

---

# Diffusion-Inspired Truncated Sampler for Text-Video Retrieval

---

Jiamian Wang<sup>1†</sup>, Pichao Wang<sup>2\*</sup>, Dongfang Liu<sup>1</sup>, Qiang Guan<sup>3</sup>, Sohail Dianat<sup>1</sup>,  
Majid Rabbani<sup>1</sup>, Raghuv eer Rao<sup>4</sup>, Zhiqiang Tao<sup>1†</sup>

<sup>1</sup>Rochester Institute of Technology, <sup>2</sup>Amazon,

<sup>3</sup>Kent State University, <sup>4</sup>DEVCOM Army Research Laboratory

## Abstract

Prevalent text-to-video retrieval methods represent multimodal text-video data in a joint embedding space, aiming at bridging the relevant text-video pairs and pulling away irrelevant ones. One main challenge in state-of-the-art retrieval methods lies in the modality gap, which stems from the substantial disparities between text and video and can persist in the joint space. In this work, we leverage the potential of Diffusion models to address the text-video modality gap by progressively aligning text and video embeddings in a unified space. However, we identify two key limitations of existing Diffusion models in retrieval tasks: The  $\mathcal{L}_2$  loss does not fit the ranking problem inherent in text-video retrieval, and the generation quality heavily depends on the varied initial point drawn from the isotropic Gaussian, causing inaccurate retrieval. To this end, we introduce a new Diffusion-Inspired Truncated Sampler (DITS) that jointly performs progressive alignment and modality gap modeling in the joint embedding space. The key innovation of DITS is to leverage the inherent proximity of text and video embeddings, defining a truncated diffusion flow from the fixed text embedding to the video embedding, enhancing controllability compared to adopting the isotropic Gaussian. Moreover, DITS adopts the contrastive loss to jointly consider the relevant and irrelevant pairs, not only facilitating alignment but also yielding a discriminatively structured embedding. Experiments on five benchmark datasets suggest the state-of-the-art performance of DITS. We empirically find that DITS can also improve the structure of the CLIP embedding space. Code is available at <https://github.com/Jiamian-Wang/DITS-text-video-retrieval>

## 1 Introduction

Text-video retrieval aims to match textual descriptions with relevant video content and vice versa, utilizing both modalities to improve ranking accuracy [Zhu et al., 2023]. This task is challenging due to the multi-modality gap, which represents the inherent differences between textual and visual data representations [Yang et al., 2021, Huang et al., 2023b, Sun et al., 2024, Wang et al., 2021]. Bridging this gap effectively requires advanced methodologies to extract and align semantic information from both modalities into a joint embedding space [Cheng et al., 2023, Xu et al., 2021, Pei et al., 2023].

Current state-of-the-art approaches have explored various strategies to narrow this gap. Advanced feature extraction techniques focus on capturing fine-grained details through temporal modeling [Liu et al., 2022b] and multi-granularity matching strategies, such as frame-level, patch-level, word-level, and token-level representations [Li et al., 2023b, Guan et al., 2023]. Additionally, leveraging large-scale data augmentation has been shown to enhance the capabilities of visual-language models [Wu

---

\*The work does not relate to author’s position at Amazon.

†Corresponding authors: Jiamian Wang (jw4905@rit.edu) & Zhiqiang Tao (zhiqiang.tao@rit.edu)

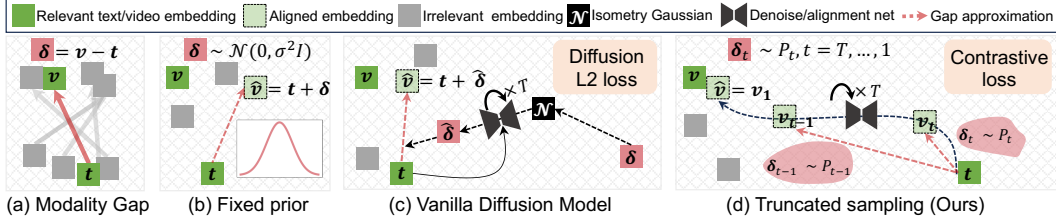


Figure 1: Multi-modality gap alignment for the text-video retrieval. **(a)** Considering the many-to-one correspondence of text-to-video, we define the modality gap stemming from the text ( $\delta = \mathbf{v} - \mathbf{t}$ ) for uniqueness. **(b)** Fixed prior alignment posits a Gaussian distribution, which is inflexible. **(c)** Diffusion-based alignment upon  $\mathcal{L}_2$  loss, which does not fit the retrieval task. The generation is heavily affected by the random samples from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ . **(d)** The proposed diffusion-inspired truncated sampler (DITS) aligns from  $\mathbf{t}$  to  $\mathbf{v}$  and gradually models the gap, guided by the contrastive loss.

et al., 2023, Wang et al., 2024b], while incorporating additional modalities like audio has also improved retrieval performance [Ibrahimi et al., 2023]. Despite the progress, the modality gap (Fig. 1 (a)) remains a persistent issue, significantly impacting the performance of retrieval systems.

A critical observation in text-video retrieval is that existing methods often struggle with the variability and complexity of the semantic alignment between text and video. This work addresses this challenge by investigating Diffusion models [Ho et al., 2019, Song et al., 2021] to bridge the modality gap. Our main contribution is developing a new Diffusion-inspired truncated sampler (DITS) for models to iteratively align video and text embeddings in a progressive manner. Unlike traditional approaches [Qiu et al., 2024, Nukrai et al., 2022] that impose fixed priors for the alignment (Fig. 1 (b)), our method starts from the text embedding and employs a truncated diffusion process to generate aligned video embeddings, addressing the inflexibility and inadequacy of fixed prior methods.

We first explore the generative capability of Diffusion models to mitigate the modality gap. Initially, we employ pre-trained text and video embeddings and model the gap using a diffusion process, where the initial state (isotropic Gaussian) represents a fixed prior, progressively denoised to align the embedding (Fig. 1 (c)). However, we found that the vanilla Diffusion model’s  $\mathcal{L}_2$  loss does not fit the ranking problem inherent in text-video retrieval. The  $\mathcal{L}_2$  loss assumes a Gaussian distribution and minimizes the mean squared error, treating all errors equally and failing to prioritize the alignment of semantically relevant pairs over irrelevant ones. Moreover, the reverse process introduces diversity through initial Gaussian noise sampling, which impacts the generation quality and causes the misalignment. Therefore, it is non-trivial and necessary to tailor the Diffusion modeling from generation (where diversity is beneficial) to ranking (where precise alignment is crucial), catering to text-video retrieval. The focus of diffusion learning should be on closing relevant embeddings while pushing apart irrelevant pairs. This requires a different loss function and sampler for better alignment.

To address the above challenges, DITS leverages the inherent proximity of text and video embeddings in the joint space [Liang et al., 2022, Zhou et al., 2023] and incorporates a truncated diffusion process into the joint space (depicted in Fig. 1 (d)). Different from the vanilla diffusion-based alignment, DITS adopts text embedding as a more meaningful intermediate latent state to initiate sampling, reducing the underlying variability and improving the alignment accuracy over Gaussian noises. The proposed DITS reformulates each diffusion step as modeling the modality gap and gradually controls the text-to-video alignment through the contrastive loss, acting on both aligned and video embeddings, which thus intrinsically promotes the alignment across timestamps in a more fine-grained manner. Empirical evidence shows that DITS can also enhance the structure of the CLIP embedding space. We summarize the contributions of this work as follows.

- This work studies the Diffusion model to bridge the modality gap of text-video retrieval, identifying two key limitations of the vanilla Diffusion model. A new sampler, namely DITS, is also proposed to enhance the multi-modality alignment and benefit the retrieval.
- DITS enables a new truncated diffusion process to conduct the video-text alignment progressively, starting from the intermediate latent state given by text embedding to approaching the relevant video embedding. Governed by the contrastive loss, the proposed DITS gradually models the gap distribution along each timestamp by learning from relevant/irrelevant pairs.

- Extensive experiments on five datasets (MSRVTT, LSMDC, DiDeMo, VATEX, and Charades) suggest that DITS achieves state-of-the-art performance. Empirical evidence shows that DITS also improves the structure of the CLIP embedding space.

## 2 Related Work

**Text-video Retrieval.** State-of-the-art text-video retrieval methods [Huang et al., 2023a, Gao et al., 2021, Yu et al., 2018, Deng et al., 2023, Jin et al., 2023, Croitoru et al., 2021] develop different strategies to facilitate the multi-modality alignment. Some adopt the temporal modeling [Li et al., 2023b, Liu et al., 2023, Li et al., 2023c, Han et al., 2022b] on word sequence or video frames, better extracting the semantic clues and enhancing the retrieval. Others extract multi-modality features at different granularity [Liu et al., 2022b, Chen et al., 2020b, Wang et al., 2023, Ma et al., 2022, Han et al., 2022a, Wang et al., 2021], uncovering hierarchical relationships between word sequences and video clips. Apart from the intricate model designs [Guan et al., 2023, Wu et al., 2021, Bain et al., 2021, Liu et al., 2021, Miech et al., 2021], incorporating the large-scale text-video data via augmentation emerges as a promising approach [Luo et al., 2022, Wu et al., 2023, Falcon et al., 2022], which effectively amplifies the potential of the visual-language foundation models [Chen et al., 2024, Ko et al., 2023] and benefiting the retrieval. Additionally, harnessing new modality data, such as the audio [Ibrahimi et al., 2023, Akbari et al., 2021, Lin et al., 2022, Miech et al., 2018, Liu et al., 2022a, Shvetsova et al., 2022], paves a new direction to bridge the text-video modalities, with abundant semantics clues. Recent advances [Fang et al., 2023, Wang et al., 2024a, Gao et al., 2024, Ji et al., 2023, Chun et al., 2021, Nukrai et al., 2022] study different representation forms of the multi-modality data, among which T-MASS [Wang et al., 2024a] achieves the outstanding performance. Despite the prosperity, the modality gap persists in the joint space. Distinguished from existing methods, this work studies how to learn the Diffusion model to model the modality gap in the joint embedding space, considering the substantial gap between the generation task and the ranking nature inherent to text-video retrieval.

**Diffusion models.** Diffusion models [Croitoru et al., 2023, Ho et al., 2019] define a transition from the isotropic Gaussian to the clean data with a learned diffusion process, emerging as powerful generative models for learning complex distributions. This approach has been extensively applied in image generation tasks [Yang et al., 2023], with models such as Denoising Diffusion Probabilistic Models (DDPM) [Ho et al., 2019], DDIM [Song et al., 2021], and LDM [Rombach et al., 2022] demonstrating impressive capabilities in generating high-fidelity images. Building upon these foundations, recent research [Zhang et al., 2023] has extended Diffusion models to the multimodal domain, not only enabling the generation of diverse and realistic data with distinct type (such as image, audio, etc.) [Yang et al., 2023], but also achieving the transition across varied modalities (*e.g.*, text-to-image, text-to-video, etc.) [Ghosh et al., 2024]. Nevertheless, leveraging the Diffusion model to the ranking tasks such as the text-video retrieval is non-trivial due to the significant difference in data, evaluation metrics, and the inherent challenges (*e.g.*, diversity, coherence, and realism for generation *v.s.* relevance scores in retrieval). To our best knowledge, previous works of DiffusionRet [Jin et al., 2023] and MomentDiff [Li et al., 2023a] combines Diffusion model with retrieval. Albeit the encouraging performance, their method adopts the Diffusion model to learn a mapping from random noise to the signal (*e.g.*, similarity score or real moments) using the diffusion  $L_1$  or  $L_2$  loss, treating the ranking problem as a generation task without addressing the problem studied in this work.

## 3 Method

### 3.1 Preliminaries

**Text-Video Retrieval.** We have a multi-modality dataset consisting of  $N_v$  video clips and  $N_t$  text descriptions, *e.g.*,  $\mathcal{D} = \{v^{(i)}, t^{(j)}\}$ , where  $i = 1, \dots, N_v$ ,  $j = 1, \dots, N_t$ . Each video clip  $v^{(i)}$  may correspond to one or more text samples  $t^{(j)}$ . For simplicity, we let  $v$  and  $t$  to denote arbitrary video and text. The retrieval system adopts a multi-modality model, such as upon CLIP [Radford et al., 2021], to abstract the video and text features into a joint embedding space, *i.e.*,  $\mathbf{v}, \mathbf{t} \in \mathbb{R}^d$ , where  $d$  denotes the embedding dimension of the joint space.

$$\mathbf{v} = F_\theta(v); \quad \mathbf{t} = G_\phi(t), \quad (1)$$

where  $F_\theta(\cdot)$  denotes the advanced visual encoder and  $G_\phi(\cdot)$  denotes the text encoder. Existing works mainly adopt the symmetric-formed cross-entropy [Oord et al., 2018] loss to train the model

$$\mathcal{L}_{\text{sce}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{s(\mathbf{t}^{(i)}, \mathbf{v}^{(i)}) \cdot \tau}}{\sum_j e^{s(\mathbf{t}^{(i)}, \mathbf{v}^{(j)}) \cdot \tau}} + \log \frac{e^{s(\mathbf{t}^{(i)}, \mathbf{v}^{(i)}) \cdot \tau}}{\sum_j e^{s(\mathbf{t}^{(j)}, \mathbf{v}^{(i)}) \cdot \tau}}, \quad (2)$$

where the cosine similarity  $s(\mathbf{t}, \mathbf{v}) = \frac{|\mathbf{t} \cdot \mathbf{v}|}{|\mathbf{t}| \cdot |\mathbf{v}|}$  is employed as a distance measuring function and  $B$  denotes the batch size. A learnable temperature scaling factor  $\tau$  is employed to control the smoothness of the loss. The first term in Eq. (2) guides the text-to-video ( $t \rightarrow v$ ) retrieval and the second term focuses on the video-to-text ( $v \rightarrow t$ ) retrieval. Notably, the scenario of  $\mathcal{L}_{\text{sce}} = 0$  presents a perfect alignment for all video and text embedding in the joint space, where the relevant  $\frac{\mathbf{v}}{|\mathbf{v}|}$  and  $\frac{\mathbf{t}}{|\mathbf{t}|}$  overlaps ( $s(\mathbf{t}, \mathbf{v}) = 1$ ), and the irrelevant text-video are orthogonal, *i.e.*,  $s(\mathbf{t}, \mathbf{v}) = 0$ . While in practice, it is hard to achieve a perfect alignment considering the substantial disparities between the  $t$  and  $v$  data. To this end, a modality gap  $\delta$  inevitably exists between the relevant text and video embedding as shown by Fig. 1 (a), which poses challenges in determining the relevancy. Nevertheless, how to model the modality gap  $\delta$  and perform alignment accordingly can benefit the retrieval but does not seem to have been fully explored. Note that there are two ways to define the  $\delta$  by either stemming from  $\mathbf{t}$  or  $\mathbf{v}$ . Considering the one-to-many correspondence of video-to-text, as exemplified in Fig. 2 ( $\mathbf{v}^{(1)}$  *v.s.*  $\{\mathbf{t}^{(1)}, \mathbf{t}^{(2)}, \mathbf{t}^{(3)}\}$ ), each  $\mathbf{t}$  associates to a unique  $\delta$  but  $\mathbf{v}$  corresponds to multiple  $\delta$  vectors. Accordingly, we consider  $\delta$  stemming from  $\mathbf{t}$  and consistently adopt  $\delta = \mathbf{v} - \mathbf{t}$  in this work.

**Fixed Prior Alignment.** We first posit a fixed prior for the multi-modality gap, similar to existing multi-modality alignment practices [Qiu et al., 2024, Nukrai et al., 2022], and have

$$\delta \sim \mathcal{N}(0, \sigma^2 I), \quad (3)$$

where  $\sigma^2$  determines the scale of the modality gap, *i.e.*, to what extent  $\mathbf{v}$  deviates from the relevant  $\mathbf{t}$ . Given the arbitrary text embedding  $\mathbf{t}$ , we draw  $\delta$  from the fixed prior as shown above and impose it onto the embedding as the alignment, *e.g.*,  $\mathbf{t} + \delta$ . Then this aligned embedding will take effect in the retrieval system as a substitution of  $\mathbf{t}$ . Unfortunately, there is a decrease in retrieval performances for variant priors (see Table 4). To elaborate, the fixed prior is far from enough to capture the delicate distribution of  $\delta$  given the complicated structure of the CLIP embedding space [Paszke et al., 2019]. Besides, the fixed prior is inflexible to adapt to varied text and video inputs, causing misalignment. These limitations naturally prompt us to resort to more advanced technologies for the gap modeling. As known, Diffusion models [Yang et al., 2023, Croitoru et al., 2023] approximate the sophisticated distributions without explicitly presuming a prior and has been proven to excel in bridging multi-modality representations [Zhou et al., 2023]. In this work, we investigate the potential of the Diffusion models for the gap modeling under text-video retrieval.

### 3.2 Proposed Diffusion-based Alignment

We define the target distribution of the Diffusion model upon the modality gap  $\delta$ . Following the previous works [Gorti et al., 2022, Guan et al., 2023], our proposed Diffusion model operates in the embedding space of the CLIP model and defines a Markov chain of the latent variables for  $\delta$ , *i.e.*,  $\delta_1, \dots, \delta_T$ , where the initial state of the Diffusion model remains an isotropic Gaussian, *i.e.*,  $\delta_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , which in this case, can also be treated as a fixed prior for the modality gap  $\delta$ , and  $\delta_0$  corresponds to the ground truth modality gap. It is non-trivial incorporating the proposed Diffusion model into existing retrieval pipeline due to different natures of the generation and retrieval. To achieve this, we reimplement the denoising network, develop a pretraining stage and a reverse process-engaged retrieval pipeline.

**Denoising Network Design.** We first design a denoising network by referring to the prevalent DiT [Peebles and Xie, 2023]. The denoising network, denoted as  $\epsilon_\gamma(\cdot)$ , consists of  $N$  DiT blocks and let  $\gamma$  denote the learnable parameters for the network. As shown in Fig. 2, the denoising network takes three inputs, including the timestamp  $t$ , the latent embedding  $\delta_t \in \mathbb{R}^d$ , and the condition  $\mathbf{c} \in \mathbb{R}^d$ . Since we define the modality gap  $\delta$  stems from  $\mathbf{t}$  as in Section 3.1, we let the text embedding  $\mathbf{t}$  as the condition to guide the generation, *i.e.*,  $\mathbf{c} = \mathbf{t}$ , which substitutes the vanilla classifier-free guidance [Ho and Salimans, 2021] in DiT. Finally, since we still perform the retrieval in the same latent space, there is no need to perform a reshape operation at the end of the denoising network. As shown in Fig. 1 (c), the denoising network  $\epsilon_\gamma(\cdot)$  approximates the noise imposed on  $\delta_t$  during the diffusion process. To facilitate the learning of  $\epsilon_\gamma(\cdot)$ , we introduce a pretraining stage as follows.

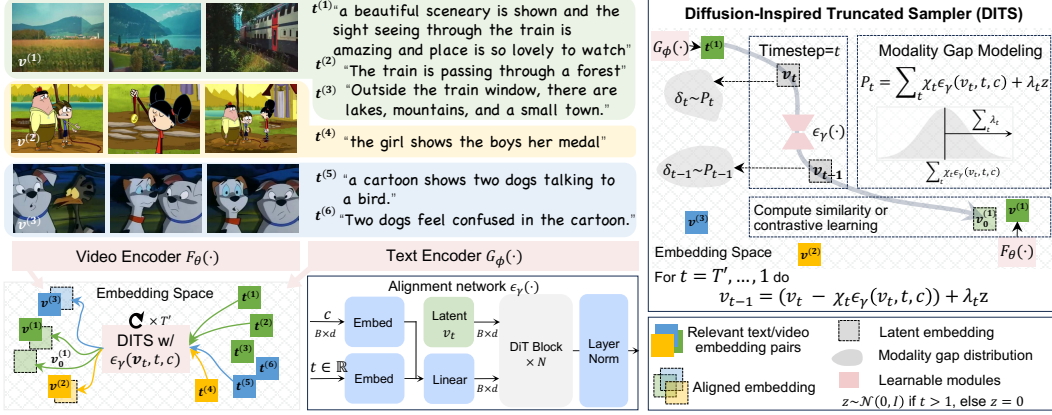


Figure 2: Overview of DITS. Given the video embedding  $\mathbf{v}$  from  $F_\theta(\cdot)$  and the text embedding  $\mathbf{t}$  from  $G_\phi(\cdot)$ , DITS performs a  $T'$ -steps sampling by starting from  $\mathbf{t}$ , to gradually approach  $\mathbf{v}$  (as exemplified by a relevant pair  $\mathbf{t}^{(1)}$ ,  $\mathbf{v}^{(1)}$ ) with the aligned embedding  $\mathbf{v}_0^{(1)}$  at the timestamp  $t = 0$ . Meanwhile the progressive alignment, DITS intrinsically approximates the distribution of the modality gap  $\delta \sim P_t$  at each timestamp  $t = T', \dots, 1$ . The contrastive loss is adopted to guide the alignment and modeling, with parameters:  $\Theta = \{\theta, \phi, \gamma\}$ . We devise  $\epsilon_\gamma(\cdot)$  upon DiT [Peebles and Xie, 2023].

**Diffusion model Pretraining.** The target distribution of  $\delta$  can be highly adaptive to  $\mathbf{v}$  and  $\mathbf{t}$  given by Eq. (1). To stabilize the learning, we adopt the pretrained retrieval model and fix  $F_\theta(\cdot)$  and  $G_\phi(\cdot)$  during the pretraining. The learning objective of  $\epsilon_\gamma(\cdot)$  derives from the diffusion process [Qiu et al., 2024, Nukrai et al., 2022], which gradually adds the Gaussian noise to  $\delta_0$  for  $T$  steps, where  $\delta_1, \dots, \delta_T$  are the latent variables of the same dimension, i.e.,  $\delta_t \in \mathbb{R}^d$

$$q(\delta_{1:T}|\delta_0) := \prod_{t=1}^T q(\delta_t|\delta_{t-1}), \quad q(\delta_t|\delta_{t-1}) := \mathcal{N}(\delta_t; \sqrt{1 - \beta_t}\delta_{t-1}, \beta_t\mathbf{I}), \quad (4)$$

where  $\beta_1, \dots, \beta_T$  can simply be held as constant [Ho et al., 2019] to form a variance schedule. By adopting the reparameterization [Kingma and Welling, 2013] upon Eq. (4),  $\delta_t$  can be analytically represented in a closed form

$$q(\delta_t|\delta_0) = \mathcal{N}(\delta_t; \sqrt{\bar{\alpha}_t}\delta_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad \text{where } \alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t := \prod_{s=1}^t \alpha_s. \quad (5)$$

To approximate  $P_\gamma(\delta_{t-1}|\delta_t)$ , we use the variational lower bound to optimize the negative log-likelihood  $-\log P_\gamma(\delta_0)$ , which can be parameterized as a  $\mathcal{L}_2$  loss [Ho et al., 2019]

$$\mathcal{L}_\gamma = \mathbb{E}_{\delta_0, t, \epsilon} [\|\epsilon - \epsilon_\gamma(\sqrt{\bar{\alpha}_t}\delta_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t, \mathbf{c})\|^2], \quad (6)$$

where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $t \sim \text{Uniform}(1, \dots, T)$  and  $\mathbf{c}$  denotes the condition. Specifically,  $\epsilon$ ,  $t$ , and  $\mathbf{c}$  are three inputs of the denoising network  $\epsilon_\gamma(\cdot)$ . By minimizing the  $\mathcal{L}_\gamma$ , the Diffusion model learns to transition from the isotropic Gaussian to the target distribution of  $\delta_0$  through a reverse process, based on which this method aligns the input text embedding  $\mathbf{t}$  (given as condition) as  $\mathbf{t} + \delta_0$ , to bridge its relevant counterpart  $\mathbf{v}$ . We introduce the reverse process-engaged retrieval pipeline in the next.

**Reverse Process-Engaged Retrieval.** We perform the vanilla DDPM [Ho et al., 2019] sampling process in this work. Starting from the isotropic Gaussian  $\delta_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , the reverse process samples the gap  $\delta_0$  through a  $T$ -step Markov chain

$$\delta_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \delta_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\gamma(\delta, t, \mathbf{c}) \right) + \sigma_t \mathbf{z}, \quad \text{where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (7)$$

where  $t$  decreases from  $T$  to 1,  $\mathbf{z} = \mathbf{0}$  if  $t = 1$ , and  $\sigma_t^2 = \beta_t$ . The reverse process of the Diffusion model will be plugged into the retrieval pipeline as the post-feature extraction alignment, collectively forming a reverse process-engaged retrieval pipeline. To this end, one can substitute  $\mathbf{t}$  with the aligned embedding  $\mathbf{t} + \delta_0$  and perform retrieval with regard to arbitrary  $\mathbf{v}$ . Since the embedding space structure has been edited by the Diffusion model, it might be better to further rearrange the multi-modality embeddings with the contrastive loss. We opt to jointly tune all of the learnable parameters  $\Theta = \{\theta, \gamma, \phi\}$  upon this reverse process-engaged retrieval pipeline with the symmetric cross-entropy given in Eq. (2). We provide more details and discussion in Section 4.1.

### 3.3 Proposed Alignment by DITS

To sum up, this work first implements a feasible approach to leverage the generative power of the Diffusion model in the retrieval pipeline. We firstly pretrain the Diffusion model with Eq. (6) and then incorporate the reverse process in Eq. (7) for the alignment. Notably, we find that despite that the Diffusion model effectively bridging the relevant text and video embedding, either applying the pretrained model for the alignment or adopting the reverse process-engaged retrieval causes obvious performance declines (shown in Table 4). This prompts us to rethink Diffusion models in the context of our problem. Based on the designs in Section 3.2, we develop DITS in the following.

**Rethinking the Vanilla Diffusion model.** We find two-fold limitations of the Diffusion model defective for the task of retrieval. **First**, the  $\mathcal{L}_2$  loss ( $\mathcal{L}_\gamma$ ) does not fit the ranking problem inherent in text-video retrieval. On one hand,  $\mathcal{L}_\gamma$  shown in Eq. (6) minimizes a mean squared error, which inherently lacks an prioritization effect for the relevant pairs over irrelevant ones. On the other, it is not clear what the desired alignment should be for the irrelevant pairs, making it hard to define a target distribution accordingly and implement  $\mathcal{L}_\gamma$  for the modeling. Nevertheless, there is no guarantee that the  $\mathcal{L}_2$  loss (e.g.,  $\mathcal{L}_\gamma$ ) can effectively pull away irrelevant pairs, even though the gap is mitigated for the relevant pairs as shown in Fig. 3. **Second**, the reverse process either fails to benefit the retrieval (Table 4) – the initial starting point of the reverse, *i.e.*, of  $\delta_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  can significantly impact the diversity and the generation quality [Zhou et al., 2023, Chung et al., 2022, Zheng et al., 2023]. This can incur inaccurate alignment and retrieval. How to determine the  $\delta_T$  without affecting the performance remains a challenge.

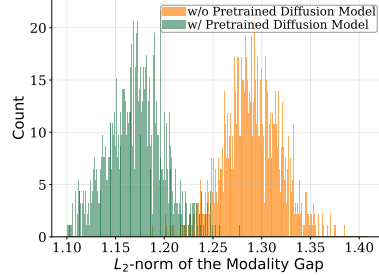


Figure 3: Modality gap distribution for the **relevant** pairs before and after the Diffusion model alignment.

**Diffusion-Inspired Truncated Sampler.** Bearing the above findings, in this work, we introduce the diffusion-inspired truncated sampler (DITS) as an effective alignment tool to benefit the retrieval. Considering the neighboring nature of the CLIP embedding space – relevant  $\mathbf{v}$  and  $\mathbf{t}$  have already been within arm’s length of the distance) [Liang et al., 2022, Nukrai et al., 2022] – there is no need to start sampling from the isotropic Gaussian as in the vanilla Diffusion models. Instead, DITS proposes a progressive alignment procedure that directly starts from the  $\mathbf{t}$  to  $\mathbf{v}$  as shown in Fig. 2 *upper right*, forming truncation. Given  $\mathbf{v}$  by  $F_\theta(\cdot)$  and  $\mathbf{t}$  by  $G_\phi(\cdot)$

$$\mathbf{v}_{t-1} = (\mathbf{v}_t - \chi_t \epsilon_\gamma(\mathbf{v}_t, t, \mathbf{c})) + \lambda_t \mathbf{z}, \text{ For } t = T', \dots, 1, \quad (8)$$

where  $\epsilon_\gamma(\cdot)$  serves as an alignment network and  $T'$  denotes the number of the truncated timestamps. The starting point  $\mathbf{v}_{T'} = \mathbf{t}$  (taking  $\mathbf{t}^{(1)}$  as an example in Fig. 2) and  $\chi_t = \frac{1-\alpha_t}{\sqrt{\alpha_t - \alpha_t \bar{\alpha}_t}}$  denotes a pre-defined schedule. Besides, we let  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$ , else  $\mathbf{z} = \mathbf{0}$ . We set  $\lambda_t = \sigma_t$ . Compared with vanilla sampling process given by Eq. (7), the initial latent of DITS becomes a text embedding point (e.g.,  $\mathbf{t}^{(i)}$ ) rather than a random sample from the isotropic Gaussian, which brings following advantages: (1) DITS gets rid of the effect of starting point random sampling. (2) the query text (such as for  $t \rightarrow v$ ) initializes the alignment in a more straightforward way, rather than solely being a condition to guide the learning. Considering this, the condition  $\mathbf{c}$  is not used in  $\epsilon_\gamma(\cdot)$ . In addition, we adopt the contrastive loss upon Eq. (2) to guide the learning, during which we substitute  $\mathbf{t}$  with the aligned embedding  $\mathbf{v}_0$ . We jointly train all of the learnable parameters  $\Theta = \{\theta, \phi, \gamma\}$  as denoted in red in Fig. 2. By observing and considering both relevant and irrelevant pairs, DITS learns the alignment that potentially benefits the overall retrieval performance.

It turns out DITS intrinsically involves the modality gap modeling along the timestamps, (*i.e.*,  $\delta \sim P_t$ ) as shown in Fig. 2. One can derive the distribution of the modality gap  $\delta$  from Eq. (8), which can be represented by using the reparameterization

$$P_t = \sum_t \chi_t \epsilon_\gamma(\mathbf{v}_t, t, \mathbf{c}) + \lambda_t \mathbf{z}, \quad (9)$$

where  $\epsilon_\gamma(\mathbf{v}_t, t, \mathbf{c})$  can be regarded as a bias of the modality gap and  $\lambda_t$  controls the scale of the gap modeling at the timestamp  $t$ . Interestingly, DITS yields a remarkable performance boost over the baseline and achieves the state-of-the-art performance (Section 4.2). We also empirically find that DITS also serves as an effective tool to help improve the CLIP embedding space. See more details in Section 4.4.

Table 1: Text-to-video retrieval performance on MSRVTT [Xu et al., 2016] and LSMDC [Rohrbach et al., 2015]. Bold denotes the best performance. “-” denotes that the result is unavailable.

Methods	MSRVTT Retrieval					LSMDC Retrieval				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
<b>CLIP-ViT-B/32</b>										
X-Pool [Gorti et al., 2022]	46.9	72.8	82.2	2.0	14.3	25.2	43.7	53.5	8.0	53.2
STAN [Liu et al., 2023]	46.9	72.8	82.8	2.0	-	23.7	42.7	51.8	9.0	-
ProST [Li et al., 2023b]	48.2	74.6	83.4	2.0	12.4	24.1	42.5	51.6	9.0	54.6
DiffusionRet [Jin et al., 2023]	49.0	75.2	82.7	2.0	12.1	24.4	43.1	54.3	8.0	<b>40.7</b>
UATVR [Fang et al., 2023]	47.5	73.9	83.5	2.0	12.3	-	-	-	-	-
UCOFIA [Wang et al., 2023]	49.4	72.1	-	-	12.9	-	-	-	-	-
TEFAL [Ibrahimi et al., 2023]	49.4	<b>75.9</b>	83.9	2.0	12.0	26.8	46.1	56.5	7.0	44.4
CLIP-ViP [Xue et al., 2023]	50.1	74.8	84.6	1.0	-	25.6	45.3	54.4	8.0	-
T-MASS [Wang et al., 2024a]	50.2	75.3	<b>85.1</b>	1.0	11.9	<b>28.9</b>	<b>48.2</b>	<b>57.6</b>	6.0	43.3
DITS (Ours)	<b>51.9</b>	75.7	84.6	<b>1.0</b>	<b>11.6</b>	28.2	47.3	56.6	<b>6.0</b>	43.7
<b>CLIP-ViT-B/16</b>										
X-Pool [Gorti et al., 2022]	48.2	73.7	82.6	2.0	12.7	26.1	46.8	56.7	7.0	47.3
UATVR [Fang et al., 2023]	50.8	76.3	85.5	1.0	12.4	-	-	-	-	-
CLIP-ViP [Xue et al., 2023]	54.2	77.2	84.8	1.0	-	29.4	50.6	59.0	5.0	-
T-MASS [Wang et al., 2024a]	52.7	77.1	85.6	1.0	10.5	30.3	52.2	<b>61.3</b>	5.0	40.1
DITS (Ours)	<b>55.0</b>	<b>79.8</b>	<b>87.1</b>	<b>1.0</b>	<b>10.0</b>	<b>31.0</b>	<b>52.4</b>	61.0	<b>5.0</b>	<b>38.4</b>

Table 2: Text-to-video retrieval performance on DiDeMo [Anne Hendricks et al., 2017] and VATEX [Wang et al., 2019]. Bold denotes the best performance. “-” denotes that the result is unavailable.

Methods	DiDeMo Retrieval					VATEX Retrieval				
	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	MnR ↓
<b>CLIP-ViT-B/32</b>										
X-Pool [Gorti et al., 2022]	44.6	73.2	82.0	2.0	15.4	60.0	90.0	95.0	1.0	3.8
ProST [Li et al., 2023b]	44.9	72.7	82.7	2.0	13.7	60.6	90.5	95.4	1.0	3.4
STAN [Liu et al., 2023]	46.5	71.5	80.9	2.0	-	-	-	-	-	-
DiffusionRet [Jin et al., 2023]	46.7	74.7	82.7	2.0	14.3	-	-	-	-	-
UATVR [Fang et al., 2023]	43.1	71.8	82.3	2.0	15.1	61.3	91.0	95.6	1.0	3.3
UCOFIA [Wang et al., 2023]	46.5	74.8	-	-	13.4	61.1	90.5	-	-	3.4
CLIP-ViP [Xue et al., 2023]	48.6	77.1	84.4	2.0	-	-	-	-	-	-
T-MASS [Wang et al., 2024a]	50.9	77.2	85.3	1.0	12.1	63.0	92.3	96.4	1.0	3.2
DITS (Ours)	<b>51.1</b>	<b>77.9</b>	<b>85.8</b>	<b>1.0</b>	<b>12.1</b>	<b>64.1</b>	<b>92.7</b>	<b>97.0</b>	<b>1.0</b>	<b>2.9</b>
<b>CLIP-ViT-B/16</b>										
X-Pool [Gorti et al., 2022]	47.3	74.8	82.8	2.0	14.2	62.6	91.7	96.0	1.0	3.4
UATVR [Fang et al., 2023]	45.8	73.7	83.3	2.0	13.5	64.5	92.6	96.8	1.0	2.8
CLIP-ViP [Xue et al., 2023]	50.5	78.4	87.1	1.0	-	-	-	-	-	-
T-MASS [Wang et al., 2024a]	53.3	80.1	<b>87.7</b>	1.0	<b>9.8</b>	65.6	93.9	97.2	1.0	2.7
DITS (Ours)	<b>55.8</b>	<b>80.5</b>	87.5	<b>1.0</b>	11.0	<b>66.4</b>	<b>94.3</b>	<b>97.5</b>	<b>1.0</b>	<b>2.7</b>

## 4 Experiment

### 4.1 Experimental Settings

**Datasets.** We employ five benchmark datasets for evaluation. Firstly, we utilize MSRVTT [Xu et al., 2016], comprising 10,000 YouTube video clips (each having 20 captions) and follow the 1K-A testing split in Liu et al. [2019]. Secondly, LSMDC [Rohrbach et al., 2015] includes 118,081 text-video pairs, providing videos with longer duration. The testing set contains 1000 videos, as per Gabeur et al. [2020], Gorti et al. [2022]. Thirdly, DiDeMo [Anne Hendricks et al., 2017] contains  $\sim 40,000$  captions and  $\sim 10,000$  video clips. We adhere to the data splits detailed in Luo et al. [2022], Jin et al. [2023]. Fourthly, VATEX [Wang et al., 2019] comprises 41,250 video clips, where each is paired with ten English and ten Chinese descriptions. We follow the split in Chen et al. [2020a]. Lastly, Charades [Sigurdsson et al., 2016] contains 9848 video clips, each with multiple text descriptions detailing daily activities and actions. We adopt the split protocol of Lin et al. [2022].

**Implementation Details.** Following previous works [Gorti et al., 2022, Li et al., 2023b], we resize the video to be the spatial size of  $224 \times 224$  and uniformly sample 12 frames from the video for all datasets. For the retrieval model, we adopt X-Pool [Gorti et al., 2022] as the baseline, where both CLIP-ViT/B-32 and CLIP-ViT/B-16 are employed as the feature extractor. The dropout is set to 0.3. Different from DiT [Peebles and Xie, 2023], we set our denoising network (also used as the alignment network in DITS) with  $N = 4$  blocks, with 16 heads and an MLP ratio of 4.0. We let the dimension  $d = 512$  for the whole model. We find that a timestamp of  $T = 10$  is enough for diffusion-based alignment. For DITS, we set the truncated timestamp  $T' = 5$  for DiDeMo and  $T' = 10$  for others. A linear variance schedule with  $\beta = 0.1$  and  $\beta = 0.99$  is adopted. All of the parameters  $\Theta = \{\theta, \phi, \gamma\}$  are trained with an AdamW [Loshchilov and Hutter, 2017] optimizer with weight decay of 0.2 and

Table 3: Text-to-video comparisons on Charades [Sigurdsson et al., 2016]. Bold denotes the best.

Methods	R@1	R@5	R@10	MdR	MnR
<i>CLIP-ViT-B/32</i>					
ClipBERT [Lei et al., 2021]	6.7	17.3	25.2	32.0	149.7
CLIP4Clip [Luo et al., 2022]	9.9	27.1	36.8	21.0	85.4
X-Pool [Gorti et al., 2022]	11.2	28.3	38.8	20.0	82.7
T-MASS [Wang et al., 2024a]	14.2	36.2	<b>48.3</b>	12.0	<b>54.8</b>
DITS (Ours)	<b>15.5</b>	<b>36.5</b>	47.8	<b>12.0</b>	55.0
<i>CLIP-ViT-B/16</i>					
CLIP4Clip [Luo et al., 2022]	16.0	38.2	48.5	12.0	54.1
X-Pool [Gorti et al., 2022]	20.7	42.5	53.5	9.0	47.4
T-MASS [Wang et al., 2024a]	26.7	51.7	63.9	5.0	30.0
DITS (Ours)	<b>27.4</b>	<b>52.0</b>	<b>64.3</b>	<b>5.0</b>	<b>29.7</b>

Table 4: Ablation study of the proposed DITS on MSRVT-1k. We adopt X-Pool [Gorti et al., 2022] as the ‘‘Baseline’’. ‘‘Diffusion’’ denotes the Diffusion-based alignment method in Section 3.2.

Methods	Alignment	$\mathcal{L}_2$	$\mathcal{L}_{\text{sce}}$	Truncation	R@1	R@5	R@10	MdR	MnR	
Baseline	$\times$	$\times$	$\checkmark$	$\times$	46.9	72.8	82.2	2.0	14.3	
Fixed Prior	$\sigma^2 = 1.0$	$\checkmark$	$\times$	$\checkmark$	$\times$	28.0	61.9	74.7	3.0	19.7
	$\sigma^2 = 0.1$	$\checkmark$	$\times$	$\checkmark$	$\times$	45.1	73.3	82.6	2.0	13.8
Diffusion	Pretrain	$\checkmark$	$\checkmark$	$\times$	$\times$	35.4	66.7	78.3	3.0	14.4
	Fine tune	$\checkmark$	$\times$	$\checkmark$	$\times$	46.5	73.9	83.0	2.0	13.3
DITS	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	<b>51.9</b>	<b>75.7</b>	<b>84.6</b>	<b>1.0</b>	<b>11.6</b>	

warmup rate of 0.1. We set the training epochs to 5 for all datasets and adopt the same seed of 24. We perform contrastive learning with a batch size of  $B = 32$  for all datasets and backbones. Same as X-Pool [Gorti et al., 2022], the learning rate of the CLIP model is initialized as  $1 \times 10^{-5}$ . The learning rate for non-CLIP modules is  $3 \times 10^{-5}$  for MSRVT [Xu et al., 2016] and  $1 \times 10^{-5}$  for all the other datasets. For a fair comparison, no post-processing techniques [Bogolin et al., 2022, Cheng et al., 2021] are employed in this work. We implement DITS with PyTorch [Paszke et al., 2019] and perform experiments on an NVIDIA A100 GPU.

**Compared Methods.** We compare DITS with X-Pool [Gorti et al., 2022], STAN [Liu et al., 2023], ProST [Li et al., 2023b], DiffusionRet [Jin et al., 2023], UATVR [Fang et al., 2023], UCOFIA [Wang et al., 2023], TEFAL [Ibrahimi et al., 2023], CLIP-ViP [Xue et al., 2023] and T-MASS [Wang et al., 2024a]. Among them, TEFAL adopts additional audio data to train the model. CLIP-ViP further adopts WebVid-2.5M [Bain et al., 2021] and HD-VILA-100M [Xue et al., 2022] as the augmentation. Besides, some methods take larger batch sizes for the training (e.g., 64 in DiffusionRet [Jin et al., 2023] and UATVR, 128 for ProST, STAN, and UCOFIA except DiDeMo). We perform evaluation using metrics such as Recall at ranks  $\{1, 5, 10\}$ , Median Rank (MdR), and Mean Rank (MnR).

## 4.2 Performance Comparison

We compare the text-to-video retrieval performance of different methods in Table 1, 2, and 3. The proposed DITS can achieve the best performance on different CLIP backbones, datasets, and metrics. For example, on MSRVT, DITS outperforms the baseline X-Pool by 5% at R@1 and outperforms T-MASS by 1.7%/2.3% at R@1 on different backbones. On DiDeMo, DITS outperforms CLIP-ViT by 2.2% with CLIP-ViT-B/32 and 5.3% with CLIP-ViT-B/16, achieving a consistent boost on different metrics. Moreover, on the challenging dataset of Charades, DITS also brings a remarkable performance boost over X-Pool on both backbones, e.g., 4.3% and 6.7% at R@1, respectively. There is one scenario that T-MASS enables better retrieval than DITS on LSMDC with CLIP-ViT-B/32. However, our method gains an advantage with a larger CLIP backbone on all datasets, indicating better scalability on the retrieval model. We also note that DITS excels in retrieval top-ranked results, such as on MSRVT, DiDeMo – a progressive alignment and gap modeling facilitates a narrower retrieval scope, encouraging retrieval precision. We provide in-depth analysis of this tendency in Section 4.4. Overall, extensive results on benchmark datasets demonstrate that DITS effectively aligns text-video embedding in the joint space, delivering promising retrieval results.

## 4.3 Ablation Study

Table 4 provides an ablation study of the proposed method. Specifically, ‘‘Baseline’’ denotes the baseline method of X-Pool [Gorti et al., 2022], based on which we provide three types of the



Table 5: Discussion on DITS. Highlighted settings are adopted for the benchmark comparison.

$T'$	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MdR $\downarrow$	MnR $\downarrow$
0	46.9	72.8	82.2	2.0	14.3
1	49.2	75.0	<b>85.1</b>	2.0	<b>11.3</b>
10	<b>51.9</b>	<b>75.7</b>	84.6	<b>1.0</b>	11.6
20	47.0	73.9	83.7	2.0	12.2
30	32.5	63.5	76.0	3.0	16.2
40	0.2	0.5	1.0	493.5	497.9

Methods	R@1	R@5	R@10	MdR	MnR
Baseline w/o DITS	46.9	72.8	82.2	2.0	14.3
DITS (Fix CLIP)	39.2	65.4	77.1	2.0	18.0
DITS (Joint train)	<b>51.9</b>	<b>75.7</b>	<b>84.6</b>	<b>1.0</b>	<b>11.6</b>

alignment methods, corresponding to Fig. 1. First, we show the naive modeling methods using two different fixed priors, such as  $\sigma^2 = 0.1, 1.0$ . Both of them bring performance descent. Fixed priors are inflexible, hardly benefiting the retrieval as expected. We notice that the sensitivity of the performance toward the prior. This also prompts us to leverage advanced methods of Diffusion model below. We provide the retrieval performance corresponding to both training stages of pretraining and fine-tuning as in Section 3.2. In pretraining, the reverse process is adopted for the generation. As shown, the pretraining with  $\mathcal{L}_2$  loss undermines the retrieval performance (e.g., 35.4% v.s. 46.9%). Despite the fact that  $\mathcal{L}_2$  loss effectively attracts relevant pairs (Fig. 3), it fails to align the irrelevant ones, leading to the performance descent. By comparison, when we fine-tune the model with the contrastive loss, we observe a remarkable boost of 11.1% at R@1 over pretraining and better results than baseline of X-Pool on metrics except R@1. This indicates that such a method, although enables a good retrieval scope (e.g., being favorable for R@5/10), but fails to perform accurate alignment and retrieval, underlying which, one main flaw lies in the random sampling in isotropic Gaussian. Drawing inspiration, the proposed DITS in the bottom line of Table 4 proposes a truncation process to alleviate the above concern, and adopts contrastive learning to guide the alignment. Encouragingly, we find an evident boost of implementing such a pipeline by following the variance schedule of the vanilla diffusion flow (e.g.,  $\alpha_t, \beta_t$ ). This prompt us to study the behavior of DITS in the following.

#### 4.4 Discussion on DITS

**Discussion on Truncated timestamps.** The number of the timestamps  $T'$  plays an important role for an accurate retrieval. We provide the performance of DITS at different timestamps in Table 5a. The  $T' = 0$  denotes the baseline of X-Pool. As shown, the performance of DITS first increases and then drops when we gradually enlarging  $T'$  from 1 to 40. DITS achieves the best performance at  $T' = 10$ . Specifically, one-step operation ( $T' = 1$ ) enables determining a rough retrieval scope, benefiting Recall over a larger range, e.g., R@10. Based on this, DITS requires more steps to narrow the scope for more accurate alignment and retrieval (e.g.,  $T' = 10$ ). However, continually enlarging  $T'$  enforces DITS to learn an extremely precise result, sacrificing the flexibility to accommodate the testing data and in the end, resulting the loss of the generalization ability.

**Improving the CLIP Embedding Space.** We also study the effect of DITS to the CLIP in Table 5b. There is a large performance gap when we fix CLIP or perform joint train for DITS. This indicates that solely performing the alignment in the original CLIP space is not enough. To this end, DITS not only plays the role of text-video embedding alignment, but also serves as a tool to guide the learning of CLIP, thus improving the CLIP space. As shown in Fig. 4, the distribution of the modality gap<sup>3</sup> between two spaces deviates from each other, characterizing the aligning effect of DITS.

**Diffusion Model Conditions.** We perform different conditions of the  $\epsilon_\gamma(\cdot)$  in DITS, including with text condition  $\mathbf{t}$ , with video condition  $\mathbf{v}$ , and with both conditions  $\mathbf{t}$  and  $\mathbf{v}$ , as well as without any conditions (see discussions in Section 3.3). As shown in Table 6a, we find that there is no need to introduce the conditions onto the network. The diffusion process for this task emphasizes “accurate alignment”, unlike the general Diffusion models highlighting diversity in creative works. The previous intuition of the diffusion condition may not be applicable in this work: (1) Due to the one-to-many

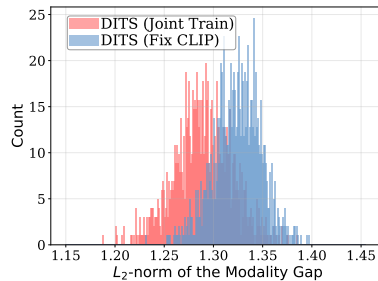


Figure 4: DITS reduces the modality gap for the **relevant** pairs, serving as a tool to improve the space by guiding the CLIP’s learning.

<sup>3</sup>Measured by  $L_2$ -norm (non-bounded), corresponding to Euclidean distance. We also provide mean absolute distance version (using  $L_1$ -norm of the same data in Fig. 6).

Table 6: Discussion on DITS. Highlighted settings are adopted for the benchmark comparison.

(a) Different model conditions for DITS on MSRVT. (b) Different types of the modality gap  $\delta$  on MSRVT.

Condition	R@1	R@5	R@10	MdR	MnR	$\delta$	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MdR $\downarrow$	MnR $\downarrow$
w/o alignment	46.9	72.8	82.2	2.0	14.3	w/o alignment	46.9	72.8	82.2	2.0	14.3
$\mathbf{t}$	42.3	70.6	80.9	2.0	14.3	$\delta = \mathbf{t} - \mathbf{v}$	39.2	65.4	77.1	2.0	18.0
$\mathbf{v}$	23.1	51.9	65.9	5.0	24.8	$\delta = \mathbf{v} - \mathbf{t}$ (Ours)	<b>50.1</b>	<b>75.5</b>	<b>84.9</b>	<b>1.0</b>	<b>12.2</b>
$\mathbf{t}, \mathbf{v}$	40.9	71.3	80.8	2.0	15.0						
w/o condition	<b>50.1</b>	<b>75.5</b>	<b>84.9</b>	<b>1.0</b>	<b>12.2</b>						

mapping of video-to-text, one video can map to multiple gap vectors. Taking the video as a condition can guide the model learn the undesired modality gap, empirically dropping the performance. (2) DITS start from the text, ensuring the alignment starts from a semantically meaningful initial point. The text condition acts as the constraint or guideline for the learning. Simultaneously applying the text embedding as the starting point and the condition can cause conflicting instructions, empirically decreasing the performance. (3) Joint usage of text and video conditions can inherit both limitations, empirically leading to an unsatisfactory performance.

**Modality Gap Stemming from video embedding.** We also provide experiments when let  $\delta = \mathbf{t} - \mathbf{v}$  in Table 6b. As shown, such a setting brings sub-optimal performance. Imposing  $\delta = \mathbf{t} - \mathbf{v}$  to the  $\mathbf{v}$  works like approximating  $\mathbf{t}$  for the retrieval, conducting the “text-to-text” retrieval, considering less semantic clues within the text, it might be hard to retrieve accurately. By comparison, the proposed method DITS implements  $\delta = \mathbf{v} - \mathbf{t}$  to the  $\mathbf{t}$  works like approximating  $\mathbf{v}$  for the retrieval, performing the “video-to-video” retrieval. Since video contains much abundant clues, it might be easier to get better performance. Interesting, we notice that the previous work of Cap4video [Wu et al., 2023] also demonstrates similar tendencies in Table 5, denoted as Query-Caption/Video Only case.

## 5 Conclusion

This work studied the task of text-video retrieval by proposing a Diffusion-Inspired Truncated Sampler (DITS) for the multi-modality alignment. The primary contribution of this work was to offer insights on tailoring the Diffusion model for the ranking task inherent in retrieval. We uncovered two-fold limitations of the vanilla Diffusion models: the vanilla  $\mathcal{L}_2$  loss was inadequate for alignment by overlooking the irrelevant pairs, the the random initial sampling in isotropic Gaussian introduces variability, causing misalignment. We proposed DITS, which leveraged the inherent proximity of text and video embedding and directly started sampling from text to alleviate the sampling variability, and proposed to adopt the contrastive learning not only to guide the iterative alignment steps over time, but also to facilitate gap modeling. Extensive experiments demonstrated the state-of-the-art performance of DITS. We found DITS could encourage an improved embedding space by guiding the CLIP’s learning. We hope DITS can inspire future explorations in studying Diffusion models in the task of text-video retrieval.

## 6 Acknowledgment

This research work was supported by the DEVCOM Army Research Laboratory under contract W911QX-21-D-0001.

## References

- Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 3
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 7
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 3, 8
- Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *CVPR*, 2022. 8
- Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020a. 7
- Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020b. 3
- Sihan Chen, Xingjian He, Handong Li, Xiaojie Jin, Jiashi Feng, and Jing Liu. Cosa: Concatenated sample pretrained vision-language foundation model. In *ICLR*, 2024. 3
- Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *CVPR*, 2023. 1
- Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv*, 2021. 8
- Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 3
- Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *CVPR*, 2022. 6
- 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*, 2023. 3, 4
- Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtex: Crossmodal generalized distillation for text-video retrieval. In *CVPR*, 2021. 3
- Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, and Qi Wu. Prompt switch: Efficient clip adaptation for text-video retrieval. In *ICCV*, 2023. 3
- Alex Falcon, Giuseppe Serra, and Oswald Lanz. A feature-space multimodal data augmentation technique for text-video retrieval. In *ACM MM*, 2022. 3
- Bo Fang, Chang Liu, Yu Zhou, Min Yang, Yuxin Song, Fu Li, Weiping Wang, Xiangyang Ji, Wanli Ouyang, et al. Uatvr: Uncertainty-adaptive text-video retrieval. In *ICCV*, 2023. 3, 7, 8
- Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 7
- Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. CLIP2TV: an empirical study on transformer-based methods for video-text retrieval. *CoRR*, 2021. 3
- Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *CVPR*, 2024. 3
- Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2024. 3

- Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*, 2022. 4, 7, 8, 17
- Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiayi Gu, Hang Xu, Songcen Xu, Youliang Yan, and Edmund Y Lam. Pidro: Parallel isomeric attention with dynamic routing for text-video retrieval. In *ICCV*, 2023. 1, 3, 4
- Ning Han, Jingjing Chen, Hao Zhang, Huanwen Wang, and Hao Chen. Adversarial multi-grained embedding network for cross-modal text-video retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2022a. 3
- Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022b. 3
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 4
- Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2019. 2, 3, 5
- J. Huang, Y. Li, J. Feng, X. Wu, X. Sun, and R. Ji. Clover: Towards a unified video-language alignment and fusion model. In *CVPR*, 2023a. 3
- Jingjia Huang, Yanan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model. In *CVPR*, 2023b. 1
- Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *ICCV*, 2023. 2, 3, 7, 8
- Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaying Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *CVPR*, 2023. 3
- Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *ICCV*, 2023. 3, 7, 8, 17
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5
- Dohwan Ko, Joonmyung Choi, Hyeong Kyu Choi, Kyoung-Woon On, Byungseok Roh, and Hyunwoo J Kim. Meltr: Meta loss transformer for learning to fine-tune video foundation models. In *CVPR*, 2023. 3
- Jie Lei, Linjie Li, Luwei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 8
- Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. In *NeurIPS*, 2023a. 3
- Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *ICCV*, 2023b. 1, 3, 7, 8
- Yi Li, Kyle Min, Subarna Tripathi, and Nuno Vasconcelos. Svitt: Temporal learning of sparse video-text transformers. In *CVPR*, 2023c. 3
- Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 2, 6
- Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, 2022. 3, 7

- Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, and Thomas H Li. Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In *ICCV*, 2023. 3, 7, 8
- Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 7
- Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *CVPR*, 2021. 3
- Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. Animating images to transfer clip for video-text retrieval. In *ACM SIGIR*, 2022a. 3
- Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, 2022b. 1, 3, 17
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7
- Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. In *Neurocomputing*, 2022. 3, 7, 8
- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, 2022. 3
- Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv*, 2018. 3
- Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, 2021. 3
- David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. In *EMNLP*, 2022. 2, 3, 4, 5, 6
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 4, 8
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4, 5, 7
- Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, 2023. 1
- Longtian Qiu, Shan Ning, and Xuming He. Mining fine-grained image-text alignment for zero-shot captioning via text-only training. In *AAAI*, 2024. 2, 4, 5
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
- Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 7
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *CVPR*, 2022. 3

- Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 7, 8
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. 2021. 2, 3
- Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. Sq-llava: Self-questioning for large vision-language assistant. In *ECCV*, 2024. 1
- Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *CVPR*, 2024a. 3, 7, 8, 17
- Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 1, 3
- Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019. 7
- Yimu Wang, Shuai Yuan, Xiangru Jian, Wei Pang, Mushi Wang, and Ning Yu. Havtr: Improving video-text retrieval through augmentation using large foundation models. *arXiv preprint arXiv:2404.05083*, 2024b. 2
- Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *ICCV*, 2023. 3, 7, 8
- Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. Hanet: Hierarchical alignment networks for video-text retrieval. In *ACM MM*, 2021. 3
- Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In *CVPR*, 2023. 1, 3, 10
- Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 1
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 7, 8
- Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 8
- Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In *ICLR*, 2023. 7, 8
- Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, 2021. 1
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023. 3, 4
- Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 3
- Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 3
- Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. In *ICLR*, 2023. 6
- Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. In *CVPR*, 2023. 2, 4, 6
- Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3, 2023. 1

## A Appendix / supplemental material

We provide in-depth discussions, more results of the proposed DITS as follows

- Limitations discussion. (Section A.1).
- Broader impacts on the proposed method. (Section A.2).
- More discussions and results on Diffusion-based alignment (Section A.3)
- More discussions on computational cost and efficiency. (Section A.4).

### A.1 Limitations

Despite the advancements brought by the proposed Diffusion-Inspired Truncated Sampler (DITS), there might be some limitations can be considered for the future research.

DITS, like other multi-modality alignment methods using contrastive learning, requires redundant amounts of data to train effectively. In scenarios where labeled text-video pairs are scarce, the performance of DITS remains unexplored. This raises a question on studying DITS on challenging text-video retrieval scenarios or zero-shot, few-shot retrieval tasks.

In addition, the robustness of DITS to noisy or incomplete data may be a potential concern. In real-world scenarios, text descriptions and video content can be noisy, incomplete, or of poor quality. While DITS aims to bridge the modality gap, its performance may still degrade in the presence of such imperfect data. This highlights the need for further research into deploying DITS to real-world retrieval systems with data imperfections.

### A.2 Broader Impacts

The proposed Diffusion-Inspired Truncated Sampler (DITS) presents a new alignment solution in the realm of text-video retrieval. By effectively modeling and aligning the modality gap, DITS elevates the performance of retrieval systems and may contribute to the broader landscape of multi-modality learning and the diffusion models.

In essence, DITS offers a novel methodology for mitigating the disparities between textual and visual data, potentially inspiring diverse applications, for example, image captioning models that can accurately describe the contents of images, image question answering models that requires precise semantics understanding capabilities of the model toward the visual information.

Beyond its impact on the multi-modality learning, DITS presents a new diffusion flow that is guided by the contrastive learning and considers nature of the ranking task. In light of this, DITS can pave the way for future innovations in the diffusion model design for learning the relationships of the multi-modality data. Besides, traditional diffusion models often struggle with maintaining alignment accuracy due to their reliance on fixed priors and isotropic Gaussian initial states. DITS introduces a truncated diffusion process that shifts from a Gaussian noise start to a meaningful intermediate state, showcasing a new way to handle complex alignment tasks in diffusion models.

Despite the encouraging performance of the proposed method. There exists a possibility that the proposed method can be utilized in the real-world applications with the illegitimate purpose.

Table 7: Discussion on  $\mathcal{L}_1$  Loss. We compare the retrieval performance with the pretrained Diffusion model as an alignment. We perform experiments on MSRVT. All the other settings are kept the same for a fair comparison.

Methods	R@1	R@5	R@10	MdR	MnR
Baseline w/o alignment	46.9	72.8	82.2	2.0	14.3
Pretrained Diffusion w/ $\mathcal{L}_1$	30.9	59.5	68.4	4.0	40.6
Pretrained Diffusion w/ $\mathcal{L}_2$	<b>35.4</b>	<b>66.7</b>	<b>78.3</b>	<b>3.0</b>	<b>14.4</b>

### A.3 More Details on the Proposed Method

**Discussion on  $L_1$  Loss.** We perform experiments on different diffusion loss of  $\mathcal{L}_1$  and  $\mathcal{L}_2$  losses for the diffusion-based alignment pretraining. As shown in Table 7, the  $\mathcal{L}_2$  provides a better performance.

Table 8: Discussion on the timestamps ( $T'$ ) on MSRVTT.

$T'$	R@1 $\uparrow$	R@5 $\uparrow$	R@10 $\uparrow$	MdR $\downarrow$	MnR $\downarrow$
0	46.9	72.8	82.2	2.0	14.3
1	49.2	75.0	85.1	2.0	11.3
5	49.6	75.1	84.9	2.0	<b>11.2</b>
10	<b>51.9</b>	75.7	84.6	<b>1.0</b>	11.6
15	51.1	<b>76.1</b>	<b>85.8</b>	1.0	11.8
20	47.0	73.9	83.7	2.0	12.2
30	32.5	63.5	76.0	3.0	16.2
40	0.2	0.5	1.0	493.5	497.9

Table 9: Modality gap (measured by  $L_1$ -norm), similarity, and performance change discussion.

Methods	Averaged Modality Gap ( $\downarrow$ )	Averaged Similarity ( $\uparrow$ )	R@1 ( $\uparrow$ )
DITS fix CLIP	23.76	0.122	39.2
DITS (Ours)	18.13	0.168	51.9

$\mathcal{L}_2$  loss penalizes larger errors more heavily than smaller ones because the error term is squared. This can be beneficial in tasks where it is crucial to reduce large deviations, as it forces the model to pay more attention to outliers or significant errors. By comparison,  $\mathcal{L}_1$  loss treats all errors equally, which might not be ideal if large errors need to be minimized more aggressively. In the context of our problem,  $\mathcal{L}_2$  ensures that these outliers are brought into alignment, reducing their impact on the overall model performance, which might be a reason for this observation.

**More Discussions on Timestamp.** In Table 8, we provide a more thorough comparison of different timestamps. As we gradually enlarge the total number timestamp  $T'$  from 1 to 40, all metrics will first get better, and then decrease. This is because DITS performs a progressive alignment. It gradually narrows down the retrieval scope when we initially increase  $T'$ , enabling generally improved performances (R@1, R@5, MdR). However, the model loses the generalization ability if we continually enlarge  $T'$ , causing a performance degradation (e.g., from 15 to 40). We also notice there are some inconsistent tendencies, such as R@10 decreases from  $T' = 1$  to  $T' = 15$ , contrary to the tendency of R@1 and R@5. This is because different recall precision (e.g., 1,5,10) can be regarded as different retrieval tasks with different focus. When the model continually narrows down the retrieval scope for the given query, it can easily affect R@10 that requires a larger scope. To balance between different Recall precision, we choose  $T' = 10$  for MSRVTT. This experiment shows the potential of the progressive alignment in adjusting the retrieval scope. Note that devising an automated method for timestamp selection may improve performance. However, our experimental results suggest that setting the timestamp to a small range, e.g.,  $1 \sim 15$  yields reasonable performances and within this small range, a general and consistent setting ( $T' = 10$ ) enables leading performances across diverse datasets, e.g, MSRVTT, LSMDC, Charades, and VATEX. Thus, we adopted the empirical timestamp setting to retain computational efficiency and method extensibility.

**Effectiveness on Learning the Modality Gap** In Table 9, we study the effect of DITS on CLIP embedding space, to uncover how DITS bridge the modality gap. We statistically compare DITS with a ‘‘DITS fix CLIP’’ baseline, in terms of the averaged modality gap, averaged similarity, and the performance. Specifically, we compute the  $L_1$ -norm of the modality gap (cosine similarity) for each relevant pair and compute the averaged value. Overall, DITS can bridge the modality gap by effectively aligning the CLIP embedding space, yielding better performance. In Fig. 5, we further provide the distribution of the similarity change. Note that Fig. 4 computes the  $L_2$ -norm of the modality gap distribution, corresponding to the Euclidean distance. The values (such as mean around 1.28) are scaled by  $L_2$ -norm and are non-bounded. We also provide the mean absolute distance version (using  $L_1$ -norm) of the same data in Fig. 6 (left). As shown,  $L_1$ -norm enables a different scale and is also non-bounded. The difference between ‘‘Joint Train’’ and ‘‘Fix CLIP’’ is more remarkable under the  $L_1$ -norm scale. We also provide the cosine similarities version of the same data in Fig. 6 (right). As shown, cosine similarities are bounded (i.e.,  $[-1, 1]$ ). The similarity values are generally enlarged with joint train, being consistent with Fig. 4 and Fig. 6 (left).

#### A.4 Computational Cost and Efficiency

In this section, we discuss the computational efficiency for both training (Table 10) and inference (Table 11). The proposed DITS does not need much additional resources for training and reports comparable inference efficiency. We specifically compare with both non-diffusion methods (e.g., X-



Table 10: Training resource usage comparison with on MSRVTT dataset.

Methods	GPU Memory (MB)	GPU Request	Training Time (h)
X-Pool [Gorti et al., 2022]	18986	1× RTX3090	14.67
DiffusionRet (Stage 1)	27255	–	105.50
DiffusionRet (Stage 2)	8654	–	106.85
DiffusionRet (Total) [Jin et al., 2023]	27255	2× RTX3090	212.35
T-MASS [Wang et al., 2024a]	20390	1× RTX3090	14.74
DITS (Ours)	20950	1× RTX3090	14.90

Table 11: Inference time efficiency and GPU usage comparison on MSRVTT dataset.

Methods	GPU Memory (MB)	Inference Time (s)	R@1
X-Pool [Gorti et al., 2022]	5452	65.35	46.9
TS2-Net [Liu et al., 2022b]	2835	119.50	47.0
DiffusionRet [Jin et al., 2023]	3375	64.23	49.0
T-MASS [Wang et al., 2024a]	5452	76.41	50.2
DITS (Ours)	5464	70.17	51.9

Pool, TS2-Net, T-MASS) and diffusion-based method (e.g., DiffusionRet). All training and inference costs are measured with the same computational platform (2 NVIDIA RTX3090 GPU-24GB, Intel i9-10900X CPU).

For the training, since the proposed truncated sampler DITS starts from an pre-aligned text embedding and empirically needs small numbers of iterations, it requires much smaller training time compared with conventional diffusion model-based method. Besides, DITS enjoys a comparable GPU memory usage, being easily deployed on a single GPU.

For inference, DITS requires comparable GPU memory usage and inference efficiency with previous methods. Since there is no need to refer to the video embedding at each iteration of the alignment, we can compute and cache all the aligned embeddings beforehand. Thus we can skip the iterative sampling when performing retrieval. We find this method makes DITS faster than non-diffusion methods such as TS2-Net and T-MASS.

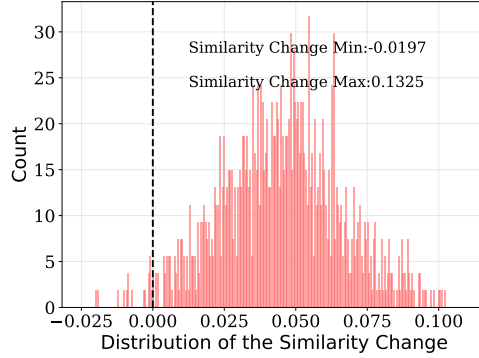


Figure 5: Similarly value subtraction between DITS and “DITS fix CLIP” (relevant pairs). DITS enables higher similarity values, yielding positive-valued histogram shown above. By comparison, DITS can effectively bridge the modality gap by aligning the CLIP embedding space.

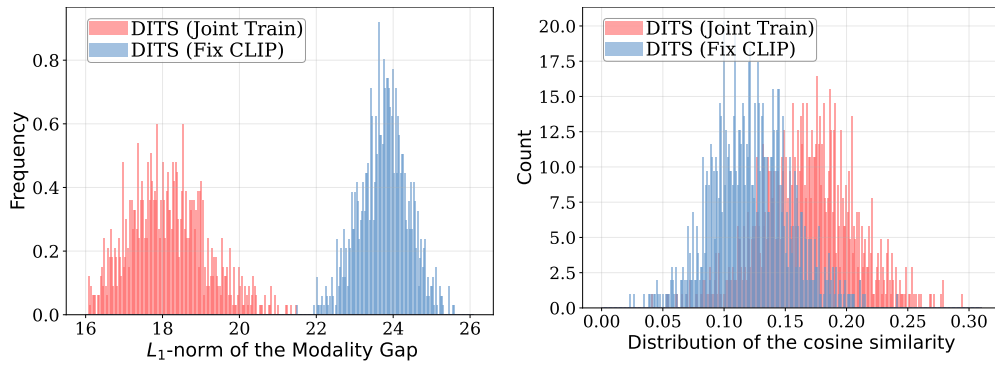


Figure 6: DITS reduce the modality gap for the **relevant** pairs. *Left*: The modality gap comparison upon  $L_1$ -norm (Mean absolute distance, unbounded). *Right*: The modality gap comparison measured by the cosine similarity (bounded, i.e.,  $[0, 1]$ ).

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discussed the limitations of the work performed by the authors in the supplementary.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose all the information needed to reproduce the main experimental results of the paper in the supplementary. We provided the code in the supplementary and claimed to further release the code and pretrained models in the abstract.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The manuscript and the supplementary provides detailed information in reproduce the results. We provided the code in the supplementary and claimed to further release the code and pretrained models in the abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The manuscript and the supplementary provides detailed information about the experimental setting/details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Error bar is not applicable to this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

#### 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computer resources has been reported.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

#### 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

#### 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discuss both potential positive societal impacts and negative societal impacts of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.



- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.