Imagine Before Concentration: Diffusion-Guided Registers Enhance Partially Relevant Video Retrieval

Supplementary Material

A. More Details on Method

A.1. Derivation of ELBO in Eq. (2)

Given the following predictive function:

$$p_{\theta,\phi}(Q|V) = \int \underbrace{p_\theta(Q|V,r)}_{\text{concentration}} \cdot \underbrace{p_\phi(r|V)}_{\text{imagination}} dr, \tag{18}$$

where $p_\phi(r|V)$ denotes the video branch that first generates global registers, and $p_\theta(Q|V,r)$ models the register-augmented cross-modal matching. Our training objective is to maximize $p_{\theta,\phi}(Q|V)$. We introduce a variational posterior $q_\varphi(r|Q_a)$ to approximate the true posterior $p(r|Q_a)$.

First, we rewrite Eq. (18) by introducing $q_\varphi(r|Q_a)$:

$$\begin{aligned} &\log p_{\theta,\phi}(Q|V) \\ &= \log \int p_\theta(Q|V,r) p_\phi(r|V) dr \\ &= \log \int p_\theta(Q|V,r) \frac{p_\phi(r|V)}{q_\varphi(r|Q_a)} q_\varphi(r|Q_a) dr. \end{aligned} \tag{19}$$

Next, we invoke Jensen's inequality [5], leveraging the concavity of the log function, which satisfies:

$$\log \mathbb{E}[X] \geq \mathbb{E}[\log X]. \tag{20}$$

Thus, the logarithm can be moved outside the integral, yielding a tractable lower bound:

$$\begin{aligned} &\log p_{\theta,\phi}(Q|V) \\ &\geq \int q_\varphi(r|Q_a) \log \left(p_\theta(Q|V,r) \frac{p_\phi(r|V)}{q_\varphi(r|Q_a)} \right) dr. \end{aligned} \tag{21}$$

We subsequently decompose the logarithmic term within the integrand:

$$\begin{aligned} &\log \left(p_\theta(Q|V,r) \frac{p_\phi(r|V)}{q_\varphi(r|Q_a)} \right) \\ &= \log p_\theta(Q|V,r) + \log \frac{p_\phi(r|V)}{q_\varphi(r|Q_a)}. \end{aligned} \tag{22}$$

Hence, the lower bound becomes:

$$\begin{aligned} \log p_{\theta,\phi}(Q|V) \geq &\int q_\varphi(r|Q_a) \log p_\theta(Q|V,r) dr \\ &- \int q_\varphi(r|Q_a) \log \frac{q_\varphi(r|Q_a)}{p_\phi(r|V)} dr. \end{aligned} \tag{23}$$

Therefore, we obtain the final form of the ELBO in Eq. (2):

$$\begin{aligned} \log p_{\theta,\phi}(Q|V) \geq &\mathbb{E}_{q_\varphi(r|Q_a)} \left[\log p_\theta(Q|V,r) \right] \\ &- \mathbb{KL} \left[q_\varphi(r|Q_a) \,\|\, p_\phi(r|V) \right]. \end{aligned} \tag{24}$$

A.2. Relationship between Eq. (7) and L_{dre}

Following [7], Bayes' rule gives:

$$q(\hat{\boldsymbol{q}}_t|\hat{\boldsymbol{q}}_{t-1}, \hat{\boldsymbol{q}}_0) = \frac{q(\hat{\boldsymbol{q}}_{t-1}|\hat{\boldsymbol{q}}_t, \hat{\boldsymbol{q}}_0) q(\hat{\boldsymbol{q}}_t|\hat{\boldsymbol{q}}_0)}{q(\hat{\boldsymbol{q}}_{t-1}|\hat{\boldsymbol{q}}_0)}. \tag{25}$$

Armed with this new equation, we can retry the derivation resuming from the ELBO in Eq. (7) by viewing $\hat{\boldsymbol{q}}$ as $\hat{\boldsymbol{q}}_0$:

$$\begin{aligned} &\log p(\hat{\boldsymbol{q}}_0) \\ &= \log \int p(\hat{\boldsymbol{q}}_{0:T}) \mathrm{d}\hat{\boldsymbol{q}}_{1:T} \\ &= \log \mathbb{E}_{q(\hat{\boldsymbol{q}}_{1:T}|\hat{\boldsymbol{q}}_0)} \left[\frac{p(\hat{\boldsymbol{q}}_{0:T})}{q(\hat{\boldsymbol{q}}_{1:T}|\hat{\boldsymbol{q}}_0)} \right] \\ &\geq \underbrace{\mathbb{E}_{q(\hat{\boldsymbol{q}}_1|\hat{\boldsymbol{q}}_0)} \left[\log p_\phi(\hat{\boldsymbol{q}}_0|\hat{\boldsymbol{q}}_1) \right]}_{\text{(reconstruction term)}} - \underbrace{\mathbb{KL}(q(\hat{\boldsymbol{q}}_T|\hat{\boldsymbol{q}}_0) \,\|\, p(\hat{\boldsymbol{q}}_T))}_{\text{(prior matching term)}} \\ &- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\hat{\boldsymbol{q}}_t|\hat{\boldsymbol{q}}_0)} \left[\mathbb{KL}(q(\hat{\boldsymbol{q}}_{t-1}|\hat{\boldsymbol{q}}_t, \hat{\boldsymbol{q}}_0) \,\|\, p_\phi(\hat{\boldsymbol{q}}_{t-1}|\hat{\boldsymbol{q}}_t)) \right]}_{\text{(denoising matching term)}}, \end{aligned} \tag{26}$$

where (i) the reconstruction term corresponds to the negative reconstruction error over $\hat{\boldsymbol{q}}_0$; (ii) the prior matching term is constant with no trainable parameters and can thus be ignored during optimization; and (iii) the denoising matching terms constrain $p_\phi(\hat{\boldsymbol{q}}_{t-1} \mid \hat{\boldsymbol{q}}_t)$ to align with the tractable ground-truth transition $q(\hat{\boldsymbol{q}}_{t-1} \mid \hat{\boldsymbol{q}}_t, \hat{\boldsymbol{q}}_0)$ [7]. Consequently, ϕ is optimized to iteratively recover $\hat{\boldsymbol{q}}_{t-1}$ from $\hat{\boldsymbol{q}}_t$. Following [4], the denoising matching terms can be simplified as

$$\sum_{t=2}^{T} \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\hat{\boldsymbol{q}}_t, t)\|_2^2 \right], \tag{27}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $\boldsymbol{\epsilon}_\phi(\hat{\boldsymbol{q}}_t, t)$ is parameterized by a neural network (e.g., U-Net [4]) to predict the noise $\boldsymbol{\epsilon}$ that generates $\hat{\boldsymbol{q}}_t$ from $\hat{\boldsymbol{q}}_0$ in the forward process [7]. A detailed derivation of Eq. (7) and Eq. (26) is provided in [7].

L_{dre} extends Eq. (27) by incorporating a conditioning variable $\boldsymbol{c}$. Inspired by [4], the reconstruction term in Eq. (26) has a relatively minor effect; hence, L_{dre} is employed to optimize the denoising matching terms in Eq. (26), thereby providing an approximate estimation of Eq. (26) for training.

A.3. Further details of DreamPRVR architecture

Diffusion Register Estimator (DRE) As illustrated in Fig. 7(a), the DRE block follows an MLP-based architecture incorporating Layer Normalization [1], activation functions, and linear projection layers. We use $N_{\text{dre}} = 2$ blocks.

Condition Generator The condition $\mathbf{c} \in \mathbb{R}^{N_r \times d}$ is obtained via a simple cross-attention mechanism between $\boldsymbol{V}_v \in \mathbb{R}^{N_v \times d}$ and learnable parameters, as illustrated in Fig. 7 (b), and can be formulated as

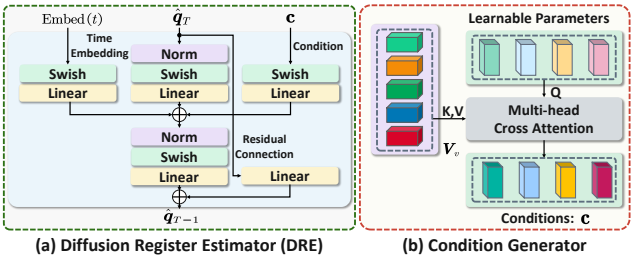

Figure 7. (a) Illustration of the proposed Diffusion Register Estimator Block (DRE). Embed(t), $\boldsymbol{q}_T$ and $\boldsymbol{c}$ denote the temporal embedding, the latent embedding corrupted by t-step noise and the guided condition, respectively. (b) Condition generator for DRE.

$$\mathbf{c} = \mathrm{CA}(LP, \boldsymbol{V}_v, \boldsymbol{V}_v) \in \mathbb{R}^{N_r \times d}, \tag{28}$$

where CA denotes cross-attention, and $LP \in \mathbb{R}^{N_r \times d}$ represents learnable parameters.

Asymmetric Attention Mask We retain the Gaussian self-attention as in [6, 9] and instead define two cross-attention patterns through a designed masking strategy, as illustrated in Fig. 2(d). Given the global registers $\boldsymbol{r}_0$ and video embeddings $\boldsymbol{V}_o$, the cross-attention configurations are defined as follows. For video embeddings:

$$\text{Query} = \boldsymbol{V}_o, \quad \text{Key} = \text{Value} = \texttt{Concat}([\boldsymbol{V}_o, \boldsymbol{r}_0]). \tag{29}$$

For global registers:

$$\text{Query} = \boldsymbol{r}_0, \quad \text{Key} = \text{Value} = \boldsymbol{V}_o. \tag{30}$$

A.4. Learning Objectives

Standard Similarity Retrieval Loss L_{sim} Following prior works [3, 6, 10], we employ the widely adopted triplet loss [2] L^{trip} and InfoNCE loss [8, 11] L^{nce} for PRVR. A text–video pair is treated as positive if the video contains a moment relevant to the query; otherwise, it is considered negative. Given a positive pair (Q, V), the triplet ranking loss over a mini-batch $\mathcal{B}$ is defined as follows:

$$L^{trip} = \frac{1}{n} \sum_{(Q,V)\in\mathcal{B}} \{\max(0, m + S(Q^-, V) - S(Q, V)) + \max(0, m + S(Q, V^-) - S(Q, V))\}, \tag{31}$$

where m denotes the margin, Q^- and V^- represent the negative text for V and the negative video for Q, respectively, and the similarity score $S(\cdot,\cdot)$ is computed as in Eq. (20). The infoNCE loss is computed as:

$$L^{nce} = -\frac{1}{n} \sum_{(Q,V)\in\mathcal{B}} \{\log(\frac{S(Q,V)}{S(Q,V) + \sum_{Q_i^-\in\mathcal{N}_Q} S(Q_i^-, V)}) + \log(\frac{S(Q,V)}{S(Q,V) + \sum_{V_i^-\in\mathcal{N}_V} S(Q, V_i^-)})\}, \tag{32}$$

where $\mathcal{N}_Q$ and $\mathcal{N}_V$ represent the negative texts and videos of V and Q within the mini-batch $\mathcal{B}$, respectively. Finally , L_{sim} is defined as:

Algorithm 1 Register Generation Process during Training

Input: Video features $\boldsymbol{V}_v$, all textual features from the video $\boldsymbol{q}$, condition features $\boldsymbol{c}$, timesteps T, noise schedule $\{\beta_t\}_{t=1}^T$, diffusion register estimator (DRE) ϵ_ϕ, probabilistic variational sampler (PVS), textual perturbation sampler (TPS)
Output: Optimal Registers $\boldsymbol{r}_0$, diffusion loss L_D
Initialize loss accumulator $L_D \leftarrow 0$
Precompute $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ for all t
Feature Encoding
$p(\boldsymbol{r}_T|\boldsymbol{V}_v) \leftarrow \text{PVS}(\boldsymbol{V}_v)$ ▷ Using Eq. (8)
Sample $\boldsymbol{r}_T \leftarrow p(\boldsymbol{r}_T|\boldsymbol{V}_v) \sim \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2 I)$
$p(\hat{\boldsymbol{q}}|\boldsymbol{q}) \leftarrow \text{TPS}(\boldsymbol{q})$ ▷ Using Eq. (6)
Sample $\hat{\boldsymbol{q}} \leftarrow p(\hat{\boldsymbol{q}}|\boldsymbol{q}) \sim \mathcal{N}(\boldsymbol{\mu}_{\hat{q}}, \boldsymbol{\sigma}_{\hat{q}}^2 I)$
Forward Diffusion Process
$\hat{\boldsymbol{q}}_0 \leftarrow \hat{\boldsymbol{q}}$
for $t = 1$ **to** T **do**
  Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $\hat{\boldsymbol{q}}_t \leftarrow \sqrt{\bar{\alpha}_t}\hat{\boldsymbol{q}}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ ▷ Add noise via Eq. (13)
  $\hat{\epsilon} \leftarrow \epsilon_\phi(\hat{\boldsymbol{q}}_t, t, \boldsymbol{c})$ ▷ Predict noise
  $L_{\text{dre}} \leftarrow ||\epsilon - \hat{\epsilon}||^2$ ▷ Calculate loss via Eq. (16)
  $L_D \leftarrow L_D + L_{\text{dre}}$ ▷ Accumulate loss
end for
Reverse Generation Process
Sample $\hat{\boldsymbol{q}}_T \leftarrow \boldsymbol{r}_T$ ▷ Start generation
for $t = T$ **to** 1 **do**
  Predict noise $\hat{\epsilon} \leftarrow \epsilon_\phi(\hat{\boldsymbol{q}}_t, t, \boldsymbol{c})$
  if $t > 1$ **then**
    Sample $z \sim \mathcal{N}(0, \mathbf{I})$
    $\hat{\boldsymbol{q}}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_k}}\left(\hat{\boldsymbol{q}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}\right) + \sqrt{\beta_t}z$
  else
    $\hat{\boldsymbol{q}}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\hat{\boldsymbol{q}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}\right)$
  end if
end for
$\boldsymbol{r}_0 \leftarrow \hat{\boldsymbol{q}}_0$
return $\boldsymbol{r}_0, L_D$ ▷ Return features and loss

$$L_{\text{sim}} = L_c^{\text{trip}} + L_f^{\text{trip}} + \lambda_c L_c^{\text{nce}} + \lambda_f L_f^{\text{nce}}, \tag{33}$$

where f and c denote the objectives of the frame-scale and clip-scale branches, respectively and λ_f and λ_c are the corresponding hyper-parameters.

Query Diversity Loss L_{div} Following Wang et al. [9], given a collection of text queries in the mini-batch $\mathcal{B}$, the query diversity loss is defined as:

$$\ell(i,j) = (1 + \cos(\boldsymbol{q}_i, \boldsymbol{q}_j)) \log(1 + e^{\omega(\cos(\boldsymbol{q}_i, \boldsymbol{q}_j)+\delta)}),$$
$$L_{\text{div}} = \frac{2}{M_q(M_q-1)} \sum_{1\le i,j\le M_q, i\neq j} \ell(i,j), \tag{34}$$

where $\delta > 0$ is a margin factor, $\omega > 0$ is a scaling factor and M_q is the number of text queries relevant to a video.

A.5. Relationship between $L_{\text{DreamPRVR}}$ and L_{total}

$L_{\text{DreamPRVR}}$ is the theoretical training objective defined in Eq. (3). It consists of two components: (i) a KL-divergence term that enforces the registers to generate global contextual semantics consistent with the textual queries, and (ii) a likelihood term that strengthens video representation learning with register guidance, thereby facilitating improved cross-modal alignment and retrieval performance.

L_{total} is the practical training objective, comprising four components: L_{tssl}, L_{pvs}, L_{dre}, and L_{sim}. Among them, L_{tssl}, L_{pvs}, and L_{dre} jointly regularize the registers to generate text-consistent representations and capture richer textual semantics. These terms promote more effective register generation and correspond to optimizing the KL-divergence term in $L_{\text{DreamPRVR}}$. In addition, L_{sim} serves as the retrieval-oriented similarity learning objective, aiming to improve retrieval performance. This term aligns with maximizing the likelihood component in $L_{\text{DreamPRVR}}$.

A.6. Register Generation Process

Training Stage Please refer to Algorithm 1.
Inference Stage The procedure follows Algorithm 1, with the forward diffusion process and TPS sampling omitted.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021. 2

[3] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. Partially relevant video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 246–257, 2022. 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[5] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906. 1

[6] Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Hlformer: Enhancing partially relevant video retrieval with hyperbolic learning. *arXiv preprint arXiv:2507.17402*, 2025. 2

[7] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 1

[8] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[9] Yuting Wang, Jinpeng Wang, Bin Chen, Tao Dai, Ruisheng Luo, and Shu-Tao Xia. Gmmformer v2: An uncertainty-aware framework for partially relevant video retrieval. *arXiv preprint arXiv:2405.13824*, 2024. 2

[10] Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. Gmmformer: Gaussian-mixture-model based transformer for efficient partially relevant video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 2

[11] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video corpus moment retrieval with contrastive learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021. 2