Temporal Calibrating and Distilling for Scene-Text Aware Text-Video Retrieval

Zhiqian Zhao[1,2*], Liang Li[2†], Lei Shen[1], Xichun Sheng[4], Yaoqi Sun[5], Fang Kang[3], Chenggang Yan[1]

[1]Hangzhou Dianzi University, Hangzhou, China,
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China,
[3]Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland,
[4]Macao Polytechnic University, Macao, China,
[5]Lishui University, Lishui, China
zhiqian.zhao@hdu.edu.cn, liang.li@ict.ac.cn

Abstract

Existing text-video retrieval methods mainly focus on single-modal video content (*i.e.*, visual entities), often overlooking heterogeneous scene text that is ubiquitous in human environments. Although scene text in videos provides fine-grained semantics for cross-modal retrieval, effectively utilizing it presents two key challenges: (1) Temporally dense scene text disrupts sync with sparse video frames, obstructing video understanding; (2) Redundant scene text and irrelevant video frames hinder the learning of discriminative temporal clues for retrieval. To address them, we propose a temporal scene-text calibrating and distilling (TCD) network for text-video retrieval. Specifically, we first design a window-OCR captioner that aggregates dense scene text into OCR captions to facilitate feature interaction. Next, we devise a heterogeneous semantics calibration module that leverages scene text as a self-supervised signal to temporally align window-level OCR captions and frame-level video features. Further, we introduce a context-guided temporal clue distillation module to learn the complementary and relevant details between scene text and video modalities, thereby obtaining discriminative temporal clues for retrieval. Extensive experiments show that our TCD achieves state-of-the-art performance on three scene-text related benchmarks.

Demo — https://tcd365.github.io

Introduction

Conventional text-video retrieval methods often struggle to distinguish visually similar but semantically distinct videos, as they primarily rely on coarse-grained visual content (*e.g.*, objects and actions). However, scene text, which is prevalent in human-centric environments, often conveys fine-grained semantic clues (*e.g.*, street names) that play a role in disambiguating similar-looking videos. In light of this, researchers have proposed scene-text aware text-video retrieval, incorporating scene text as a complementary modality to improve retrieval accuracy (Wu et al. 2025).

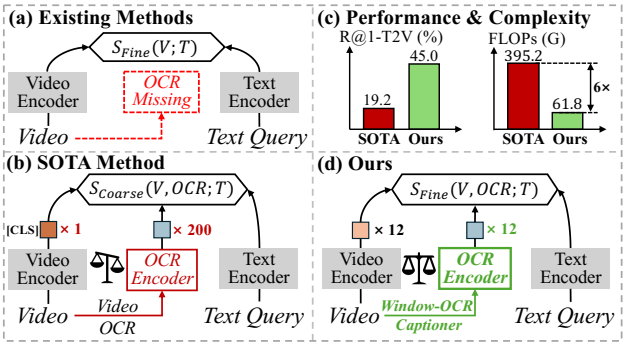

Figure 1: (a) Existing methods overlook fine-grained scene text in videos. (b) SOTA method (Wu et al. 2025) treats scene text as isolated instances, resulting in token imbalance issue that limits fine-grained interactions. (d) Our window-OCR captioner condenses scene text into OCR captions, achieving better performance with lower complexity (c).

While leveraging the discriminative yet noisy scene text in text-image retrieval is already considered a challenging task (Mafla et al. 2020), extending it to the video domain presents even greater difficulties. One key challenge is how to effectively encode the temporally dense scene text in videos for fine-grained retrieval. Existing temporal scene text encoding methods adopt either frame-level sub-sampling or video-level full-sampling, both of which have limitations. Sub-sampling methods (Tom et al. 2023; Zhou et al. 2024, 2025) typically employ uniformly spaced sampling to select a fixed number of frames, with each frame retaining a subset of its scene text instances. Nonetheless, this strategy is often suboptimal, as it may overlook temporally dense scene text in unselected frames, thereby hindering cross-modal matching and compromising retrieval performance. Therefore, full-sampling methods (Zhao et al. 2022a; Wu et al. 2025) aim to extract scene text words from the entire video, thereby capturing a more comprehensive representation of semantics. However, they treat redundant scene text as independent instances, resulting in dense OCR tokens (*i.e.*, 200). The token imbalance between dense OCR

*This work is done during the intern in VIPL group, ICT, CAS.
†Corresponding Author

inputs and sparse video frames hinders fine-grained feature interaction, leading to an incomplete understanding of video semantics (Fig. 1(b)) and high computation (Fig. 1(c)).

Another challenge lies in the obscurity of semantic clues across videos, making it difficult to extract query-relevant signals for effective retrieval. In feature space, key temporal clues are often drowned by redundant scene text and irrelevant video frames. Moreover, scene text in videos frequently suffers from recognition errors caused by occlusions or motion blur, rendering it even more challenging to extract discriminative and coherent temporal clues for retrieval.

In this paper, we propose a novel **T**emporal **C**alibrating and **D**istilling (TCD) network that aligns scene text with video frames while capturing discriminative temporal clues for text-video retrieval. Concretely, we first introduce a window-OCR captioner to transform temporal dense scene text instances into condensed OCR captions through video frame-centered non-overlap sliding windows. As shown in Fig. 1, this new window-level strategy reduces the input length from the previous 200 tokens to 12, matching the number of video frames, thereby enabling fine-grained interactions for video understanding and reducing computational cost. Then, we design a heterogeneous semantics calibration (HSC) module to align frame-level video frames and window-level OCR captions. It leverages heterogeneous scene text as a self-supervised signal and employs a cycle-consistency constraint to bridge the temporal scale discrepancy between these two modalities. Next, we apply the co-attention to enable fine-grained cross-modal interactions, so as to learn the most informative details for retrieval. In this way, we can obtain synergistic representation from fine-grained scene text and coarse-grained video content to distinguish visually similar videos for accurate retrieval.

Furthermore, we design a context-guided temporal clue distillation (CTCD) module to capture discriminative temporal clues for retrieval. Concretely, we employ learnable modality-specific clue tokens to aggregate contextual signals from both video frames and OCR captions. These tokens integrate complementary details across modalities via inter-modal attention, and capture intra-video and intra-OCR relevance through intra-modal attention. This helps mitigate the impact of recognition errors within OCR captions (*e.g.*, "cok" and "oke" $\rightarrow$ "coke") and capture motion dynamics across video frames. In this manner, we can obtain robust and discriminative temporal clues for effective text-video retrieval. The main contributions of this paper are as follows:

- We propose TCD, a scene-text enhanced text-video retrieval network that aligns dense scene text with sparse video frames while capturing discriminative temporal clues for accurate text-video retrieval.
- We design an HSC that leverages heterogeneous scene text as a self-supervised signal to dynamically align video-OCR sequences with different temporal scales for comprehensive video understanding.
- Extensive experiments demonstrate the superiority of our method on three datasets. Moreover, our window-OCR captioner is a plug-and-play module that can endow CLIP-based models with text-reading capability, achieving an average gain of 13.4% on T2V-R@1.

Related Work

Text-Video Retrieval. With the success of CLIP (Radford et al. 2021), several works (Xue et al. 2022; Jin et al. 2023b,a; Liu et al. 2022b; Zhao et al. 2022b; Fang et al. 2023; Ma et al. 2024; Wang et al. 2023) have expanded it from static images to dynamic videos for text-video retrieval. For example, CLIP4Clip (Luo et al. 2022) investigates the feasibility of transferring the CLIP model into text-video retrieval. X-CLIP (Ma et al. 2022) designs a multi-grained interaction between video and text for matching the same semantics. To extract query-related frame features, X-Pool (Gorti et al. 2022) applies cross-attention among query text and video frames. Cap4Video (Wu et al. 2023) brings a new sight that the caption of videos can enhance retrieval performance. T-MASS (Wang et al. 2024a) models text as a stochastic embedding to enrich textual content. However, existing methods overlook fine-grained scene text in videos, leading to ineffectiveness in text-related scenarios. We address this issue by integrating scene text into CLIP-based models through the proposed window-OCR captioner.

Leveraging Scene Text in Vision and Language Tasks. With the development of deep learning (Li et al. 2022, 2025; Zhang et al. 2024; Tu et al. 2024; Liu et al. 2022a; Cui et al. 2025; Zhang et al. 2025; Yin et al. 2025), scene text has gained much attention in image-language tasks, such as Text-VQA (Singh et al. 2019; Hu et al. 2020) and Text-Caption (Sidorov et al. 2020; Yang et al. 2020). Recent works (Mafla et al. 2020; Cheng et al. 2022; Zhao et al. 2025) focus on leveraging fine-grained scene text to enhance text-image retrieval. In video domain, scene text has wide applications in many VideoQA tasks, such as TextVideoQA (Zhao et al. 2022a; Tom et al. 2023), Grounded TextVideoQA (Zhou et al. 2024), and EgoTextVQA (Zhou et al. 2025), which target dynamic, grounded, and egocentric video scenarios, respectively. Recently, StarVR (Wu et al. 2025) introduces TextVR, a new dataset that incorporates scene text for text-video retrieval. However, existing methods struggle with computational explosion due to the temporal density of scene text. Therefore, we devise a new encoding way for video scene text that significantly reduces computational complexity.

Cycle-Consistency Learning. Leveraging the transitive nature of relationships, cycle-consistency provides a valuable learning signal for various computer vision tasks, including image alignment and translation (Zhu et al. 2017; Zhou et al. 2016, 2015), spatial-temporal alignment (Dwibedi et al. 2019; Wang et al. 2019; Jabri, Owens, and Efros 2020), visual question answering (Shah et al. 2019), and image captioning (Guo et al. 2019; Wang, Deng, and Jia 2024). COOT (Ging et al. 2020) adopts cycle-consistency into video-text domain for video representation learning. TCC (Dwibedi et al. 2019) employs cycle-consistency to align two video sequences at frame level. Different from them, we leverage the heterogeneous scene text as a bridging learning signal to calibrate sequences of various temporal scales across different modalities.

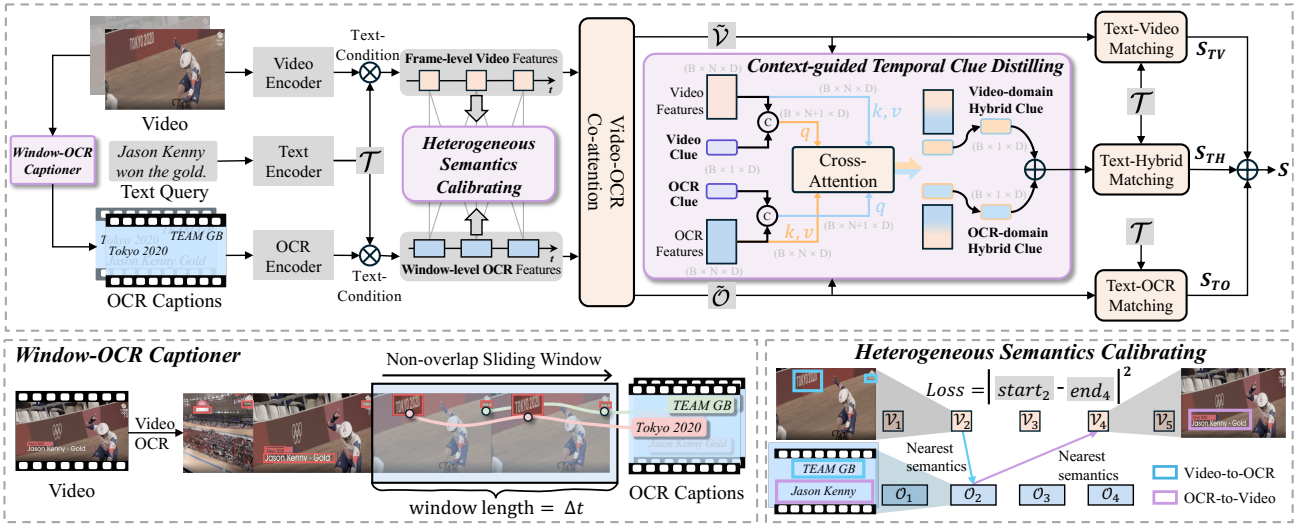

Figure 2: Pipeline of **TCD**. Window-OCR captioner condenses abundant scene text into OCR captions. Heterogeneous semantics calibrating leverages scene text present in both video and OCR captions as a self-supervised signal for temporal alignment. Context-guided temporal clue distilling employs clue tokens to learn clear temporal clues from inter- and intra-modal contexts.

Methodology

Preliminaries. Scene-text aware text-video retrieval dataset $\mathcal{D} = \{T^i; (V^i, O^i)\}_{i=1}^{|\mathcal{D}|}$ consists of a text query T^i, a target video V^i, and its associated OCR information O^i extracted by a video OCR model. Given a text query T^i, the objective is to find the most similar video-OCR pair (V^i, O^i) from the gallery $\{(V^k, O^k)\}_{k=1}^{|\mathcal{D}|}$ by computing similarity scores.

Overview. The structure of TCD framework is depicted in Fig. 2. Concretely, we first employ the proposed window-OCR captioner to aggregate raw OCR information into condensed OCR captions. Then, window-level OCR features and frame-level video features are extracted separately. To align these features with different temporal scales, we introduce a heterogeneous semantics calibration module. Next, video-OCR co-attention is employed to enable fine-grained cross-modal interactions. To capture discriminative temporal clues, we devise a context-guided temporal clue distillation module that learns complementary and relevant video-OCR features. Finally, we perform multi-modal matching to obtain similarity scores with multiple perspectives.

Window- & Frame-level Feature Extraction

In this section, we introduce our window-OCR captioner and describe the process of feature extraction.

Window-OCR Captioner. Current frame-level and video-level encoding strategies treat scene text as separate instances, leading to a large number of OCR tokens (*e.g.*, 12 frames × 15 = 180 or 200), which hinders fine-grained interactions between video and OCR elements. To address this, we propose a window-OCR captioner to reduce the number of OCR tokens from 200 in previous methods (Wu et al. 2025) to $N_O = 12$. Given a raw video V, we first obtain the OCR results consisting of N_w words and their frame indices by a video OCR model as $O = \{\mathbf{o}_n^{word}, \mathbf{o}_n^{frame}\}_{n=0}^{N_w}$, here N_w is usually larger than 200. Then, we group these words around selected video frames into N_O scene text groups,

$$G_m = \left\{ \mathbf{o}_i^{word} \mid i \in \mathbb{N} \right\}, \quad m = 1, 2, \ldots, N_O, \tag{1}$$

by non-overlap sliding windows with length $\Delta t = L/N_O$, where L denotes the video's duration, resulting in $N_O = N_V = 12$, as illustrated in Fig. 2. If there is no scene text, G_m can be an empty set ϕ. Note that if repeated scene text words appear within the same window, they are merged into one word, which helps reduce redundancy. Moreover, representing scene text as a complete sentence enables more compact and informative expression without semantic loss. Inspired by this, we transform each sampled scene text group into an OCR caption to reduce the token length. For each group, we construct an OCR caption using the pre-defined template: *There are scene texts: [G_m] in this frame.* If G_m is ϕ, we instead use *There is no scene text in this frame.* Thereafter, we obtain N_O window-level OCR captions.

Feature Extraction. We adopt the powerful CLIP model as our backbone. The backbone is used to extract features from text query, video frames, and OCR captions by their [CLS] token, formulating feature sequence $\mathcal{T} \in \mathbb{R}^{1\times D}$, $\mathcal{V} \in \mathbb{R}^{N_V \times D}$ and $\mathcal{O} \in \mathbb{R}^{N_O \times D}$, where N_V, N_O are the number of video frames and OCR captions, D is the feature dimension. Video frames and OCR captions often contain semantics that are irrelevant to the text query. To address this, we utilize the text query as a conditioning signal to extract the most relevant semantic features. Concretely, given video frames $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_{N_V}\}$ and OCR captions $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, ..., \mathcal{O}_{N_O}\}$, we compute their relevance to the text query $\mathcal{T}$ via temperature parameter τ as:

$$p_n = \frac{\exp((\mathcal{T})^\top \mathcal{V}_n/\tau)}{\sum_{n=1}^{N_V} \exp((\mathcal{T})^\top \mathcal{V}_n/\tau)}, \quad q_n = \frac{\exp((\mathcal{T})^\top \mathcal{O}_n/\tau)}{\sum_{n=1}^{N_O} \exp((\mathcal{T})^\top \mathcal{O}_n/\tau)}, \tag{2}$$

where p_n and q_n are the importance scores of the n-th frame and caption, respectively. Each frame-level video feature is then re-weighted as $\mathcal{V}_n \leftarrow p_n\mathcal{V}_n$, and each window-level OCR feature as $\mathcal{O}_n \leftarrow q_n\mathcal{O}_n$.

Heterogeneous Semantics Calibrating

We propose heterogeneous semantics calibrating (HSC) to align video and OCR sequences with different time-scales. Ideally, the cycle of the nearest semantics between these two sequences should form a closed-loop mapping ($V_2 \rightarrow O_2 \rightarrow V_2$). However, due to the time-scale differences in sampling strategies, OCR caption may include irrelevant scene text from other video frames (*i.e.*, "Jason Kenny Gold"), leading to incorrect mappings ($V_2 \rightarrow O_2 \rightarrow V_4$). Inspired by this, we propose leveraging the above cycle consistency and using heterogeneous scene text as a self-supervised signal to achieve temporal alignment and preserve prior knowledge.

Given a video sequence $\{\mathcal{V}_n\}_{n=1}^{N_V}$ and an OCR sequence $\{\mathcal{O}_n\}_{n=1}^{N_O}$, starting from node $\mathcal{V}_i$, its nearest semantics $\mathcal{O}_k$ is calculated as $\left[\mathcal{O}_k = \arg\min_{\mathcal{O}\in\{\mathcal{O}_n\}_{n=1}^{N_O}} dist(\mathcal{V}_i, \mathcal{O})\right]$, where $dist$ is a distance function (*e.g.*, cosine similarity or L2 distance). However, this hard matching operation is non-differentiable. To enable end-to-end optimization, we adopt a soft nearest neighbor formulation (Rocco et al. 2018; Goldberger et al. 2004; Dwibedi et al. 2019):

$$\bar{\mathcal{O}}_{\mathcal{V}_i} = \textstyle\sum_{j=1}^{N_O} \alpha_j \mathcal{O}_j, \text{where } \alpha_j = \frac{dist(\mathcal{V}_i, \mathcal{O}_j)}{\sum_{n=1}^{N_O} dist(\mathcal{V}_i, \mathcal{O}_n)}, \quad (3)$$

here α_j is the similarity score between $\mathcal{V}_i$ and $\mathcal{O}_j$. Then, similarly, we map the fused OCR representation $\bar{\mathcal{O}}_{\mathcal{V}_i}$ back to the video sequence $\{\mathcal{V}_n\}_{n=1}^{N_V}$, obtaining a soft index μ:

$$\mu = \textstyle\sum_{j=1}^{N_V} \beta_j j, \text{where } \beta_j = \frac{dist(\bar{\mathcal{O}}_{\mathcal{V}_i}, \mathcal{V}_j)}{\sum_{n=1}^{N_V} dist(\bar{\mathcal{O}}_{\mathcal{V}_i}, \mathcal{V}_n)}, \quad (4)$$

here β_j is the similarity score between $\bar{\mathcal{O}}_{\mathcal{V}_i}$ and $\mathcal{V}_j$. After that, we complete the semantic alignment cycle from the start index i to the end index μ. According to their index, this process is optimized by the calibration loss as:

$$\mathcal{L}_{hsc} = \|i - \mu\|^2, \quad (5)$$

so as to learn a semantically consistent representation.

Video-OCR Co-attention

To help cross-modal matching, we apply co-attention (Lu et al. 2019) to learn discriminative video and OCR representations through bidirectional interaction. We first take text-condition video and OCR features $\mathcal{V}$ and $\mathcal{O}$ as inputs, and then compute video-to-OCR and OCR-to-video attention.

Video-to-OCR Attention. Video frames can alleviate the semantic polysemy of scene text, enabling more accurate understanding. For example, the word "*Apple*" has distinct meanings in a mobile store and a fruit shop. Surrounding information from video frames can help clarify such differences. Specifically, we first project them into the common feature space as $Q_{\mathcal{V}} = \mathcal{V}W_Q$, $K_{\mathcal{O}} = \mathcal{O}W_K$, $V_{\mathcal{O}} = \mathcal{O}W_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{D\times D}$ are projection matrices, then the refined OCR features $\tilde{\mathcal{O}} \in \mathbb{R}^{N_O\times D}$ are calculated:

$$\tilde{\mathcal{O}} = \text{Softmax}\left(\frac{Q_{\mathcal{V}}K_{\mathcal{O}}^T}{\sqrt{D}}\right)V_{\mathcal{O}}. \quad (6)$$

OCR-to-Video Attention. Scene text also carries vital clues for grasping video semantics. For instance, the words "*Eiffel Tower*" in a sign indicate its location, providing context about the video's background. Therefore, we leverage OCR-to-video attention as below:

$$\tilde{\mathcal{V}} = \text{Softmax}\left(\frac{Q_{\mathcal{O}}K_{\mathcal{V}}^T}{\sqrt{D}}\right)V_{\mathcal{V}}, \quad (7)$$

where $\tilde{\mathcal{V}} \in \mathbb{R}^{N_V\times D}$ are refined video features and $Q_{\mathcal{O}} = \mathcal{O}W'_Q, K_{\mathcal{V}} = \mathcal{V}W'_K, V_{\mathcal{V}} = \mathcal{V}W'_V$ are projected by the learnable matrices $W'_Q, W'_K, W'_V \in \mathbb{R}^{D\times D}$, respectively.

Context-guided Temporal Clue Distilling

We propose context-guided temporal clue distilling (CTCD) to capture a discriminative context-guided temporal clue for accurate retrieval. To aggregate contextual information, we initialize the modal-specific clues as learnable embedding vectors $\mathcal{C}_{\mathcal{V}} \in \mathbb{R}^{1\times D}$ and $\mathcal{C}_{\mathcal{O}} \in \mathbb{R}^{1\times D}$ for video features and OCR features, respectively. We regard these modal-specific clues as another modality and then add different type embeddings, $\tilde{\mathcal{C}}^{type} \in \mathbb{R}^{1\times D}$, $\tilde{\mathcal{V}}^{type} \in \mathbb{R}^{N_V\times D}$, $\tilde{\mathcal{O}}^{type} \in \mathbb{R}^{N_O\times D}$ as:

$$\begin{bmatrix} \mathcal{C}'_{\mathcal{V}} & \mathcal{C}'_{\mathcal{O}} \\ \mathcal{V}' & \mathcal{O}' \end{bmatrix} = \begin{bmatrix} \mathcal{C}_{\mathcal{V}} & \mathcal{C}_{\mathcal{O}} \\ \tilde{\mathcal{V}} & \tilde{\mathcal{O}} \end{bmatrix} + \begin{bmatrix} \mathcal{C}^{type} & \mathcal{C}^{type} \\ \tilde{\mathcal{V}}^{type} & \tilde{\mathcal{O}}^{type} \end{bmatrix}. \quad (8)$$

Then, instead of performing vanilla cross-attention, which neglects intra-modality relevance within the query itself, we concatenate the modal-specific clue $\mathcal{C}'_{\mathcal{O}} \in \mathbb{R}^{1\times D}$ with OCR features $\mathcal{O}' \in \mathbb{R}^{N_V\times D}$ to leverage the intra-modality relevance within the query. Subsequently, we employ a multi-head cross-attention (MHCA) (Vaswani 2017) to mine the inter-modality complementary features with video $\hat{\mathcal{V}} \in \mathbb{R}^{N_V\times D}$. Above process leverages both intra- and inter-interactions to obtain the context OCR-domain hybrid clue $\mathcal{C}^*_{\mathcal{O}} \in \mathbb{R}^{1\times D}$:

$$\begin{gathered} [\mathcal{C}^*_{\mathcal{O}}; \mathcal{O}^*] = \text{MHCA}([\mathcal{C}'_{\mathcal{O}}; \mathcal{O}'], \hat{\mathcal{V}}, \hat{\mathcal{V}}), \\ \mathcal{C}^*_{\mathcal{O}} \leftarrow [\mathcal{C}^*_{\mathcal{O}}; \mathcal{O}^*] \cdot \hat{\mathbf{e}}_{\mathcal{O}}, \end{gathered} \quad (9)$$

where $\hat{\mathbf{e}}_{\mathcal{O}} = [1, 0, \ldots, 0]^\top \in \mathbb{R}^{1+N_O}$ is a one-hot vector to extract $\mathcal{C}^*_{\mathcal{O}}$. In this manner, OCR-domain hybrid clue is encouraged to leverage intra-modality relevance to alleviate OCR recognition errors caused by object occlusion or motion blur (*e.g.*, 'coff' and 'offee' in adjacent OCR captions are associated as 'coffee'). The video-domain hybrid clue $\mathcal{C}^*_{\mathcal{V}} \in \mathbb{R}^{1\times D}$ also follows the same procedure as:

$$\begin{gathered} [\mathcal{C}^*_{\mathcal{V}}; \mathcal{V}^*] = \text{MHCA}([\mathcal{C}'_{\mathcal{V}}; \mathcal{V}'], \hat{O}, \hat{O}), \\ \mathcal{C}^*_{\mathcal{V}} \leftarrow [\mathcal{C}^*_{\mathcal{V}}; \mathcal{V}^*] \cdot \hat{\mathbf{e}}_{\mathcal{V}}, \end{gathered} \quad (10)$$

here $\hat{\mathbf{e}}_{\mathcal{V}} = [1, 0, \ldots, 0]^\top \in \mathbb{R}^{1+N_V}$, and $\mathcal{V}^* \in \mathbb{R}^{N_V\times D}$. In this way, the video-domain hybrid clue aids in capturing detailed motion across intra-modality relevant continuous video frames. Finally, we integrate them into the tempo-

Method	TextVR					M4-ViteVQA				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Transformer-based										
ClipBERT (Lei et al. 2021)	6.3	16.3	26.7	30.0	183.3	-	-	-	-	-
Frozen (Bain et al. 2021)	7.4	22.7	32.5	26.0	132.3	-	-	-	-	-
StarVR (Wu et al. 2025)	19.2	38.2	47.5	13.0	138.2	22.5	44.3	55.9	8.0	47.9
CLIP-based (ViT-B/32)										
CLIP4Clip (Luo et al. 2022)	25.4	47.6	56.2	7.0	97.1	26.0	44.2	54.4	8.0	42.0
DiCoSA (Jin et al. 2023b)	26.4	48.7	58.0	6.0	75.6	40.9	61.7	70.0	3.0	25.8
X-CLIP (Ma et al. 2022)	28.3	50.2	58.0	5.0	108.4	41.5	61.1	69.7	3.0	35.6
X-Pool (Gorti et al. 2022)	23.5	42.5	51.7	9.0	104.8	40.5	58.8	67.5	3.0	38.4
TS2-Net (Liu et al. 2022b)	27.9	49.8	59.0	6.0	94.5	36.4	57.1	65.0	3.0	33.1
UATVR (Fang et al. 2023)	28.5	50.0	58.9	5.0	77.4	41.1	59.3	67.8	3.0	33.1
UCoFiA (Wang et al. 2023)	27.7	50.5	60.4	5.0	83.1	38.6	60.7	69.3	3.0	31.7
TeachCLIP (Tian et al. 2024)	25.3	46.3	55.1	7.0	87.1	32.9	51.1	60.3	5.0	33.6
T-MASS (Wang et al. 2024a)	28.4	48.9	57.7	6.0	113.3	44.5	62.4	72.1	2.0	25.3
+ Our Window-OCR Captioner										
CLIP4Clip (Luo et al. 2022)	33.5 +8.1	57.8 +10.2	66.3 +10.1	3.0	69.8	41.7 +15.7	61.7 +17.5	70.2 +15.8	3.0	31.4
X-CLIP (Ma et al. 2022)	37.4 +9.1	61.7 +11.5	71.0 +13.0	3.0	51.1	47.4 +5.9	66.4 +5.3	75.8 +6.1	2.0	23.2
X-Pool (Gorti et al. 2022)	36.3 +12.8	59.6 +17.1	67.8 +16.1	3.0	62.3	58.4 +17.9	74.0 +15.2	81.6 +14.1	**1.0**	16.0
T-MASS (Wang et al. 2024a)	45.0 +16.6	66.8 +17.9	74.4 +16.7	2.0	45.9	65.6 +21.1	78.5 +16.1	83.9 +11.8	**1.0**	17.2
TCD (Ours)	**60.4**	**82.8**	**88.2**	**1.0**	**8.0**	**68.8**	**85.4**	**89.4**	**1.0**	**8.2**

Table 1: Text-to-video (T2V) retrieval performance on scene-text aware text-video retrieval (TVR) datasets, TextVR (Wu et al. 2025) and M4-ViteVQA (Zhao et al. 2022a). **Green numbers** indicate performance gains achieved through our window-OCR captioner by incorporating OCR information for retrieval.

ral hybrid clue $\mathcal{C} \in \mathbb{R}^{1\times D}$ to enhance the clarity of the video event signal for further retrieval:

$$\mathcal{C} = (\mathcal{C}^*_{\mathcal{V}} + \mathcal{C}^*_{\mathcal{O}})/2. \tag{11}$$

Training Objective

Multi-modal Matching. To enhance retrieval robustness, we perform multi-modal matching, including text-video, text-OCR, and text-hybrid matching. As shown in Fig. 2, given processed features $\mathcal{T}^i$, $\mathcal{C}^i$, $\tilde{V}^i$, and $\tilde{O}^i$, we apply mean pooling to obtain $\mathcal{V}^i = \frac{1}{N_V}\sum_n \tilde{V}^i_n$ and $\mathcal{O}^i = \frac{1}{N_O}\sum_n \tilde{O}^i_n$. Then, similarity function $S(\mathcal{T}^i, (\mathcal{V}^i, \mathcal{O}^i))$ is defined as:

$$S(\mathcal{T}^i, (\mathcal{V}^i, \mathcal{O}^i)) = (S_{TH} + S_{TV} + S_{TO})/3, \tag{12}$$

where S_{TH}, S_{TV}, and S_{TO} are calculated as:

$$S_{TH} = \frac{\mathcal{T}^i(\mathcal{C}^i)^\top}{\|\mathcal{T}^i\|\,\|\mathcal{C}^i\|}, S_{TV} = \frac{\mathcal{T}^i(\mathcal{V}^i)^\top}{\|\mathcal{T}^i\|\,\|\mathcal{V}^i\|}, S_{TO} = \frac{\mathcal{T}^i(\mathcal{O}^i)^\top}{\|\mathcal{T}^i\|\,\|\mathcal{O}^i\|}, \tag{13}$$

Loss. The proposed network is trained in an end-to-end manner. Given a batch B of triplets $\{\mathcal{T}^k, (\mathcal{V}^k, \mathcal{O}^k)\}_{k=1}^B$, our network will predict a $B \times B$ similarity matrix. Then we use the InfoNCE loss to optimize the network as:

$$\begin{aligned}
\mathcal{L}_{\mathcal{T}\to\mathcal{V}} &= -\frac{1}{B}\sum_{i=1}^{B} log\frac{\exp(S(\mathcal{T}^i, (\mathcal{V}^i, \mathcal{O}^i))/\eta)}{\sum_j \exp(S(\mathcal{T}^i, (\mathcal{V}^j, \mathcal{O}^j))/\eta)}, \\
\mathcal{L}_{\mathcal{V}\to\mathcal{T}} &= -\frac{1}{B}\sum_{i=1}^{B} log\frac{\exp(S(\mathcal{T}^i, (\mathcal{V}^i, \mathcal{O}^i)/\eta)}{\sum_j \exp(S(\mathcal{T}^j, (\mathcal{V}^i, \mathcal{O}^i))/\eta)}, \\
\mathcal{L}_{ce} &= \frac{1}{2}(\mathcal{L}_{\mathcal{T}\to\mathcal{V}} + \mathcal{L}_{\mathcal{V}\to\mathcal{T}}),
\end{aligned} \tag{14}$$

where η is the learnable scaling parameter, $\mathcal{L}_{\mathcal{T}\to\mathcal{V}}$ and $\mathcal{L}_{\mathcal{V}\to\mathcal{T}}$ represents text-to-video retrieval loss and video-to-text retrieval loss, respectively. Besides, the network is self-supervised by the loss of heterogeneous semantics calibrating. Hence, the final loss function is optimized as:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{hsc}, \tag{15}$$

here λ is a trade-off parameter, and its effect is analyzed in the experiments.

Experiments

Datasets and Metrics

We conduct experiments on both scene-text rich TVR datasets (TextVR and M4-ViteVQA) and a scene-text sparse dataset (*i.e.*, MSR-VTT) to comprehensively evaluate our method under scene-text related scenarios. In brief, **TextVR** has 80% videos contain scene text more than 20 instances, **M4ViteVQA** reaches nearly 90% scene-text videos, and the videos in **MSR-VTT** only has 21% scene-text videos. For evaluation metric, we report Recall at Rank (R@K, K={1, 5, 10}), Median Rank (MdR), and Mean Rank (MnR) to show the retrieval performance.

Implementation Details

We use CLIP's visual encoder (ViT-B/32) as our video encoder and its textual encoder as text and OCR encoder. The feature dimension D is set to 512, the number of frames N_V and OCR captions N_O are set to 12, temperature parameter $\tau = 5$. Adam optimizer (Kingma 2014) is used to train the model with a batch size of 42 for 30 epochs. We apply a cosine schedule strategy with a linear warmup (Goyal 2017) to decay the learning rate. The initial learning rate is set to 1e-4. For the parameter λ of the HSC loss $\mathcal{L}_{hsc}$, we set $\lambda = 0.1$ for TextVR and M4-ViteVQA, and $\lambda = 0.05$ for MSR-VTT. Our model is trained on one NVIDIA RTX 4090 GPU.

Row	Core Modules			Text→Video				Video→Text			
	CTCD	*Co-Attn*	$\mathcal{L}_{hsc}$	R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
1				53.3	77.5	84.1	13.8	46.0	70.7	80.3	14.4
2	✓			54.9	79.2	85.2	10.8	47.3	76.0	84.2	9.9
3	✓	✓		58.9	81.6	86.8	10.9	49.2	76.9	85.0	10.5
4	✓		✓	56.0	79.3	85.8	11.7	48.9	76.1	84.5	11.0
5	✓	✓	✓	**60.4**	**82.8**	**88.2**	**8.0**	**51.9**	**78.3**	**86.5**	**7.7**

Table 2: Ablation study of each module on TextVR. Row 1 denotes the model where CTCD is replaced by cross-attention.

Method	MSR-VTT			
	R@1↑	R@5↑	R@10↑	MnR↓
CLIP-based (ViT-B/32)				
CLIP4Clip (Luo et al. 2022)	44.5	71.4	81.6	15.3
DiCoSA (Jin et al. 2023b)	47.5	74.7	83.8	13.2
Cap4Video (Wu et al. 2023)	49.3	74.3	83.8	12.0
X-CLIP (Ma et al. 2022)	46.1	73.0	83.1	13.2
X-Pool (Gorti et al. 2022)	46.9	72.8	82.2	14.3
TS2-Net (Liu et al. 2022b)	47.0	74.5	83.8	13.0
UATVR (Fang et al. 2023)	47.5	73.9	83.5	12.3
ECLIPSE (Lin et al. 2022)	44.9	71.3	81.6	15.0
TEFAL (Ibrahimi et al. 2023)	49.4	75.9	83.9	12.0
T-MASS (Wang et al. 2024a)	50.2	75.3	85.1	11.9
DIST (Wang et al. 2024b)	51.9	75.7	84.6	11.6
MUSE (Tang et al. 2024)	50.9	76.7	85.6	10.9
NarVid (Hur et al. 2025)	52.7	77.7	85.6	12.3
BIA (Bai et al. 2025)	52.1	76.8	86.3	9.7
BiMa (Le et al. 2025)	53.5	78.6	86.5	-
AVIGATE (Jeong et al. 2025)	50.2	74.3	83.2	-
TCD (w/o OCR)	54.5	77.5	84.0	8.5
TCD (w OCR)	**59.7**	**83.2**	**89.3**	**7.2**

Table 3: Text-to-video retrieval performance on MSR-VTT.

OCR Ratio	MSR-VTT		
	R@1	R@5	R@10
0.0	54.5	77.5	84.0
0.1	56.5	80.1	86.2
0.2	57.6	81.2	88.6
0.4	58.5	80.3	87.4
0.8	58.9	81.8	**89.4**
1.0	**59.7**	**83.2**	89.3

Table 4: Different OCR ratio.

Method	MSR-VTT		
	R@1	R@5	R@10
OCR-Free Setting (w/o OCR)			
Full	54.5	77.5	84.0
w/o Co-Attn	49.8	75.2	84.8
w/o $\mathcal{L}_{hsc}$	53.9	76.4	83.6
w/o CTCD	49.3	73.0	80.9
w/o TC	38.6	65.7	75.9

Table 5: Ablation study.

Comparisons with SOTA Methods

To show the effectiveness of scene text in TVR, we evaluate our method on TextVR, M4-ViteVQA, and MSR-VTT. As shown in Tab. 1, adding our window-OCR captioner yields **+5.3%–21.1%** gains in T2V-R@1, validating its effect. In addition, TCD achieves SOTA performance, benefiting from HSC's effective temporal alignment and CTCD's ability to distill discriminative temporal clues from redundant scene text. In Tab. 3, TCD surpasses existing modality-enhanced SOTA methods (*e.g.*, TEFAL, AVIGATE), showing the discriminative value of scene text.

Ablation Study

Ablation Study of Each Module on TextVR dataset. In Tab. 2, using heterogeneous semantics as a self-supervised signal (rows 4-5) yields better performance, indicating that

Method	Text→Video			Video→Text		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Addition	57.1	78.8	85.4	48.0	73.1	82.3
A×B+A	56.4	79.8	86.2	49.3	73.9	81.9
Concat+MLP	57.4	79.7	86.0	50.5	74.3	83.2
Cross-Attention	58.6	80.2	85.8	50.8	75.7	83.2
$\mathcal{C}^*_{\mathcal{V}}$	59.4	80.2	86.2	51.2	77.3	85.0
$(\mathcal{C}^*_{\mathcal{V}} + \mathcal{C}^*_{\mathcal{O}})$ / 2	**60.4**	**82.8**	**88.2**	**51.9**	**78.3**	**86.5**

Table 6: Ablation study of CTCD on TextVR. $\mathcal{C}^*_{\mathcal{V}}$ and $\mathcal{C}^*_{\mathcal{O}}$ denote video and OCR-domain hybrid clues, respectively.

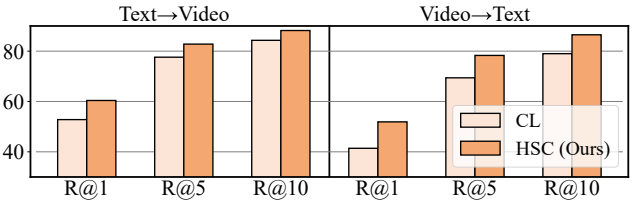

Figure 3: Ablation study of HSC on TextVR. CL is short for contrastive learning. CL regards same-index pairs as positives, all others are negatives for temporal alignment.

$\mathcal{L}_{hsc}$ allows temporal alignment and preserves video–OCR synchrony. Moreover, co-attention enables bidirectional fine-grained interaction, producing more discriminative representations for accurate retrieval.

Ablation Study on MSR-VTT dataset. For a comprehensive evaluation, we conduct experiments on different OCR ratios and perform ablations by removing scene text. As shown in Tab. 4, retrieval performance consistently improves as the OCR ratio increases, verifying the effectiveness of our method in leveraging scene text. Moreover, the results in Tab. 5 demonstrate the effectiveness of each component, especially the proposed CTCD module, which captures discriminative temporal clues for accurate retrieval.

Ablation Study of CTCD. In Tab. 6, we study CTCD with different fusion strategies on TextVR. We observe that replacing cross-attention with a learnable video clue $\mathcal{C}^*_{\mathcal{V}}$ allows the query to capture intra-video cues for motion, thereby achieving better performance. Moreover, adding an OCR clue $\mathcal{C}^*_{\mathcal{O}}$ helps model to handle recognition errors from motion blur or object occlusion. Further, the proposed CTCD gains **+1.8%** R@1 over SOTA fusion designs.

Ablation Study of HSC. Fig. 3 compares our HSC module with contrastive learning for preserving synchrony between two temporal sequences. In the setting of contrastive

Multi-modal Matching			Text→Video			
T-Video	*T-OCR*	*T-Hybrid*	R@1↑	R@5↑	R@10↑	MnR↓
✓			54.0	78.3	84.0	13.9
	✓		54.1	76.7	83.4	16.4
		✓	56.8	79.8	86.6	8.8
✓	✓		58.9	80.9	86.5	11.1
	✓	✓	58.7	80.7	86.9	9.7
✓		✓	60.0	82.1	87.0	**7.9**
✓	✓	✓	**60.4**	**82.8**	**88.2**	8.0

Table 7: Ablation study of varying granularity matching for text-to-video retrieval on TextVR dataset.

Method	T2V-R@1	GFLOPs	$\mathcal{O}(N^2d)$	Training Time
StarVR	19.2	395.2	$\mathcal{O}(40,401d)$	76.3h
TCD (Ours)	**60.4**	61.8	$\mathcal{O}(576d)$	8.5h

Table 8: Complexity and performance comparison.

λ	0.01	0.05	**0.1**	0.5	1	5	10
RSUM	442.6	431.9	**448.1**	446.6	443.5	446.3	414

Table 9: Ablation study of weight λ for $\mathcal{L}_{hsc}$ on TextVR.

learning, same-index pairs are positives and others are negatives. The results indicate that HSC performs better, showing that heterogeneous scene-text semantics offer a more flexible self-supervised signal than rigid contrastive alignment.

Effect of Multi-modal Matching. Tab. 7 shows a range of matching strategies with varying levels of granularity. We note that hybrid modality outperforms video and OCR alone by **+2.8%** and **+2.7%** R@1, indicating that the hybrid modality offers richer temporal clues by capturing contextual information. Notably, strategies that using the hybrid modality consistently achieve higher performance.

Effect of Scene Text for Retrieval. Tab. 10 shows that scene text in videos provides fine-grained, discriminative cues for retrieval. We observe that relying solely on video content yield low performance (25.3% and 35.8% R@1), while adding OCR modalities significantly improves performance(**+9.8%** and **+24.6%** R@1), indicating that scene text offers complementary information. Meanwhile, the text condition helps the model capture query-relevant semantics, thereby outperforming SOTA methods.

Computational Comparison. Tab. 8 shows that TCD outperforms StarVR by **+41.2%** with lower complexity. This efficiency comes from our window-OCR captioner, which reduces sequence length N from 201 to 24, thereby alleviating quadratic complexity $\mathcal{O}(N^2d)$ from $\mathcal{O}(40,401d)$ to $\mathcal{O}(576d)$ in attention, obtaining a **70×** complexity reduction and **9×** speed-up in training time.

Parameter Sensitivity. Tab. 9 shows that our TCD achieves the best results at $\lambda = 0.1$ on TextVR. Moreover, we report additional results on other datasets in appendix. The results show that λ is a data-dependent parameter. In detail, datasets containing more scene text need a larger λ to calibrate video-OCR sequences in the HSC module.

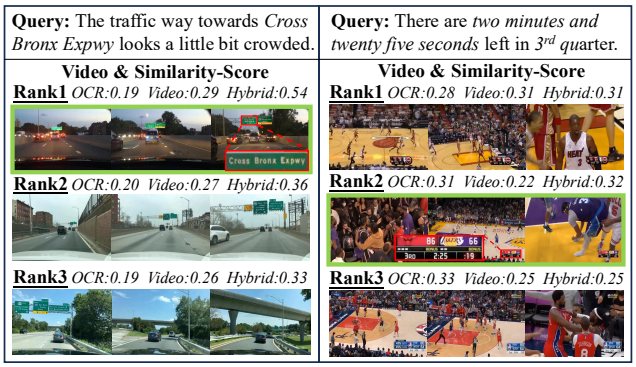

Figure 4: Visualization of text-to-video retrieval with **top-3 ranked** results on TextVR. Please zoom in for a better view.

TC	Modality	R@1↑	R@5↑	R@10↑	MnR↓
✗	OCR	25.2	45.4	52.7	169.4
✗	Video	25.3	48.0	56.5	96.3
✗	Video+OCR	35.1	57.9	67.0	64.0
✓	OCR	43.3	62.2	68.9	51.7
✓	Video	35.8	65.3	76.5	19.2
✓	Video+OCR	**60.4**	**82.8**	**88.2**	**8.0**

Table 10: Ablation study of text-condition (denoted as TC) and modality enhancement on TextVR with T2V results.

Qualitative Analysis

To better understand the proposed method, we present the visualization of retrieval results in Fig. 4. We can see that TCD can capture fine-grained scene text in videos (*e.g.*, "*Cross Bronx Expwy*" and "*2:25*") for accurate retrieval. This is attributed to the window-OCR captioner that balances token distribution between video and OCR modalities for fine-grained video understanding. On the left, TCD retrieves the ground truth video in rank 1 with a high hybrid similarity score. This demonstrates that the hybrid modality provides a discriminative temporal clue for retrieval by capturing contextual information between video and OCR. This demonstrates that the hybrid modality provides a discriminative temporal clue for retrieval by capturing contextual information between video and OCR.

Conclusion

This paper proposes a TCD network that aligns scene text with video frames while capturing a discriminative temporal clue for TVR. We first introduce a window-OCR captioner to aggregate abundant scene text from videos into continuous captions, thereby reducing computation complexity. Then, we apply scene text as a signal in a cycle manner to calibrate video-OCR sequences so as to preserve their prior sync for accurate video understanding. Further, we devise a context-guided temporal clue distillation module to capture crucial temporal clues by leveraging complementarity and relevance among redundant scene text and irrelevant video frames for accurate retrieval. Extensive experiments show that our TCD achieves SOTA results on three benchmarks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62322211), the National Key Research and Development Program of China under Grant (2023YFB4502800), the "Pioneer" and "Leading Goose" RD Program of Zhejiang Province(2024C01023, 2024C01107, 2023C01030, 2023C01046).

References

Bai, Z.; Xiao, T.; He, T.; Wang, P.; Zhang, Z.; Brox, T.; and Shou, M. Z. 2025. Bridging information asymmetry in text-video retrieval: A data-centric approach. In *The Thirteenth International Conference on Learning Representations*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Cheng, M.; Sun, Y.; Wang, L.; Zhu, X.; Yao, K.; Chen, J.; Song, G.; Han, J.; Liu, J.; Ding, E.; et al. 2022. Vista: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5184–5193.

Cui, Y.; Li, L.; Yin, H.; Gao, Y.; Sun, Y.; and Yan, C. 2025. Debiased Teacher for Day-to-Night Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2577–2587.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2019. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1801–1810.

Fang, B.; Wu, W.; Liu, C.; Zhou, Y.; Song, Y.; Wang, W.; Shu, X.; Ji, X.; and Wang, J. 2023. Uatvr: Uncertainty-adaptive text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13723–13733.

Ging, S.; Zolfaghari, M.; Pirsiavash, H.; and Brox, T. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33: 22605–22618.

Goldberger, J.; Hinton, G. E.; Roweis, S.; and Salakhutdinov, R. R. 2004. Neighbourhood components analysis. *Advances in neural information processing systems*, 17.

Gorti, S. K.; Vouitsis, N.; Ma, J.; Golestan, K.; Volkovs, M.; Garg, A.; and Yu, G. 2022. X-pool: Cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5006–5015.

Goyal, P. 2017. Accurate, large minibatch SG D: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Guo, L.; Liu, J.; Yao, P.; Li, J.; and Lu, H. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4204–4213.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9992–10002.

Hur, C.; Hong, J.-h.; Lee, D.-h.; Kang, D.; Myeong, S.; Park, S.-h.; and Park, H. 2025. Narrating the Video: Boosting Text-Video Retrieval via Comprehensive Utilization of Frame-Level Captions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24077–24086.

Ibrahimi, S.; Sun, X.; Wang, P.; Garg, A.; Sanan, A.; and Omar, M. 2023. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12054–12064.

Jabri, A.; Owens, A.; and Efros, A. 2020. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33: 19545–19560.

Jeong, B.; Park, J.; Kim, S.; and Kwak, S. 2025. Learning Audio-guided Video Representation with Gated Attention for Video-Text Retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26202–26211.

Jin, P.; Huang, J.; Xiong, P.; Tian, S.; Liu, C.; Ji, X.; Yuan, L.; and Chen, J. 2023a. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2472–2482.

Jin, P.; Li, H.; Cheng, Z.; Huang, J.; Wang, Z.; Yuan, L.; Liu, C.; and Chen, J. 2023b. Text-video retrieval with disentangled conceptualization and set-to-set alignment. *arXiv preprint arXiv:2305.12218*.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Le, H.; Chung, N.; Kieu, T.; Nguyen, A.; and Le, N. 2025. BiMa: Towards Biases Mitigation for Text-Video Retrieval via Scene Element Guidance. *arXiv preprint arXiv:2506.03589*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7331–7341.

Li, L.; Cong, G.; Qi, Y.; Zha, Z.-J.; Wu, Q.; Sheng, Q. Z.; Huang, Q.; and Yang, M.-H. 2025. Dubbing Movies via Hierarchical Phoneme Modeling and Acoustic Diffusion Denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, L.; Gao, X.; Deng, J.; Tu, Y.; Zha, Z.-J.; and Huang, Q. 2022. Long short-term relation transformer with global gating for video captioning. *IEEE Transactions on Image Processing*, 31: 2726–2738.

Lin, Y.-B.; Lei, J.; Bansal, M.; and Bertasius, G. 2022. Eclipse: Efficient long-range video retrieval using sight and sound. In *European Conference on Computer Vision*, 413–430. Springer.

Liu, X.; Li, L.; Wang, S.; Zha, Z.-J.; Li, Z.; Tian, Q.; and Huang, Q. 2022a. Entity-enhanced adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3003–3018.

Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; and Jin, Q. 2022b. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, 319–335. Springer.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Ma, Z.; Zhang, Z.; Chen, Y.; Qi, Z.; Yuan, C.; Li, B.; Luo, Y.; Li, X.; Qi, X.; Shan, Y.; et al. 2024. EA-VTR: Event-Aware Video-Text Retrieval. *arXiv preprint arXiv:2407.07478*.

Mafla, A.; Rezende, R. S.; Gomez, L.; Larlus, D.; and Karatzas, D. 2020. StacMR: Scene-Text Aware Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2220–2230.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; and Sivic, J. 2018. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31.

Shah, M.; Chen, X.; Rohrbach, M.; and Parikh, D. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6649–6658.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: a Dataset for Image Captioning with Reading Comprehension. *ArXiv*, abs/2003.12462.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Tang, H.; Cao, M.; Huang, J.; Liu, R.; Jin, P.; Li, G.; and Liang, X. 2024. Muse: Mamba is efficient multi-scale learner for text-video retrieval. *arXiv preprint arXiv:2408.10575*.

Tian, K.; Zhao, R.; Xin, Z.; Lan, B.; and Li, X. 2024. Holistic Features are almost Sufficient for Text-to-Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17138–17147.

Tom, G.; Mathew, M.; Garcia-Bordils, S.; Karatzas, D.; and Jawahar, C. 2023. Reading Between the Lanes: Text VideoQA on the Road. In *International Conference on Document Analysis and Recognition*, 137–154. Springer.

Tu, Y.; Li, L.; Su, L.; Zha, Z.-J.; and Huang, Q. 2024. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4926–4943.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J.; Sun, G.; Wang, P.; Liu, D.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 2024a. Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16551–16560.

Wang, J.; Wang, P.; Liu, D.; Guan, Q.; Dianat, S.; Rabbani, M.; Rao, R.; and Tao, Z. 2024b. Diffusion-inspired truncated sampler for text-video retrieval. *Advances in Neural Information Processing Systems*, 37: 3882–3906.

Wang, N.; Deng, J.; and Jia, M. 2024. Cycle-Consistency Learning for Captioning and Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5535–5543.

Wang, N.; Song, Y.; Ma, C.; Zhou, W.; Liu, W.; and Li, H. 2019. Unsupervised deep tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1308–1317.

Wang, Z.; Sung, Y.-L.; Cheng, F.; Bertasius, G.; and Bansal, M. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2816–2827.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Wu, W.; Zhao, Y.; Li, Z.; Li, J.; Zhou, H.; Shou, M. Z.; and Bai, X. 2025. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157: 110818.

Xue, H.; Sun, Y.; Liu, B.; Fu, J.; Song, R.; Li, H.; and Luo, J. 2022. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florêncio, D. A. F.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2020. TAP: Text-Aware Pre-training for Text-VQA and Text-Caption. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8747–8757.

Yin, J.; Li, L.; Zhang, J.; Gao, Y.; Yan, C.; and Sheng, X. 2025. Progressive Homeostatic and Plastic Prompt Tuning for Audio-Visual Multi-Task Incremental Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022–2033.

Zhang, B.; Li, L.; Wang, S.; Cai, S.; Zha, Z.-J.; Tian, Q.; and Huang, Q. 2024. Inductive state-relabeling adversarial active learning with heuristic clique rescaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Z.; Li, L.; Yan, C.; Liu, C.; van den Hengel, A.; and Qi, Y. 2025. Prosody-Enhanced Acoustic Pre-training and Acoustic-Disentangled Prosody Adapting for Movie Dubbing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 172–182.

Zhao, M.; Li, B.; Wang, J.; Li, W.; Zhou, W.; Zhang, L.; Xuyang, S.; Yu, Z.; Yu, X.; Li, G.; et al. 2022a. Towards video text visual question answering: Benchmark and baseline. *Advances in Neural Information Processing Systems*, 35: 35549–35562.

Zhao, S.; Zhu, L.; Wang, X.; and Yang, Y. 2022b. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 970–981.

Zhao, Z.; Li, L.; Zhang, J.; Sun, Y.; Sheng, X.; Yin, H.; and Jiang, S. 2025. Heterogeneous Prompt-Guided Entity Inferring and Distilling for Scene-Text Aware Cross-Modal Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10537–10545.

Zhou, S.; Xiao, J.; Li, Q.; Li, Y.; Yang, X.; Guo, D.; Wang, M.; Chua, T.-S.; and Yao, A. 2025. EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 3363–3373.

Zhou, S.; Xiao, J.; Yang, X.; Song, P.; Guo, D.; Yao, A.; Wang, M.; and Chua, T.-S. 2024. Scene-Text Grounding for Text-Based Video Question Answering. *arXiv preprint arXiv:2409.14319*.

Zhou, T.; Jae Lee, Y.; Yu, S. X.; and Efros, A. A. 2015. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q.; and Efros, A. A. 2016. Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 117–126.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.