Distribution-Alignment Bridge for Uncertainty-Aware Text-to-Video Retrieval

Kyeongmo Chae[1], Jihoon Lee[1], and Sangtae Ahn[1*]

School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, South Korea
{ckm1994, leejh98123, stahn}@knu.ac.kr

Abstract. This paper proposes the Distribution-Alignment Bridge (DAB), a framework that reconceptualizes text-to-video retrieval as a distribution alignment task rather than traditional deterministic point matching. By modeling both text and video embeddings as Gaussian distributions defined by mean and variance, DAB explicitly accounts for modality-specific uncertainty. We employ a deterministic, diffusion-inspired bridge to iteratively refine text distributions toward their target video distributions through a truncated refinement process. This approach unifies probabilistic embedding and distributional transformation into a cohesive, end-to-end trainable system. To optimize cross-modal similarity, we introduce a distribution-aware contrastive loss based on Kullback–Leibler divergence. Extensive evaluations on MSR-VTT, MSVD, and VATEX benchmarks confirm that DAB significantly outperforms existing probabilistic and diffusion-based baselines, while providing calibrated uncertainty-aware ranking through bridge-induced distributional margins.

Keywords: Text-to-Video Retrieval · Probabilistic Embedding · Distribution Bridge · Uncertainty Modeling · Cross-Modal Alignment · Contrastive Learning

1 Introduction

Text-to-video retrieval (TVR), a pivotal task at the intersection of computer vision and natural language processing, aims to identify the most semantically relevant video from a large corpus given a textual query. The main challenge lies in the inherent modality gap, i.e., the representational disparity between the symbolic, structured nature of text and the dense, spatio-temporal nature of video. While recent methods leverage pretrained vision–language models such as CLIP [21] to project both modalities into a shared latent space, this modality gap persists and continues to hinder retrieval accuracy, especially under linguistic ambiguity and one-to-many correspondences.

A notable step forward is the Diffusion-Inspired Truncated Sampler (DITS) [28], which reframes alignment as iterative refinement rather than one-shot projection. DITS demonstrated that directly applying generative diffusion models to

* Corresponding author: Sangtae Ahn (stahn@knu.ac.kr)

retrieval suffers from instability: random noise initialization and $\mathcal{L}_2$-based objectives are ill-suited for ranking. Instead, DITS performs a deterministic truncated refinement that starts from meaningful text features and gradually evolves them toward target video embeddings under contrastive supervision, achieving both stability and efficiency. However, despite these advances, DITS remains fundamentally point-wise: it treats text–video alignment as a mapping between single vectors and thus overlooks semantic uncertainty and one-to-many relations intrinsic to multimodal data. A single textual query can validly describe visually diverse scenes, and a single video can correspond to multiple semantically distinct captions. Collapsing such diversity into a single embedding point oversimplifies cross-modal semantics and restricts representational flexibility.

To overcome this limitation, we propose the Distribution-Alignment Bridge (DAB), a framework that casts alignment as a distribution-to-distribution transformation. Instead of deterministic vectors, DAB encodes both text and video as Gaussian distributions–the mean captures core semantics, while the variance reflects ambiguity and contextual diversity. At the heart of DAB is a Distribution Bridge, a diffusion-inspired yet deterministic transformation that iteratively refines the textual distribution toward the target video distribution through smooth, drift-guided updates. Unlike generative diffusion models, the bridge is sampling-free and therefore stable and efficient for retrieval. Specifically, DAB first extracts textual and frame-level video embeddings using CLIP, then aggregates frame features through a transformer-based cross-attention module to form text-conditioned video representations. A shared probabilistic encoder maps each modality to Gaussian parameters $(\mu, \log \sigma^2)$, explicitly separating semantic centers from uncertainty. The bridge predicts drift terms $(\Delta\mu, \Delta \log \sigma^2)$ over truncated refinement steps to evolve the textual distribution toward its video counterpart. Crucially, DAB is trained with a distribution-aware contrastive objective, in which the pairwise cost is a divergence between distributions–KL divergence by default (with W_2 as a comparator)– so that alignment becomes sensitive to both mean and variance, i.e., to information content as well as geometry. This yields systematic distribution-space refinement of $(\mu, \log \sigma^2)$: rather than mapping a query to a single point or simply reducing variance, DAB learns a deterministic, sampling-free trajectory that jointly updates semantic centers and uncertainty scales while preserving modality-specific uncertainty. Empirically, this not only improves top-k recall but also yields a substantial reduction in Mean Rank (MnR), demonstrating a globally more coherent embedding space rather than isolated top-rank gains.

The main contributions are as follows.

- We identify the limitations of point-wise refinement (e.g., DITS) and introduce deterministic distribution-space refinement of $(\mu, \log \sigma^2)$, which jointly transports semantic centers and uncertainty scales without stochastic sampling.
- We formulate a bidirectional contrastive objective over directional KL costs, providing ranking-sensitive alignment of both mean and variance.

– Extensive experiments demonstrate state-of-the-art retrieval performance and show that DAB produces calibrated uncertainty signals and a semantically coherent embedding space.

2 Related Work

2.1 Text-to-Video Retrieval

Text-to-video retrieval (TVR) aims to find the most semantically relevant video for a given textual query. Early studies [18,19] learned a shared embedding space where similarity is computed using cosine or dot-product similarity. With large-scale vision–language pre-training, CLIP-based frameworks such as CLIP4Clip [17] transferred image–text alignment knowledge to the video domain, achieving strong generalization. X-Pool [7] introduced a text-conditioned pooling module that aggregates frame features via cross-attention guided by textual context, while subsequent works extended this to hierarchical interaction [6, 15, 20] and multimodal cues such as audio [12]. Despite such progress, the modality gap between compact textual descriptions and complex, dynamic video content remains an open challenge.

2.2 Diffusion-based Retrieval

Diffusion models [5, 9, 26] progressively refine random noise into structured data through a learned denoising process. Although originally developed for high-fidelity generation, diffusion mechanisms have been adopted for multimodal synthesis, including text-to-image [22, 24] and text-to-video [10, 25]. However, retrieval differs fundamentally from generation: it requires stable similarity estimation rather than stochastic sampling. DiffusionRet [13] attempted to model similarity as a generative denoising process, but its random sampling introduces unnecessary variance. DITS [28] improved upon this by employing a truncated, deterministic refinement that begins from meaningful text embeddings instead of random noise, demonstrating that diffusion-inspired iterative refinement can enhance alignment without full generative sampling. Nevertheless, DITS still operates on deterministic points–it refines feature vectors, not distributions–and thus cannot explicitly capture uncertainty or diversity.

2.3 Probabilistic and Uncertainty-Aware Retrieval

The inherent one-to-many nature of multimodal alignment makes uncertainty modeling crucial. Probabilistic embedding frameworks [3, 4, 6] represent each modality as a Gaussian distribution whose mean encodes semantics and variance reflects ambiguity. UATVR [6] extends this idea to TVR, allowing variance to capture semantic diversity. However, UATVR relies on sampled prototypes to compute similarity, which partially collapses probabilistic representations into deterministic comparisons. UAN [8] addresses uncertainty under

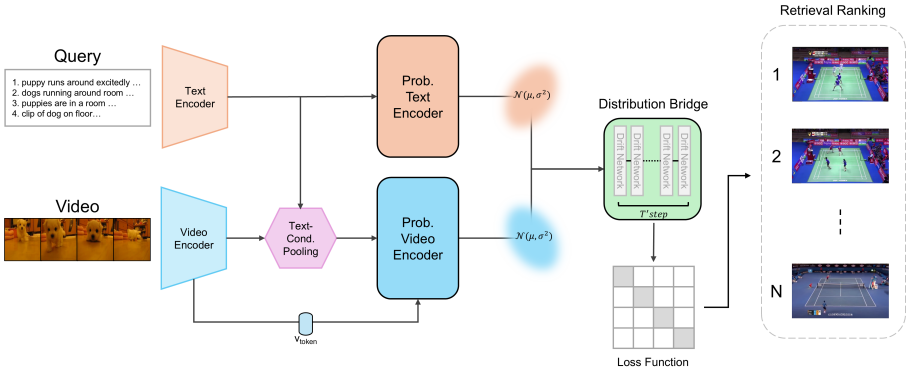

Fig. 1: Overall architecture of the proposed Distribution-Alignment Bridge (DAB). DAB reformulates text-to-video retrieval as a distribution alignment framework. It encodes text and video embeddings via CLIP, transforms them into Gaussian distributions through a probabilistic encoder, and employs a deterministic distribution bridge to progressively align text distributions toward video distributions for contrastive training.

cross-domain adaptation, but aligns features for domain transfer rather than learning modality-intrinsic distributional refinement. Recent fine-grained TVR methods such as Video-ColBERT [23] and Hybrid-Tower [14] strengthen late interaction or pseudo-query interaction at the feature level. DAB is complementary: it eliminates sampling and learns a deterministic bridge over $(\mu, \log \sigma^2)$, directly refining both mean and variance with directional KL costs.

3 Methods

Our proposed DAB reframes text-to-video retrieval from a deterministic point-matching problem into a distribution-alignment framework. As illustrated in Fig. 1, the model consists of three main components: (1) a foundational feature representation module that extracts text-conditioned video embeddings, (2) a probabilistic encoder that converts embeddings into Gaussian distributions to explicitly model semantic uncertainty, and (3) a distribution bridge that deterministically transforms the text distribution toward the video distribution through iterative, drift-based refinement (see Fig. 3). This bridge serves as the core of our framework, enabling stable and efficient distribution alignment without stochastic sampling.

3.1 Foundational Feature Representation

CLIP Encoding. Given a batch of text–video pairs, the pretrained CLIP model (ViT-B/32 and ViT-B/16) encodes text and video frames into a shared semantic space. The text encoder $G_\phi(\cdot)$ produces a sentence-level embedding $t \in \mathbb{R}^{B\times D}$,

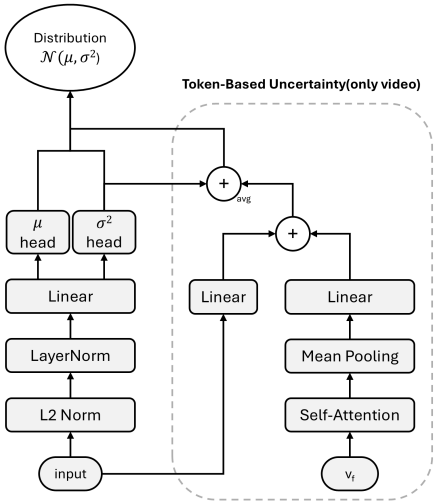

Fig. 2: Probabilistic Embedding Module. An input embedding is normalized and linearly projected, then split into two heads that predict the mean (μ) and variance (σ^2) of a Gaussian distribution. For video inputs, a token-based uncertainty branch refines the variance by applying self-attention and mean pooling over frame tokens. The fused variance and predicted mean form the final distribution $\mathcal{N}(\mu, \sigma^2)$.

while the visual encoder $F_\theta(\cdot)$ extracts frame-level features $\{v_f\}_{f=1}^{F}$ for each video, resulting in $V \in \mathbb{R}^{B\times F\times D}$. Here, B is the batch size, F is the number of sampled frames per video, and D is the embedding dimension. These features serve as the foundation for subsequent distributional modeling.

Text-Conditioned Frame Pooling. To obtain a representative video embedding that reflects textual context, we apply a transformer-based cross-attention pooling module [7]. For each text–video pair, the text embedding t serves as the *Query*, while frame embeddings $\{v_f\}_{f=1}^{F}$ serves as the *Keys* and *Values*. The resulting attention outputs are aggregated into text-conditioned video representations $v_c \in \mathbb{R}^{B\times B\times D}$, where each element $v_c^{(i,j)}$ corresponds to video j pooled conditioned on text i. Diagonal entries $v_c^{(i,i)}$ thus represent the integrated embeddings of matched text–video pairs. This cross-modal attention mechanism allows the model to focus on semantically relevant frames, laying the foundation for the subsequent probabilistic encoding and alignment stages.

3.2 Probabilistic Embedding Module

Probabilistic Head. As illustrated in Fig. 2, the probabilistic embedding module predicts Gaussian parameters through separate mean and log-variance heads, with an additional token-based uncertainty branch for video inputs. To capture

semantic ambiguity, each text and video embedding is modeled as a Gaussian distribution $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ rather than a deterministic point. Before projection, the input embedding x is L2-normalized as $\tilde{x} = x/(\|x\|_2 + 10^{-6})$. Two independent heads—each composed of a LayerNorm followed by a Linear layer—map the normalized feature to the mean and log-variance:

$$\begin{aligned} \mu &= W_\mu \,\mathrm{LN}(\tilde{x}) + b_\mu, \\ \log \sigma^2 &= W_{\ell v} \,\mathrm{LN}(\tilde{x}) + b_{\ell v}. \end{aligned} \tag{1}$$

The log-variance bias is initialized to -2.0 to encourage low initial uncertainty, and the predicted $\log \sigma^2$ is numerically clipped within $[-6, 2]$ for stability. The Probabilistic Head outputs $\mu, \log \sigma^2 \in \mathbb{R}^{B\times D}$ for each modality.

Token-Based Uncertainty for Video. When token-level representations $\{v_f\}_{f=1}^{L}$ (e.g., frame embeddings) are available, we introduce a *Token-Based Uncertainty Head* to refine variance estimation based on intra-video variability. The text-conditioned frame sequence is first processed by a single-head self-attention layer to capture contextual relations:

$$\mathbf{H} = \mathrm{SelfAttn}(\{v_f\}), \quad \mathbf{H} \in \mathbb{R}^{B\times L\times D}. \tag{2}$$

The token outputs are mean-pooled to summarize the spatio-temporal variability:

$$\bar{h} = \frac{1}{L}\sum_{f=1}^{L} H_f, \quad \bar{h} \in \mathbb{R}^{B\times D}. \tag{3}$$

Two linear projections are applied to the pooled feature and the global video embedding v_c, and their residual sum predicts the token-based log-variance:

$$\log \sigma^2_{v,\mathrm{tok}} = W_1 \bar{h} + b_1 + W_2 v_c + b_2. \tag{4}$$

Finally, the token-based and base variances are averaged to produce the final video uncertainty:

$$\log \sigma_v^2 = \tfrac{1}{2}\big(\log \sigma^2_{v,\mathrm{base}} + \log \sigma^2_{v,\mathrm{tok}}\big). \tag{5}$$

For stability, all log-variance outputs are clipped to the range $[-6, 2]$ during training.

This formulation decouples semantic content (mean) from spatio-temporal uncertainty (variance), allowing the model to represent diverse visual evidence across frames. Both modalities share the same *Probabilistic Encoder* that converts embeddings into Gaussian distributions. For text, $(\mu_t, \log \sigma_t^2)$ is obtained from the sentence-level embedding, whereas for video, $(\mu_v, \log \sigma_v^2)$ is derived from the pooled representation v_{repr} and refined using frame-level tokens to capture temporal uncertainty.

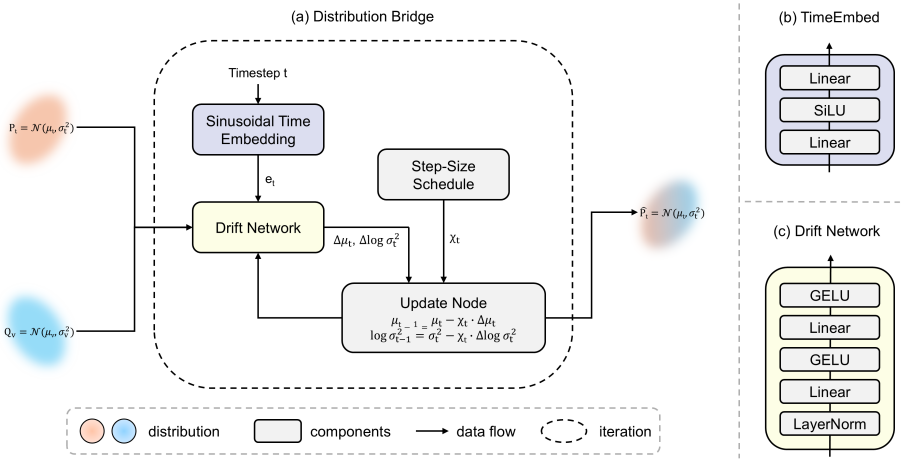

Fig. 3: Detailed structure of the Distribution Bridge module. The bridge iteratively updates the mean and variance of text distributions toward those of the corresponding video distributions using a time-conditioned drift network. Through multi-step deterministic refinement, it achieves stable distribution alignment without stochastic sampling.

3.3 Distribution Bridge

Architecture. Given a text distribution $P_t = \mathcal{N}(\mu_t, \sigma_t^2)$ and a target video distribution $Q_v = \mathcal{N}(\mu_v, \sigma_v^2)$, the distribution bridge deterministically transforms P_t toward Q_v through iterative refinement inspired by diffusion processes. It uses discrete timesteps and updates parameters across T' steps. The bridge comprises three components: a sinusoidal time embedding, a drift network that predicts parameter offsets, and a step-size schedule χ_t controlling the update magnitude.

Sinusoidal Time Embedding. Each timestep t is encoded as a continuous embedding e_t via sinusoidal positional encoding followed by a lightweight MLP with SiLU activation:

$$\text{TimeEmbed}(t) = \text{MLP}\left(\left[\begin{matrix}\sin\left(\frac{t}{10000^{2k/d}}\right)\\ \cos\left(\frac{t}{10000^{2k/d}}\right)\end{matrix}\right]_{k=0}^{d/2-1}\right), \quad e_t \in \mathbb{R}^{d_t}. \tag{6}$$

This embedding provides the drift network with temporal context, allowing parameter updates to vary smoothly with the refinement stage. The time embedding is parameterized by d, the embedding dimensionality, while k represents the index of the frequency components, ensuring that each dimension of e_t corresponds to a unique sinusoidal wavelength.

Drift Network. The drift module (*DriftDistMLP*) predicts mean and variance shifts given the current text and target video distributions, together with the time embedding:

$$[\Delta\mu_t,\, \Delta(\log\sigma_t^2)] = f_\theta^{\text{drift}}\big([\mu_t,\, \sigma_t,\, \mu_v,\, \sigma_v,\, e_t]\big), \\ \Delta\mu_t,\, \Delta(\log\sigma_t^2) \in \mathbb{R}^{B\times D}. \tag{7}$$

The network consists of a LayerNorm followed by two GELU-activated Linear layers, and two independent output heads for $\Delta\mu_t$ and $\Delta(\log\sigma_t^2)$. All output weights are zero-initialized to suppress abrupt drift during the early stages of training.

Step-Size Schedule. Each refinement step is scaled by a precomputed coefficient χ_t derived from a linear β-schedule, following the truncated diffusion strategy [28]:

$$\chi_t = \text{Norm}\left(\sqrt{1-\prod_{s=1}^{t}(1-\beta_s)}\right), \\ \beta_s = \beta_{\min} + (\beta_{\max}-\beta_{\min})\tfrac{s}{T}. \tag{8}$$

Here $\beta_{\min} = 10^{-4}$ and $\beta_{\max} = 2\times 10^{-2}$. The resulting χ_t values are normalized into the range $[0.2, 0.6]$ for stable gradient scaling, and only the last T' truncated steps are used to balance efficiency and stability.

Iterative Refinement. At each timestep, the parameters are updated as:

$$\mu_{t-1} = \mu_t - \chi_t\cdot\Delta\mu_t, \\ \log\sigma_{t-1}^2 = \log\sigma_t^2 - \chi_t\cdot\Delta(\log\sigma_t^2). \tag{9}$$

After each update, $\log\sigma^2$ values are clipped to a fixed range $[-6, 2]$ to maintain numerical stability. No stochastic noise is injected, producing a smooth and deterministic trajectory in distribution space. After T' iterations, the refined distribution $\hat{P}_t = \mathcal{N}(\mu_0, \sigma_0^2)$ is closely aligned with the target Q_v. This stepwise process enables gradual correction even for large initial modality gaps and yields stable, sampling-free distribution alignment.

This update also provides a simple stability intuition. If the PAC-trained drift approximates the residual $x_t - x^*$, where $x = (\mu, \log\sigma^2)$, then $x_{t-1} - x^* \approx (1-\chi_t)(x_t - x^*)$; since $\chi_t \in [0.2, 0.6]$, each step behaves as a contraction with rate at most 0.8 under this approximation. Together with zero-initialized drift heads and clipped log-variance, this keeps the deterministic refinement trajectory bounded in practice.

3.4 Optimization and Loss Function

Distributional Cost Matrix. Given a batch of B paired samples (t_i, v_i), we compute a pairwise divergence-based cost matrix between the bridged text

distributions $\hat{P}_{t_i}$ and the target video distributions Q_{v_j}:

$$C_{ij} = D\big(Q_{v_j} \,\|\, \hat{P}_{t_i}\big). \tag{10}$$

where $D(\cdot\|\cdot)$ denotes a probabilistic divergence such as the Kullback–Leibler (KL) divergence or, alternatively, a Wasserstein-based metric.

Distribution Alignment Metric. For two diagonal Gaussian distributions $P = \mathcal{N}(\mu_P, \mathrm{diag}(\sigma_P^2))$ and $Q = \mathcal{N}(\mu_Q, \mathrm{diag}(\sigma_Q^2))$ in $\mathbb{R}^D$, the closed-form KL divergence is:

$$D_{\mathrm{KL}}(P \parallel Q) = \tfrac{1}{2} \sum_{d=1}^{D} \left[\frac{(\mu_{Q,d} - \mu_{P,d})^2}{\sigma_{Q,d}^2} + \frac{\sigma_{P,d}^2}{\sigma_{Q,d}^2} - 1 + \log \frac{\sigma_{Q,d}^2}{\sigma_{P,d}^2} \right]. \tag{11}$$

This asymmetric form $D(Q_v\|\hat{P}_t)$ reflects the bridge direction from text to video distributions, measuring how well the refined text distribution approximates the target video distribution. In practice, all log-variance outputs are clipped within a fixed range for numerical stability. The KL divergence effectively penalizes both mean and variance mismatches, providing stable and uncertainty-aware updates for deterministic bridge refinement.

Probabilistic Alignment Contrastive (PAC) Loss. Using the cost matrix C, we adopt a bidirectional contrastive objective over directional KL costs:

$$\mathcal{L}_{\mathrm{PAC}} = -\frac{1}{B} \sum_{i=1}^{B} \left[\log \frac{\exp(-C_{ii}/\tau)}{\sum_{j=1}^{B} \exp(-C_{ij}/\tau)} + \log \frac{\exp(-C_{ii}/\tau)}{\sum_{j=1}^{B} \exp(-C_{ji}/\tau)} \right]. \tag{12}$$

where τ is a temperature parameter controlling contrastive sharpness. The first term ranks videos for each text, and the second ranks texts for each video. Importantly, the pairwise cost remains directional, $C_{ij} = D_{\mathrm{KL}}(Q_{v_j}\|\hat{P}_{t_i})$; thus DAB does not optimize symmetric KL or JS divergence. The bidirectional objective only requires matched directional costs to be smaller than in-batch negatives, and does not force text and video variances to become identical.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We evaluate DAB on three public benchmarks: MSR-VTT, MSVD, and VATEX. (1) MSR-VTT [32] contains 10K videos with 20 captions each. We train on the *9K* split and report on the 1K-A test set [17]. (2) MSVD [1] has 1,970 YouTube videos with ∼80K captions using the standard 1,200/100/670 split. (3) VATEX [29] includes ∼35K videos with multiple captions; we follow the split in [2]. We report *Recall@*$\{1, 5, 10\}$, *MdR*, and *MnR* (higher recall and lower ranks are better).

Table 1: Text-to-Video Retrieval on **MSR-VTT**.

Method	Venue	R@1	R@5	R@10	MdR	MnR	RSum
CLIP-ViT-B/32							
CLIP4Clip [17]	Neurocomputing'22	44.5	71.4	81.6	2	15.3	197.5
X-Pool [7]	CVPR'22	46.9	72.8	82.2	2	14.3	201.9
UATVR [6]	ICCV'23	47.5	73.9	83.5	2	12.9	204.9
Video-ColBERT [23]	CVPR'25	48.1	74.9	83.9	–	–	206.9
PIG [14]	ICCV'25	48.6	72.8	81.6	–	–	203.0
DiffusionRet [13]	ICCV'23	49.0	75.2	82.7	2	12.1	206.9
UCOFIA [30]	ICCV'23	49.4	72.1	–	–	12.9	–
T-MASS [27]	CVPR'24	50.2	75.3	85.1	2	11.9	210.6
AVIGATE [12]	CVPR'25	50.2	74.3	83.2	–	–	207.7
NarVid [11]	CVPR'25	50.8	76.4	84.7	1	11.6	211.9
DITS [28]	NeurIPS'24	51.9	75.7	84.6	1	11.6	212.2
DAB	Proposed	**56.2**	**85.4**	**92.4**	**1**	**4.1**	**234.0**
CLIP-ViT-B/16							
X-Pool [7]	CVPR'22	48.2	73.7	82.6	2	12.7	204.5
UATVR [6]	ICCV'23	50.8	76.3	85.5	1	12.4	212.6
Video-ColBERT [23]	CVPR'25	51.0	77.1	85.5	–	–	213.6
AVIGATE [12]	CVPR'25	52.1	76.4	85.2	–	–	213.7
NarVid [11]	CVPR'25	52.7	77.7	85.6	1	12.3	216.0
T-MASS [27]	CVPR'24	52.7	77.1	85.6	1	10.5	215.4
DITS [28]	NeurIPS'24	55.0	79.8	87.1	1	10.0	221.9
DAB	Proposed	**57.6**	**86.2**	**92.4**	**1**	**4.17**	**236.2**

Table 2: Text-to-Video Retrieval on **MSVD**. Gray row denotes ViT-B/16 backbone.

Method	Venue	R@1	R@5	R@10	MdR	MnR	RSum
CLIP4Clip [17]	Neurocomputing'22	45.2	75.5	84.3	2	10.0	205.0
Video-ColBERT [23]	CVPR'25	46.0	75.0	84.0	–	–	205.0
UATVR [6]	ICCV'23	46.0	76.3	85.1	2	–	207.4
X-Pool [7]	CVPR'22	47.2	77.4	86.0	2	9.3	210.6
UCOFIA [30]	ICCV'23	47.4	77.6	–	–	–	–
DiffusionRet [13]	ICCV'23	47.9	77.2	84.8	2	15.6	209.9
PIG [14]	ICCV'25	47.9	75.9	83.4	–	–	207.2
Cap4Video [31]	CVPR'23	51.8	80.8	88.3	1	–	220.9
NarVid [11]	CVPR'25	53.1	81.4	88.8	1	–	223.3
DAB (ViT-B/32)	Proposed	**54.5**	**87.8**	**93.9**	**1**	**3.4**	**236.2**

Table 3: Text-to-Video Retrieval on **VATEX**. Gray rows denote ViT-B/16 backbone.

Method	Venue	R@1	R@5	R@10	MdR	MnR	RSum
CLIP4Clip [17]	Neurocomputing'22	55.9	89.2	95.0	1	–	240.1
X-Pool [7]	CVPR'22	60.0	90.0	95.0	1	3.8	245.0
UCOFIA [30]	ICCV'23	61.1	90.5	–	1	3.4	–
UATVR [6]	ICCV'23	61.3	91.0	95.6	1	3.3	247.9
T-MASS [27]	CVPR'24	63.0	92.3	96.4	1	3.2	251.7
AVIGATE [12]	CVPR'25	63.1	90.7	95.5	1	–	249.3
PIG [14]	ICCV'25	64.0	91.5	96.6	1	–	252.1
DITS [28]	NeurIPS'24	64.1	92.7	97.0	1	2.9	253.8
Video-ColBERT [23]	CVPR'25	66.8	92.9	96.8	1	–	256.5
NarVid [11]	CVPR'25	68.4	94.0	97.1	1	–	259.5
DAB (ViT-B/32)	Proposed	**65.4**	**93.1**	**97.2**	**1**	**2.4**	**255.7**
DAB (ViT-B/16)	Proposed	**70.7**	**95.7**	**98.3**	**1**	**1.9**	**264.7**

Implementation Details. We build on X-Pool [7]. Frames are resized to 224×224 and we uniformly sample F=12 frames per clip. CLIP ViT-B/32 and CLIP ViT-B/16 [21] are used for both text and video encoders, with a d=512 projection. A lightweight transformer pooling aggregates temporal information; dropout is 0.3 for MSR-VTT/VATEX and 0.4 for MSVD. We use AdamW [16] with weight decay λ=0.02 and a warm-up ratio of 0.1. Learning rates are 1e−6 (CLIP) and 1e−4 (probabilistic/bridge modules). Batch size is B=32; seed is 24. We train for 10 epochs on all datasets.

For the bridge, we adopt N=1000 total steps and truncated steps T'=32 with a linear schedule $\beta_t \in [10^{-4}, 2 \times 10^{-2}]$ following diffusion-inspired alignment [6,28]. Parameters $\Theta = \{\theta, \phi, \gamma\}$ are optimized jointly. Inference uses global text/video chunk sizes of 128/512. All experiments are in PyTorch on a single NVIDIA A6000 GPU.

4.2 Comparisons

We compare DAB with recent SOTA models on MSR-VTT, MSVD, and VATEX, including discriminative CLIP-based [7, 17], probabilistic [6, 27], and diffusion-inspired methods [13, 28]. Unless noted, results are obtained under consistent settings.

MSR-VTT. As shown in Table 1, DAB achieves the best overall performance among all diffusion-based and probabilistic approaches. It substantially improves R@1 to 56.2, surpassing DITS [28] by +4.3, and reduces MnR from 11.6 to 4.14, representing a 64% relative reduction. This indicates that DAB forms a globally coherent embedding space in which relevant items cluster near the top ranks, rather than relying on isolated top-1 matches. Compared to probabilistic

UATVR [6] and generative DiffusionRet [13], which rely on stochastic sampling, DAB's deterministic, sampling-free distribution alignment produces both stable convergence and strong generalization. Even against T-MASS [27], which introduces stochastic text diffusion, DAB achieves higher recall, confirming the efficacy of its distribution-to-distribution refinement.

MSVD. Table 2 shows consistent performance gains across all recall metrics. DAB reaches 54.5/87.8/93.9 at R@1/5/10, outperforming previous diffusion-based or probabilistic models by large margins. Despite the smaller dataset scale and simpler sentence structures, DAB's deterministic bridge improves R@1 over CLIP4Clip [17] by +9.3 and reduces MnR to 3.4. We found that higher dropout (0.4) and slightly longer training (10 epochs) help mitigate overfitting on MSVD, while the coarse-to-fine refinement of DAB naturally regularizes the feature space.

VATEX. As shown in Table 3, DAB further improves the R@1 score to 65.4, surpassing DITS by +1.3 and T-MASS by +2.4, while achieving the lowest MnR among B/32-style comparisons. With ViT-B/16, DAB reaches 70.7/95.7/98.3 at R@1/5/10 and obtains the best overall VATEX result. The gains on VATEX suggest that DAB benefits from richer textual descriptions. Attribute-rich captions provide more semantic cues for estimating uncertainty, while the deterministic bridge tightens cross-modal alignment without collapsing modality-specific variance. This helps explain the consistent improvements in both recall and MnR. These gains are achieved without post-processing techniques such as inverted softmax or dual normalization, showing that sampling-free distribution alignment alone is highly effective.

In summary, DAB delivers strong state-of-the-art performance across MSR-VTT, MSVD, and VATEX, achieving leading results on MSR-VTT and MSVD among the compared methods and the best overall VATEX result with the ViT-B/16 backbone. Beyond higher recall, the pronounced reductions in MnR consistently show that DAB constructs a globally coherent and uncertainty-aware embedding space through distribution-level refinement.

4.3 Ablation Study

We analyze the effect of key components and hyperparameters of DAB on MSR-VTT (trained for 5 epochs) with CLIP ViT-B/32, reporting text-to-video results.

Effect of Truncated Steps T'. Table 4 examines the impact of the truncated refinement steps. As T' increases, the iterative bridge refines alignment more precisely, and performance steadily improves, peaking at $T' = 32$. Although $T' = 64$ slightly improves R@5 and R@10, its latency increases from 51.8s to 68.9s compared with $T' = 32$. FLOPs scale nearly linearly with T' because each step applies the same lightweight drift MLP. We therefore adopt $T' = 32$ as

Table 4: Accuracy-efficiency trade-off for truncated steps T' on MSR-VTT (5 epochs).

T'	R@1	R@5	R@10	MdR	MnR	RSum	Latency(s)	FLOPs(T)
8	49.0	82.0	89.0	2	5.8	220.0	39.5	69.73
16	50.6	80.7	89.4	1	5.6	220.7	43.9	139.47
32	**52.6**	82.7	90.8	1	**5.3**	226.1	51.8	278.93
64	51.9	**84.2**	**91.1**	1	5.6	**227.2**	68.9	557.85

the default setting, as it achieves the best R@1 while maintaining a favorable accuracy-efficiency trade-off.

Effect of Probabilistic Embedding and Distribution Bridge. The component ablation in Table 5 highlights the contribution of each module. Starting from X-Pool, incorporating the probabilistic embedding substantially reduces MnR and enhances overall recall by modeling semantic uncertainty. Applying the bridge alone produces limited improvement, suggesting that distribution-level refinement is most effective only when guided by uncertainty. Combining both components results in the best overall performance, validating that DAB's strength lies in its deterministic yet uncertainty-aware distribution refinement.

Table 5: Component ablation and distributional cost comparison on MSR-VTT (5 epochs).

(a) Component Analysis

Method	R@1	R@5	R@10	MnR
Baseline	46.9	72.8	82.2	14.3
w/ Prob.	44.3	76.6	87.2	7.4
w/ Bridge	43.2	76.6	86.3	8.5
DAB	52.6	82.7	90.8	5.3

(b) Distributional Cost

Cost	R@1	R@5	R@10	MnR
KL	52.6	82.7	90.8	5.3
W_2	46.8	79.7	88.4	6.6

The cost ablation further shows that replacing directional KL with W2 degrades R@1 from 52.6 to 46.8 and increases MnR from 5.3 to 6.6, supporting KL as a more ranking-sensitive divergence for DAB.

4.4 Uncertainty and Semantic-Neighborhood Diagnostics

We further examine whether DAB's distributional variables provide meaningful uncertainty signals beyond recall. First, the learned variances remain modality-asymmetric at convergence. As shown in Table 6, the average video variance is 5.87 times that of the text variance, while the bridge variance lies between the two endpoints. This indicates that DAB does not homogenize text and video uncertainty, but navigates between a low-entropy textual anchor and a higher-entropy video distribution.

For uncertainty calibration, we use the bridge-induced KL margin as a confidence signal. When queries are sorted by the top-1–top-2 KL margin, R@1 reaches 85.6% at 25% coverage, compared with 56.2% under random ordering, and further reaches 94.0% at 10% coverage. This shows that DAB's uncertainty signal emerges from the bridge-induced distributional gap rather than from text

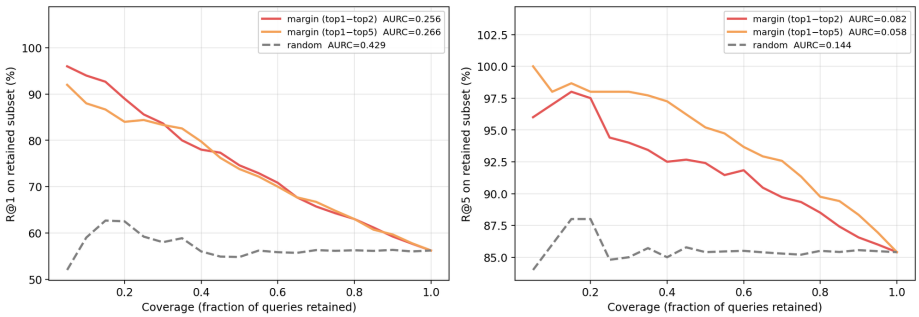

Fig. 4: Uncertainty calibration on MSR-VTT. Queries are ranked by the bridge-induced KL margin between the best and competing retrieved candidates, and retrieval accuracy is evaluated under selective prediction. Larger margins correspond to more reliable retrieval decisions, yielding substantially lower risk and AURC than random ordering. This indicates that DAB provides a calibrated confidence signal through distributional refinement.

Table 6: Learned variance asymmetry on MSR-VTT. Avg. σ^2 denotes the average variance of each distribution after training. The bridge variance remains between text and video variances, indicating that DAB preserves modality-specific uncertainty rather than collapsing both modalities to a shared variance scale during distribution refinement.

	Text	Bridge	Video
Avg. σ^2	0.507	0.935	2.979
Ratio	1.00	1.84	5.87

Table 7: Local semantic positives recovered in the top-10 on MSR-VTT. A video is counted as a semantic positive if its visual feature has a cosine similarity of at least τ with the ground-truth video. DAB consistently retrieves more semantic positives than random ranking across all thresholds tested.

τ	Elig.	Pos.	Rand.	DAB
0.3	921	50.05	0.50	5.06
0.4	864	17.75	0.18	3.22
0.5	655	7.28	0.07	2.18

variance alone. As illustrated in Fig. 4, This trend is consistent across the full risk–coverage curve. Compared with random ordering, KL-margin ranking substantially reduces the area under the risk–coverage curve (AURC), from 0.429 to 0.256 for R@1 and from 0.144 to 0.082 for R@5, indicating that the bridge-induced margin provides a reliable confidence estimate across a wide range of operating points.

Finally, we evaluate local one-to-many behavior by counting semantic positives in the top-10, where positives are videos whose visual features have cosine similarity at least τ to the ground-truth video. As shown in Table 7, DAB retrieves substantially more local semantic positives than random ranking across all thresholds. This supports local semantic-neighborhood organization rather than exhaustive global multimodal coverage.

This aggregate trend is also visible at the instance level. Fig. 5 visualizes the retrieval-score distribution for the query "*a little girl does gymnastics*" over

1,000 MSR-VTT test videos. Scores are negative KL-based costs, so larger values indicate closer distributional alignment. DAB ranks the annotated ground-truth video first and also places several gymnastics-related clips among the top results, including "*some girls are practicing gymnastics*", "*video of gymnasts practicing to roll*", and "*a girl doing gymnastics in the front yard*". This example qualitatively illustrates the same local-neighborhood behavior measured in Table 7: DAB does not only separate one annotated positive from unrelated negatives, but organizes semantically related videos near the query in the ranking space.

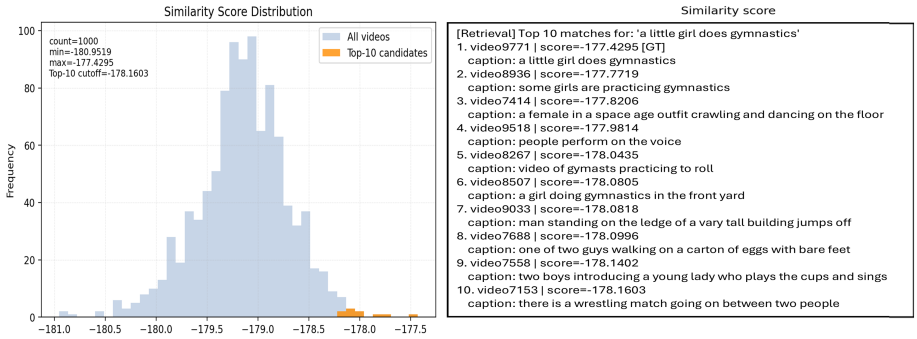

Fig. 5: Qualitative local-neighborhood example. For the query "*a little girl does gymnastics*", DAB ranks the annotated ground-truth video first and retrieves multiple gymnastics-related clips near the top. The highlighted top-ranked region provides an instance-level illustration of the local semantic-neighborhood behavior quantified in Table 7.

5 Conclusion

We introduced DAB, a framework that recasts text-to-video retrieval as distribution-to-distribution alignment. By encoding both modalities as Gaussians and refining them via a deterministic Distribution Bridge, DAB learns a sampling-free trajectory in the space of $(\mu, \log \sigma^2)$, jointly updating semantic centers and uncertainty scales. Central to this success is a bidirectional contrastive objective over directional KL costs, which improves ranking-sensitive alignment compared with W2 and preserves meaningful uncertainty signals throughout refinement. Empirically, DAB achieves strong results on MSR-VTT, MSVD, and VATEX. Further diagnostics show that learned variance asymmetry is preserved, bridge-induced KL margins provide calibrated retrieval confidence, and DAB organizes local semantic neighborhoods for ambiguous queries. These results support DAB as an uncertainty-aware retrieval framework whose uncertainty signal emerges from distributional refinement rather than from variance magnitude alone.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2025-02214941). This research was supported by the Regional Innovation System Education(RISE) Glocal 30 program through the Daegu RISE Center, funded by the Ministry of Education(MOE) and the Daegu, Republic of Korea(2025-RISE-03-001).

References

1. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
2. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10638–10647 (2020)
3. Chun, S.: Improved probabilistic image-text representations. arXiv preprint arXiv:2305.18171 (2023)
4. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8415–8424 (2021)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
6. Fang, B., Wu, W., Liu, C., Zhou, Y., Song, Y., Wang, W., Shu, X., Ji, X., Wang, J.: Uatvr: Uncertainty-adaptive text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13723–13733 (2023)
7. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5006–5015 (2022)
8. Hao, X., Zhang, W.: Uncertainty-aware alignment network for cross-domain video-text retrieval. Advances in Neural Information Processing Systems **36**, 38284–38296 (2023)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
11. Hur, C., Hong, J.h., Lee, D.h., Kang, D., Myeong, S., Park, S.h., Park, H.: Narrating the video: Boosting text-video retrieval via comprehensive utilization of frame-level captions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24077–24086 (2025)
12. Jeong, B., Park, J., Kim, S., Kwak, S.: Learning audio-guided video representation with gated attention for video-text retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26202–26211 (2025)
13. Jin, P., Li, H., Cheng, Z., Li, K., Ji, X., Liu, C., Yuan, L., Chen, J.: Diffusionret: Generative text-video retrieval with diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2470–2481 (2023)

14. Lan, B., Xie, R., Zhao, R., Sun, X., Kang, Z., Yang, G., Li, X.: Hybrid-tower: Fine-grained pseudo-query interaction and generation for text-to-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24497–24506 (2025)
15. Lee, S., Yu, Y., Kim, G., Breuel, T., Kautz, J., Song, Y.: Parameter efficient multimodal transformers for video representation learning. arXiv preprint arXiv:2012.04124 (2020)
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
18. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019)
19. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: Proceedings of the 2018 ACM on international conference on multimedia retrieval. pp. 19–27 (2018)
20. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
22. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
23. Reddy, A., Martin, A., Yang, E., Yates, A., Sanders, K., Murray, K., Kriz, R., De Melo, C.M., Van Durme, B., Chellappa, R.: Video-colbert: Contextualized late interaction for text-to-video retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19691–19701 (2025)
24. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
25. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
27. Wang, J., Sun, G., Wang, P., Liu, D., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Text is mass: Modeling as stochastic embedding for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16551–16560 (2024)
28. Wang, J., Wang, P., Liu, D., Guan, Q., Dianat, S., Rabbani, M., Rao, R., Tao, Z.: Diffusion-inspired truncated sampler for text-video retrieval. Advances in Neural Information Processing Systems **37**, 3882–3906 (2024)

29. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
30. Wang, Z., Sung, Y.L., Cheng, F., Bertasius, G., Bansal, M.: Unified coarse-to-fine alignment for video-text retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2816–2827 (2023)
31. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10704–10713 (2023)
32. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)