along with the corresponding video features, and optimize parameters through the loss function.

Testing Process In the testing process, we consider how to generate accurate importance scores through denoising. We observe that DDPM faces performance degradation with limited training data. To address this issue, we combine an unsupervised video summarization model (Zhou, Qiao, and Xiang 2018) with DDPM, using its output as the start of denoising, providing an alternative to early denoising of DDPM. In addition, before denoising, we linearly scale the importance scores generated by the unsupervised model to $[-1, 1]$ to ensure consistency with the training process. Formally, it can be defined as:

$$\begin{aligned} x_u &= \text{Unsuper}(f), \\ x_T &= \text{Scale}(x_u), \end{aligned} \quad (7)$$

Algorithm 2: Testing Process

Input: Video features.
Output: Video summary.

```
for  $n \in \{1, 2, \dots, N\}$  do
  Extract video features  $f$  of the  $n$ -th video.
  Generate importance scores  $x_T$  using Eq. 7.
  for  $t \in \{T, \dots, 1\}$  do
     $z \sim \mathcal{N}(0, 1)$  if  $t > 1$ , else  $z = 0$ .
    Denoise using Eq. 8.
  end for
  Generate video summary through importance scores  $x_0$ .
end for
```

where $\text{Unsuper}(\cdot)$ represents the unsupervised model, x_u represents its generated importance scores, and x_T represents the importance scores after scaling.

To adapt to the introduction of the unsupervised model, we reduce the maximum noise addition steps, allowing DDPM to start denoising from a relatively clear intermediate result instead of a Gaussian noise. Specifically, we treat the noising result x_t in Eq. 5 as a weighting of importance scores and noise, where $\sqrt{\bar{\alpha}_t}$ is the weight of importance scores and $\sqrt{1 - \bar{\alpha}_t}$ is the weight of the noise. When the maximum noise addition steps set to 1000 and the variance schedule $\beta_1, \dots, \beta_t$ linearly increasing from 10^{-4} to 0.02 (Ho, Jain, and Abbeel 2020), the value of $\sqrt{\bar{\alpha}_t}$ is close to 0, while the value of $\sqrt{1 - \bar{\alpha}_t}$ is close to 1, thus the noising result is approximate to Gaussian noise. We adjust the weight of importance scores and noise by reducing the maximum noise addition steps so that the value of $\sqrt{\bar{\alpha}_t}$ gradually increases, while the value of $\sqrt{1 - \bar{\alpha}_t}$ gradually decreases, thus achieving a relatively clear intermediate noising result.

A single denoising process can be defined as follows:

$$\begin{aligned} x_{t-1} &= \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, f, t) \right) + \sigma_t z, \\ \sigma_t &= \sqrt{\frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}} \beta_t, \end{aligned} \quad (8)$$

where z is the random noise sampled from a standard normal distribution, and the complete testing process is shown in Algorithm 2. For each video, we use the output of the unsupervised model as the starting point and generate more accurate importance scores through gradual denoising.

Experiment

Dataset

We conduct experiments on three benchmark datasets: TVSum (Song et al. 2015), SumMe (Gong et al. 2014) and FPVSum (Ho, Chiu, and Wang 2018), in which TVSum consists of 50 videos, SumMe consists of 25 videos, and FPVSum consists of 56 labeled videos and 42 unlabeled videos. Following the protocol in (Zhang et al. 2016), we build three settings for TVSum and SumMe: canonical, augmented and transfer. "canonical" is the standard supervised learning setting that divides the dataset into a training set and

Method	TVSum			SumMe		
	Can	Aug	Tran	Can	Aug	Tran
dppLSTM (Zhang et al. 2016)	54.7	59.6	58.7	38.6	42.9	41.8
VASNet (Fajtl et al. 2019)	61.4	62.4	-	49.7	51.1	-
SUM-FCN (Rochan, Ye, and Wang 2018)	56.8	59.2	58.2	47.5	51.1	44.1
A-AVS (Ji et al. 2019)	59.4	60.8	-	43.9	44.6	-
M-AVS (Ji et al. 2019)	61.0	61.8	-	44.4	46.1	-
DSNet _{ab} (Zhu et al. 2020)	62.1	63.9	59.4	50.2	50.7	46.5
DSNet _{af} (Zhu et al. 2020)	61.9	62.2	58.0	51.2	53.3	47.6
RSGN (Zhao et al. 2021)	60.1	61.1	60.0	45.0	45.7	44.0
3DST-UNet _{sup} (Liu et al. 2022)	58.3	58.9	56.1	47.4	49.9	47.9
RR-STG (Zhu et al. 2022)	63.0	63.6	59.7	53.4	54.8	45.4
CFT-GIB _{sup} (Zhong et al. 2023)	62.7	60.3	58.0	56.0	54.8	44.0
VSS-Net (Zhang et al. 2023)	61.0	61.4	58.5	51.5	52.8	48.4
LMVS (Nam et al. 2024)	60.5	-	-	45.8	-	-
MPFN (Khan et al. 2024)	62.4	-	-	51.9	-	-
SUM-GAN (Mahasseni, Lam, and Todorovic 2017)	50.8	58.9	-	38.7	41.7	-
DR-DSN (Zhou, Qiao, and Xiang 2018)	57.6	58.4	57.8	41.4	42.8	42.4
ACGAN (He et al. 2019)	58.5	58.9	57.8	46.0	47.0	44.5
SUM-GAN-AAE (Apostolidis et al. 2020b)	58.3	-	-	48.9	-	-
3DST-UNet _{unsup} (Liu et al. 2022)	58.3	58.4	58.0	44.6	49.5	45.7
CFT-GIB _{unsup} (Zhong et al. 2023)	62.5	60.4	57.4	55.0	54.0	42.9
SSPVS (Li et al. 2023)	60.3	61.8	57.8	48.7	50.4	45.8
AMFM (Zhang, Liu, and Wu 2024)	61.0	60.8	58.6	51.8	52.8	46.4
PRLVS (Wang, Wu, and Yan 2024)	63.0	59.2	57.0	46.3	49.7	47.6
DMFF (Yu et al. 2024)	61.2	-	-	53.0	-	-
Ours	64.8	65.0	60.9	58.7	59.2	50.3

Table 1: Performance comparison (F-score) with state-of-the-art video summarization methods on the TVSum and SumMe datasets under the canonical (Can), augmented (Aug), and transfer (Tran) settings.

Method	F-score
Random (Ho, Chiu, and Wang 2018)	16.3
Uniform (Ho, Chiu, and Wang 2018)	15.1
TDCNN (Yao, Mei, and Rui 2016)	28.6
DSN (Bousmalis et al. 2016)	22.7
FPVS (Ho, Chiu, and Wang 2018)	35.3
Ours	46.1

Table 2: Performance comparison (F-score) with state-of-the-art video summarization methods on the FPVSum dataset.

a testing set. "augmented" additionally introduces YouTube dataset (De Avila et al. 2011) and Open Video Project (OVP) dataset (De Avila et al. 2011) to explore the impact of the increased amount of annotations on performance. "transfer" divides completely different datasets into a training set and a testing set to simulate the cross-domain scenarios. Following the protocol in (Ho, Chiu, and Wang 2018), we build the FPVSum setting as a supplement to the transfer setting. We divide videos with different points of view, introducing the third-person videos in TVSum and SumMe into the training set, and the first-person videos into the testing set. We perform validation experiments with 5 randomly created data splits and report the average results.

Method	TVSum		SumMe	
	τ	ρ	τ	ρ
dppLSTM	0.042	0.055	0.071	0.101
HSA	0.082	0.088	0.064	0.066
DSNet _{ab}	0.108	0.129	0.051	0.059
DSNet _{af}	0.113	0.138	0.037	0.046
RSGN	0.083	0.090	0.083	0.085
SSPVS	0.177	0.233	0.178	0.240
Ours	0.179	0.238	0.221	0.252

Table 3: Performance comparison (Kendall's τ and Spearman's ρ correlation coefficients) with state-of-the-art video summarization methods on the TVSum and SumMe datasets.

Implementation Detail

The training and testing processes are implemented using Pytorch. In training, the maximum step of noise addition is set to 200, and the ground-truth of importance scores are normalized to the range of -1 to 1 before adding Gaussian noise. In testing, the denoising step is set to 200, and the inverse process of training is conducted, which scales the generated importance scores to the range of 0 to 1. In addition, we use Adam as the optimizer and set the learning rate to 0.0002 and the weight decay to 0.01. The model is trained

Method	TVSum			SumMe			FVPSum
	Can	Aug	Tran	Can	Aug	Tran	
w/o DDPM	62.0	61.5	55.3	51.1	51.1	46.5	39.1
w/o unsup	64.1	62.7	57.6	51.9	57.5	48.1	42.3
Ours	64.8	65.0	60.9	58.7	59.2	50.3	46.1

Table 4: Ablation study results (F-score) on the TVSum, SumMe, and FVPSum datasets.

in 100 epochs, and a warmup strategy is used in the first 10 epochs.

Quantitative Evaluation

Comparison Results with the State-of-the-art Methods

We compare our method with several state-of-the-art methods under different settings, including supervised methods (dppLSTM (Zhang et al. 2016), VASNet (Fajtl et al. 2019), SUM-FCN (Rochan, Ye, and Wang 2018), A-AVS (Ji et al. 2019), M-AVS (Ji et al. 2019), DSNet (Zhu et al. 2020), RSGN (Zhao et al. 2021), 3DST-UNet_{sup} (Liu et al. 2022), RR-STG (Zhu et al. 2022), CFT-GIB_{sup} (Zhong et al. 2023), VSS-Net (Zhang et al. 2023), LMVS (Nam et al. 2024), MPFN (Khan et al. 2024), TDCNN (Yao, Mei, and Rui 2016), DSN (Bousmalis et al. 2016) and FPVS (Ho, Chiu, and Wang 2018)), and unsupervised methods (SUM-GAN (Mahasseni, Lam, and Todorovic 2017), DR-DSN (Zhou, Qiao, and Xiang 2018), ACGAN (He et al. 2019), SUM-GAN-AAE (Apostolidis et al. 2020b), 3DST-UNet_{unsup} (Liu et al. 2022), CFT-GIB_{unsup} (Zhong et al. 2023), SSPVS (Li et al. 2023), AMFM (Zhang, Liu, and Wu 2024), PRLVS (Wang, Wu, and Yan 2024), DMFF (Yu et al. 2024) Random (Ho, Chiu, and Wang 2018) and Uniform (Ho, Chiu, and Wang 2018)).

Table 1 and 2 show the comparison results on the TVSum, SumMe, and FVPSum datasets. It is interesting to observe that our method performs best under all the settings on all the datasets, which indicates that compared with the discriminative methods, our generative method is more resistant to subjective annotation noise. In addition, our method achieves much better results with large gains on the transfer settings on SumMe and TVSum as well as FVPSum dataset, which suggests that our method is less prone to overfitting the training data and has stronger generalization ability.

Besides the F-Score, Kendall’s τ and Spearman’s ρ correlation coefficients are applied in (Otani et al. 2019) as alternatives for assessing the importance score. Table 3 shows the comparison results of our method with the state-of-the-art methods using Kendall’s τ and Spearman’s ρ correlation coefficients on TVSum and SumMe, including dppLSTM (Zhang et al. 2016), HSA (Zhao, Li, and Lu 2018), DSNet_{ab} (Zhu et al. 2020), DSNet_{af} (Zhu et al. 2020), RSGN (Zhao et al. 2021), SSPVS (Li et al. 2023). We observe that the proposed method also achieve the best result, which further demonstrates the effectiveness of our method.

Ablation Study To perform an in-depth analysis of each individual component of our method, we conduct extensive ablation studies on TVSum, SumMe, and FVPSum. We em-

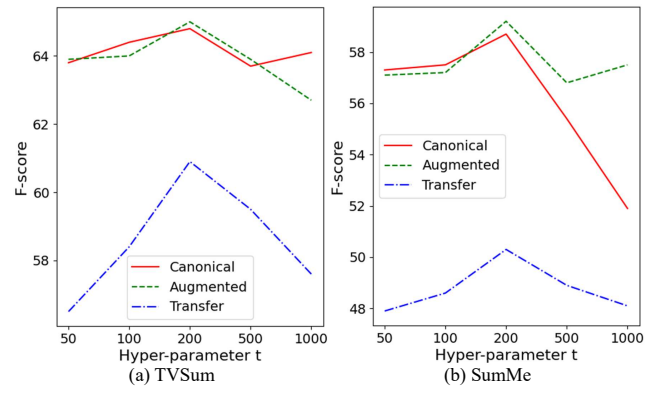


Figure 4: Results (F-score) of experiment with different hyper-parameter t on the TVSum and SumMe datasets.

Method	TVSum			SumMe		
	Can	Aug	Tran	Can	Aug	Tran
SUM-GAN-AAE	64.1	63.9	58.4	56.9	58.1	47.6
AC-SUM-GAN	63.5	65.3	58.5	56.1	58.5	48.8
DR-DSN	64.8	65.0	60.9	58.7	59.2	50.3

Table 5: Results (F-score) of experiment with different unsupervised methods on the TVSum and SumMe datasets.

ploy variants of our method for comparison, including “w/o DDPM” and “w/o unsup”. “w/o DDPM” represents removing DDPM and using only the unsupervised video summarization model, in order to evaluate the effectiveness of DDPM. “w/o unsup” represents removing the unsupervised video summarization model and using only DDPM, in order to evaluate the effectiveness of unsupervised video summarization model, which is achieved by setting the maximum noise addition steps to 1000 during training, and using Gaussian noise as the starting point for the denoising process during testing. From the results in Table 4, we observe that our method achieves better performance than “w/o DDPM”, indicating that DDPM succeeds in generating high-quality summaries by the denoising process. Our method also has advantages over “w/o unsup”, indicating that using an unsupervised method to replace the early denoising process of DDPM is an effective strategy under data scarcity.

In addition, we evaluate the impact of different unsupervised models (SUM-GAN-AAE (Apostolidis et al. 2020b), AC-SUM-GAN (Apostolidis et al. 2020a), and DR-DSN (Zhou, Qiao, and Xiang 2018)) on TVSum and SumMe. The results are shown in Table 5 and we choose the DR-DSN with the best overall performance as the unsupervised model for our method. We observe that in some settings, different unsupervised models exhibit similar performance, indicating that our method is not limited to a specific unsupervised model and has robustness to different unsupervised models.

Analysis of Hyper-parameter To analyze the denoising ability of DDPM in video summarization, we evaluate the impact of hyper-parameter t on TVSum and SumMe. The re-

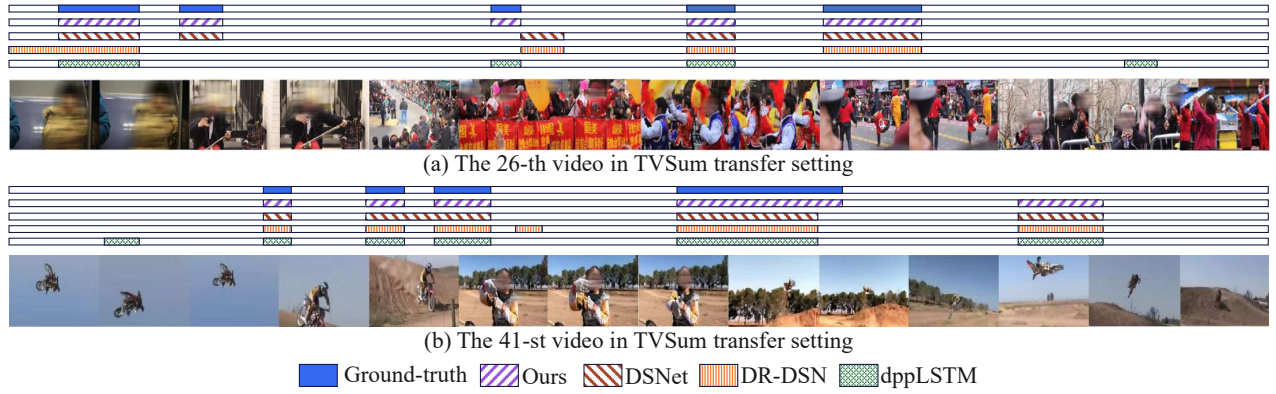


Figure 5: Qualitative results of different video summarization methods. The line segments denote the selected segments and the frames are shown below.

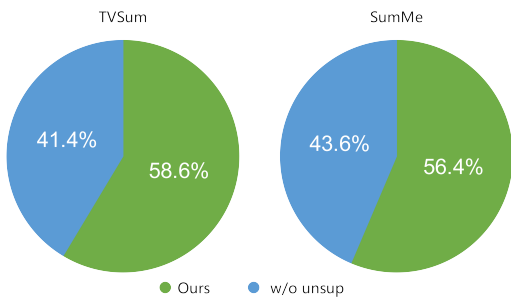


Figure 6: Human evaluation of video summaries generated by 'w/o unsup' and our method. The blue sector represents the percentage of summaries chosen by users from 'w/o unsup' and the green sector represents the percentage of summaries chosen by users from our method.

sults are shown in Figure 4. When $t=1000$, DDPM denoises from Gaussian noise without unsupervised model. It is interesting to observe that for most settings, our method achieves the best performance when t equals 200. This indicates that the importance scores generated by the unsupervised video summarization model can be approximated as the ground-truth with 200 steps of noise addition, and DDPM can generate better quality summaries on this basis by 200 steps of denoising. But, when t equals to 50 or 1000, both the performance significantly degrades. The former shows that the importance scores generated by the unsupervised video summarization model are not accurate enough and it is infeasible to obtain a satisfactory summary using only 50 steps of denoising, while the latter shows that the denoising ability of DDPM is not sufficient to directly generate summaries from completely Gaussian noise.

Qualitative Evaluation

Qualitative Comparison Results Figure 5 demonstrates several examples of generated video summaries by DSNet (Zhu et al. 2020), DR-DSN (Zhou, Qiao, and Xiang 2018), dppLSTM (Zhang et al. 2016), our method and ground-truth. We can observe that in the cases, the sum-

maries generated by our method have more overlaps with the ground-truth summaries, which demonstrates that our method effectively reduces the interference of subjective annotation noise and generates high-quality summaries. In the cross-domain scenarios shown in Figure 5 (b), our method outperforms other methods, which exhibits stronger generalization ability.

User Study We also conduct human evaluation to analyze the effectiveness of our method. Concretely, for all videos in TVSum and SumMe datasets, we show each user the video summaries generated by "w/o unsup" and the video summaries generated by our method. 10 users are asked to choose the video summaries they are more interested in. Figure 6 shows the human evaluation results. For each dataset, the blue sector represents the percentage of summaries generated by "w/o unsup" that are chosen by users and the green sector represents the percentage of summaries generated by our complete method that are chosen by users. We observe that the video summaries generated by our method are more preferred by users for both datasets, which further indicates that our method has advantages over removing the unsupervised video summarization model and using only DDPM.

Conclusion

We have presented a novel diffusion method for video summarization based on Denoising Diffusion Probabilistic Model (DDPM). It learns the probability distribution of training data through noise prediction, and generates summaries by iterative denoising. Our method can resist subjective annotation noise with better robustness and less overfit the training data with stronger generalization, compared to the discriminative method. Extensive experiments on different settings of various datasets demonstrate the effectiveness of our method.

In the future, we are going to apply our method to query-based video summarization, introducing user preferences into the summarization process in the form of text queries, and further exploring the impact of user subjectivity on video summarization.

Acknowledgments

This work was supported in part by the Natural Science Foundation of Shenzhen under Grant No. JCYJ20230807142703006 and the Natural Science Foundation of China (NSFC) under Grant No 62072041.

References

- Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2020a. AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3278–3292.
- Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2020b. Unsupervised video summarization via attention-driven adversarial learning. In *Proceedings of the International Conference on MultiMedia Modeling*, 492–504. Springer.
- Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. *Advances in Neural Information Processing Systems*, 29.
- Carrillo, H.; Clément, M.; Bugeau, A.; and Simo-Serra, E. 2023. Diffusart: Enhancing line art colorization with conditional diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3486–3490.
- De Avila, S. E. F.; Lopes, A. P. B.; da Luz Jr, A.; and de Albuquerque Araújo, A. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1): 56–68.
- Fajtl, J.; Sokeh, H. S.; Argyriou, V.; Monekosso, D.; and Remagnino, P. 2019. Summarizing videos with attention. In *Proceedings of the Asian Conference on Computer Vision*, 39–54. Springer.
- Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10021–10030.
- Gong, B.; Chao, W.-L.; Grauman, K.; and Sha, F. 2014. Diverse sequential subset selection for supervised video summarization. *Advances in Neural Information Processing Systems*, 27.
- He, X.; Hua, Y.; Song, T.; Zhang, Z.; Xue, Z.; Ma, R.; Robertson, N.; and Guan, H. 2019. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2296–2304.
- Ho, H.-I.; Chiu, W.-C.; and Wang, Y.-C. F. 2018. Summarizing first-person videos from third persons’ points of view. In *Proceedings of the European Conference on Computer Vision*, 70–85.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.
- Hsu, T.-C.; Liao, Y.-S.; and Huang, C.-R. 2023. Video summarization with spatiotemporal vision transformer. *IEEE Transactions on Image Processing*, 32: 3013–3026.
- Ji, Z.; Xiong, K.; Pang, Y.; and Li, X. 2019. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6): 1709–1717.
- Khan, H.; Hussain, T.; Khan, S. U.; Khan, Z. A.; and Baik, S. W. 2024. Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, 237: 121288.
- Laganière, R.; Bacco, R.; Hocevar, A.; Lambert, P.; Païs, G.; and Ionescu, B. E. 2008. Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TREC Vid Video Summarization Workshop*, 144–148.
- Li, H.; Ke, Q.; Gong, M.; and Drummond, T. 2023. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5584–5593.
- Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.
- Liang, G.; Lv, Y.; Li, S.; Zhang, S.; and Zhang, Y. 2022. Video summarization with a convolutional attentive adversarial network. *Pattern Recognition*, 131: 108840.
- Liu, T.; Meng, Q.; Huang, J.-J.; Vlontzos, A.; Rueckert, D.; and Kainz, B. 2022. Video summarization through reinforcement learning with a 3D spatio-temporal u-net. *IEEE Transactions on Image Processing*, 31: 1573–1586.
- Mahasseni, B.; Lam, M.; and Todorovic, S. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 202–211.
- Nam, Y.; Lehari, A.; Yang, D.; Bose, D.; Swayamdipta, S.; and Narayanan, S. 2024. Does Video Summarization Require Videos? Quantifying the Effectiveness of Language in Video Summarization. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 8396–8400. IEEE.
- Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, 8162–8171. PMLR.
- Otani, M.; Nakashima, Y.; Rahtu, E.; and Heikkilä, J. 2019. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7596–7604.
- Ren, Z.; Yan, J.; Ni, B.; Liu, B.; Yang, X.; and Zha, H. 2017. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Rochan, M.; Ye, L.; and Wang, Y. 2018. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision*, 347–363.
- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In

- Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. Springer.
- Singh, A.; and Kumar, M. 2024. Bayesian fuzzy clustering and deep CNN-based automatic video summarization. *Multimedia Tools and Applications*, 83(1): 963–1000.
- Song, Y.; Vallmitjana, J.; Stent, A.; and Jaimes, A. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5179–5187.
- Terbouche, H.; Morel, M.; Rodriguez, M.; and Othmani, A. 2023. Multi-annotation attention model for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3143–3152.
- Wang, G.; Wu, X.; and Yan, J. 2024. Progressive reinforcement learning for video summarization. *Information Sciences*, 655: 119888.
- Wu, C. H.; and De la Torre, F. 2023. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7378–7387.
- Yao, T.; Mei, T.; and Rui, Y. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 982–990.
- Yu, Q.; Yu, H.; Sun, Y.; Ding, D.; and Jian, M. 2024. Unsupervised Video Summarization Based on the Diffusion Model of Feature Fusion. *IEEE Transactions on Computational Social Systems*.
- Zhang, K.; Chao, W.-L.; Sha, F.; and Grauman, K. 2016. Video summarization with long short-term memory. In *Proceedings of the European Conference on Computer Vision*, 766–782. Springer.
- Zhang, Y.; Liu, Y.; Kang, W.; and Tao, R. 2023. VSS-Net: visual semantic self-mining network for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhang, Y.; Liu, Y.; and Wu, C. 2024. Attention-guided multi-granularity fusion model for video summarization. *Expert Systems with Applications*, 249: 123568.
- Zhao, B.; Li, H.; Lu, X.; and Li, X. 2021. Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2793–2801.
- Zhao, B.; Li, X.; and Lu, X. 2017. Hierarchical recurrent neural network for video summarization. In *Proceedings of the 25th ACM international conference on Multimedia*, 863–871.
- Zhao, B.; Li, X.; and Lu, X. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7405–7414.
- Zhong, R.; Wang, R.; Yao, W.; Hu, M.; Dong, S.; and Munteanu, A. 2023. Semantic representation and attention alignment for Graph Information Bottleneck in video summarization. *IEEE Transactions on Image Processing*.
- Zhou, K.; Qiao, Y.; and Xiang, T. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Zhu, W.; Han, Y.; Lu, J.; and Zhou, J. 2022. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31: 3017–3031.
- Zhu, W.; Lu, J.; Li, J.; and Zhou, J. 2020. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30: 948–962.