MoVA: Learning Asymmetric Dual Projections for Modular Long Video-Text Alignment

Peiyuan Zhu[1], Shaoan Xie[1,2], Zijian Li[1,2], Yifan Shen[1], Namrata Deka[2], Harsh Shrivastava[1], Guangyi Chen[1,2], and Kun Zhang[1,2]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{peiyuan.zhu,zijian.li,yifan.shen}@mbzuai.ac.ae
{harsh.shrivastava,guangyi.chen}@mbzuai.ac.ae
[2] Carnegie Mellon University, Pittsburgh, PA, USA
shaoan@cmu.edu, ndeka@cs.cmu.edu, kunz1@cmu.edu

Abstract. Contrastive pre-training has propelled video-text alignment, yet models often inherit the critical limitations of their image-text predecessors like CLIP, resulting in entangled representations. These challenges are severely exacerbated by two fundamental properties in the video domain: Temporal Misalignment, where textual descriptions often correlate only to specific, constrained temporal windows, leaving other frames text-irrelevant; and Semantic Asymmetry, which dictates a sparse, bidirectional, and non-equivalent relevance between frame-level visual details and caption-level concepts. This failure persists whether captions are short and temporally disjoint, creating ambiguity, or long and detailed, fostering entanglement between static objects and their temporal evolution. In this paper, we establish theoretical conditions that enable flexible alignment between video and text representations across the temporal dimension and at varying levels of granularity. Building on these theoretical insights, we introduce MoVA—Modular Long Video–Text Alignment—which learns dual asymmetric projections: a text-side projection that adaptively selects frame-aware subspaces of the caption, and a video-side projection that disentangles text-relevant visual concepts. Our framework ensures that the model can preserve global cross-modal semantics while disentangling evolving, frame-specific concepts and scale naturally to long captions and videos. Empirical evaluations show that MoVA outperforms existing methods in multiple video-text alignment tasks, demonstrating the effectiveness of our method.

Keywords: Video-Text Retrieval · Multimodality · Representation Learning

1 Introduction

Developing generalizable video–text representations remains an important problem in modern computer vision. Driven by rapid advances in vision–language pretraining (VLP), large-scale cross-modal representation learning—coupled with

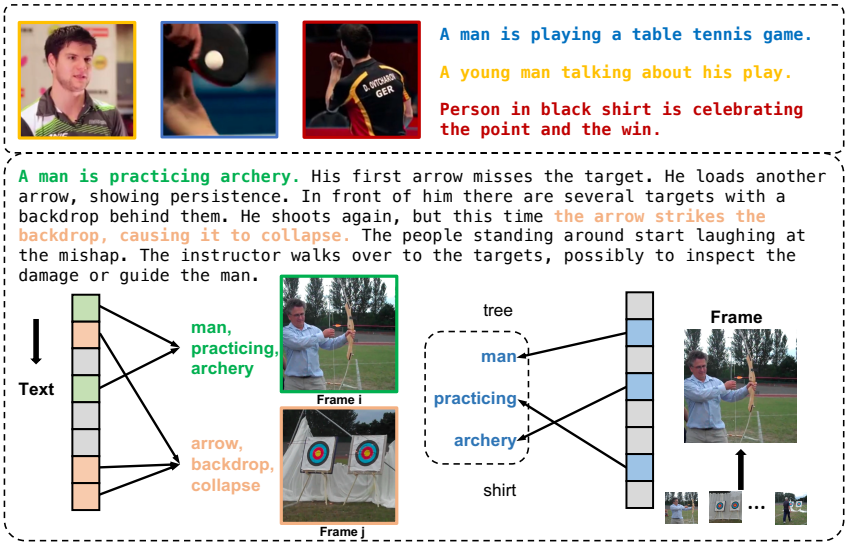

Fig. 1: Illustration of two key challenges for video–language alignment. (1) Temporal Misalignment: Video captions are inherently temporally misaligned with the underlying visual content. Multiple paired captions emphasize different moments along the video timeline—the first frame focuses on the man talking, while the second centers on the table tennis game. A single caption may describe an action occurring within only a short temporal window, leaving other frames potentially text-irrelevant. (2) Semantic Asymmetry: Regardless of caption length, only a sparse subset of the caption is relevant to any given frame. Text induces selective relevance, influencing each frame's attention toward different textual components, whereas each frame preserves richer yet underdetermined information toward the correlated text, forming an inherent bidirectional asymmetry between the two modalities.

explicit video–text alignment objectives—yields aligned embeddings that transfer broadly across a diverse spectrum of downstream tasks, including video-text retrieval [12,54,55,59,67], video captioning [21,43], and action recognition [11,73].

The success of CLIP-style contrastive learning [10, 35, 37] motivated many video–text systems to reuse image–text encoders and perform lightweight post-pretraining. However, CLIP itself can suffer from information misalignment in many image–text datasets and tends to learn entangled representations [17, 27, 28, 32, 45, 71]: different captions for the same image may highlight different concepts, and a single caption may involve multiple concepts. SmartCLIP [61] addresses this issue with adaptive masking and a modular contrastive objective at the image-level, encouraging disentangled representations of images and texts.

When extending to video, additional temporal misalignment arises because a video is a sequence of frames, risking the loss of salient frame-level semantics and misinterpretation of temporal dynamics. On the one hand, short captions for a single video may describe disjoint temporal segments, leaving the model uncertain about which parts of the text to attend to for a given frame. On the

other hand, directly aligning long captions with videos can preserve entangled details, preventing the learning of disentangled, atomic concepts at both the object and temporal levels, and ultimately limiting generalization on video tasks that demand fine-grained understanding across time and entities. We further characterize this as semantic asymmetry, shown in Figure 1.

To mitigate the impact of the above factors, we formulate video–text alignment as a latent-variable identification problem over temporally indexed frames and textual spans. We establish conditions under which frame–text correspondences are recoverable from weak clip–caption supervision, enabling our framework to preserve information as it evolves over time and to disentangle object-centric from motion-centric factors. Building on these theoretical insights, we introduce MoVA—a Modular Video–Text Alignment framework for CLIP-style video-text models. Operationally, MoVA decomposes the text representation into concept modules and learns a frame-aware mask selector that adaptively activates the relevant textual components per frame. The modular contrastive objective is computed at the frame level and aggregated across clips with lightweight temporal-consistency and coverage regularization. We empirically demonstrate that MoVA achieves competitive results across a range of downstream tasks without extra post-pretraining data, showcasing its effectiveness in addressing alignment challenges.

Our main contributions are summarized as follows.

- Theory for video–text misalignment and disentanglement. We formally characterize the challenges of information misalignment and semantic asymmetry in video–text alignment and cast the problem in a latent–variable framework and derive identification conditions that guarantee recovery of frame–text correspondences and concept-level factors.
- We propose dual asymmetric projections with modular contrastive learning procedures for video-text alignment, promoting disentangled, compositional representations.
- We conduct extensive experiments across diverse tasks, including long and short video-text retrieval, text-to-video generation, and concept visualization. MoVA consistently achieves competitive results, demonstrating its efficacy and validating our theoretical contributions.

2 Related Work

2.1 Video-Text Alignment

Modern video-text alignment research is heavily influenced by the wave of contrastive learning in the image-text domain. CLIP demonstrated the immense potential of natural language supervision for training robust visual models. This image-text pre-training paradigm was rapidly adapted to the video domain in an end-to-end manner, outperforming traditional non-CLIP methods [14, 36] and serving as effective video learners [38]. The pioneering work CLIP4Clip [30]

investigates various similarity calculation mechanisms and the effect of post-pretraining on video-language datasets, which spurs a significant volume of subsequent research into video-text alignment. Regarding encoder extraction [16, 50, 65] focus on building more precise and comprehensive encoders for video and text modalities. For interaction modeling between text and video modalities, works like [31, 46, 58, 67] learn various combinations of coarse- and fine-grained, as well as global and local representations. For instance, DGL [67] utilizes a shared latent space to generate dynamic local prompts and employs a global-local attention mechanism. For temporal sequence alignment, DTW (Dynamic Time Warping) [40] utilizes a symmetric distance definition and imposes strong temporal constraints to align the two sequences. VT-TWINS [25] aligns noisy video-text pairs via a differentiable DTW that handles weak correlations using local neighborhood smoothing. Other works, such as [22, 51], have focused on learning disentangled representations for video-text alignment. More recently, many approaches train larger models—even foundation models—on substantially richer video corpora, extending video–text alignment far beyond retrieval to domains such as video understanding [9, 49, 56, 57] and video generation [2, 48]. Long video–caption pairs are also becoming crucial: datasets like LVD-2M [62], VideoUFO [53], and UltraVideo [66] offer longer, more detailed, and richer video-caption pairs. Long captions are critical as they tend to retain entangled details [61, 72], and original videos with sufficient change and clear content help expand downstream applications of video–text alignment beyond the confines of retrieval. Our method focuses on addressing the temporal misalignment and semantic asymmetry inherent in video-text pairs through mask modeling to learn temporally interpretable concepts for long video–text alignment.

2.2 Latent Variable Identification

Latent variable identification is the process of statistically inferring and uniquely recovering unobservable, hidden variables or constructs from a set of observable, measured data. Identifying the latent structure that mediates video–text correspondence is a principled route to robust alignment and generalization. A large body of work shows that, under suitable auxiliary signals or structural conditions, semantic factors become recoverable from observations. Recent research further demonstrates that latent causal variables become identifiable once additional structure is imposed on the learning problem. Temporal regularities—modeling how latent factors evolve causally over time or exhibit temporal sparsity—supply such structure and can break indeterminacies [20, 24, 69, 70]. Complementary advances constrain the generative mechanism itself: sparsity and related structural priors shrink equivalence classes and enable recovery of the underlying factors [63, 74, 75]. Cross-modal alignment also provides anchors: paired or multi-view observations allow contrastive or grouping-based objectives to tie representations across modalities and render the latent structure identifiable even under partial observability [6, 15, 18, 20, 33, 34, 44, 68]. Under appropriate conditions, the identification of disentangled and manipulable latents enables controllable image and video generation [23, 42, 60]. For video–text alignment, the

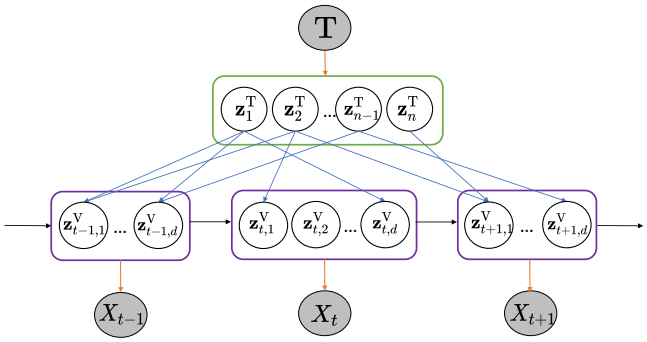

Fig. 2: The **data-generating process**. The video-text pair originates from its corresponding representation pair. The mapping from the text representation $\mathbf{z}^{\mathrm{T}}$ to the sequence of frame representations $\mathbf{z}^{\mathrm{V}}$ over time is sparse. The subset of text representations associated with a particular frame representation contains only partial information about the vision representation.

objective is to learn high-level semantics from low-level observational video–text pairs. These paired data can be treated as multi-view observations that enable identification of information shared across views. Previous work [15, 47] adopts flexible assumptions on the underlying distribution to identify blocks of latent variables directly shared by two views induced by data augmentations, and [68] extends this to the multi-view setting. However, these approaches face persistent grouping ambiguity (e.g., whether captions from different videos belong to the same group). SmartCLIP [61] approaches alignment from the image–text side, where the grouping of multiple captions—short or long—is coherently determined by preserved cross-modal information, allowing the image to unilaterally select aspects of the textual description. In contrast, video–text pairs are inherently bidirectional: the text abstracts transformations across frames and can even determine which frames are unimportant, rendering a simple one-to-one image-to-text mapping untenable. In our theoretical analysis, we show that by properly leveraging the data-generating process, we can obtain the desired identification results for video–text alignment.

3 Problem Formulation

In this section, we formalize the data-generating process underlying video–text alignment as the basis for subsequent theoretical analysis.

Notations. We indicate the dimensionality of a vector with $d(\cdot)$ and index a subset of its components by $[\mathbf{z}]_{\mathcal{B}}$ for an index set $\mathcal{B}$. For any mask vector $\mathbf{m}$, we write the support (nonzero indices) as $\mathcal{B}(\mathbf{m}) := \{i \in [d(\mathbf{m})] : [\mathbf{m}]_i \neq 0\}$. We denote the element-wise (Hadamard) product by $\odot$.

Data-generating process. The data-generating process is shown in Figure 2 and specified in (1)-(2). A video is $\mathbf{V} := (\mathbf{X}_1, \ldots, \mathbf{X}_T)$ with frames $\mathbf{X}_t \in$

$\mathcal{X} \subset \mathbb{R}^{d(\mathbf{X}_t)}$ and a paired caption $\mathbf{T} \in \mathcal{T} \subset \mathbb{R}^{d(\mathbf{T})}$. Both modalities live in a *shared* latent space $\mathcal{Z} \subset \mathbb{R}^{d(\mathbf{z})}$: a text latent $\mathbf{z}^{\mathrm{T}} \in \mathcal{Z}$ and per-frame latents $\mathbf{z}_t^{\mathrm{V}} \in \mathcal{Z}$ for $t = 1, \ldots, T$. Since each word can be represented by a continuous word embedding vector in practice [3], we model the text caption as continuous variables. We assume that each video-text pair $(\mathbf{V}, \mathbf{T})$ originates from semantic latents together with modality-specific nuisance variables through modality-specific generators: the video generator g^{V} and the text generator g^{T}. Concretely, frames are generated by $g^{\mathrm{V}} : (\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}}) \mapsto \mathbf{X}_t$ and the caption is generated by $g^{\mathrm{T}} : (\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}}) \mapsto \mathbf{T}$. To capture temporal locality and semantic asymmetry, we introduce *dual sparse asymmetric* projections:

i Text→frame mask $\mathbf{m}_t^{\mathrm{T}} \in \mathcal{M}^{\mathrm{T}} \subset \{0,1\}^{d(\mathbf{z})}$, which selects a subset of text concepts relevant to frame t (global text → local frame).

ii Frame→text mask $\mathbf{m}_t^{\mathrm{V}} \in \mathcal{M}^{\mathrm{V}} \subset \{0,1\}^{d(\mathbf{z})}$, which selects the subset of frame-level visual factors accountable to the subset of text concepts (local frame → local text).

We assume modality-specific generators with nuisance variations $\boldsymbol{\epsilon}_t^{\mathrm{V}}$ (e.g., motion blur, illumination) and $\boldsymbol{\epsilon}^{\mathrm{T}}$ (e.g., syntax, tense), and encode frame-aware dual selection as a per-frame semantic-consistency constraint, as shown in (2).

$$\mathbf{X}_t := g^{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}}), \qquad \mathbf{T} := g^{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}}), \tag{1}$$

$$\mathbf{m}_t^{\mathrm{T}} \odot \mathbf{z}^{\mathrm{T}} = \mathbf{z}_t^{\mathrm{V}} \odot \mathbf{m}_t^{\mathrm{V}}, \quad t = 1, \ldots, T. \tag{2}$$

Goal. Our goals can be formalized as follows.

a *Preserve global cross-modal semantics over time*: recover the complete caption-level latent $\mathbf{z}^{\mathrm{T}}$ and maintain its consistency with the span $\mathbf{z}_t^{\mathrm{V}} \,|_{t=1}^{T}$.

b *Disentangle frame-specific concepts at multiple granularities*: identify and separate temporally localized factors within $\mathbf{z}_t^{\mathrm{V}}$ that correspond to sparse, caption-conditioned subspaces selected by $(\mathbf{m}_t^{\mathrm{T}}, \mathbf{m}_t^{\mathrm{V}})$, even when such atomic factors are unseen during training.

Examples. As illustrated in Figure 2, the global caption contains two events: "A man is practicing archery." and "The arrow strikes the backdrop, causing it to collapse." Consider two frames at different time steps, indexed by i and j ($i < j$ in this case). From the text→frame perspective, for the frame $\mathbf{X}_i$, the text→frame mask $\mathbf{m}_i^{\mathrm{T}}$ activates $\{\text{man}, \text{practicing}, \text{archery}\}$; for the later frame $\mathbf{X}_j$, $\mathbf{m}_j^{\mathrm{T}}$ activates $\{\text{arrow}, \text{backdrop}, \text{collapse}\}$. From the frame→text perspective, each frame may include additional visual content beyond what is explicitly mentioned (e.g., trees, shirt for $\mathbf{X}_i$). Accordingly, $\mathbf{m}_i^{\mathrm{V}}$ selects the subset of visual factors in $\mathbf{X}_i$ that correspond to the text subset chosen at time i (e.g., archer's posture, bow, arrow motion rather than background trees), while $\mathbf{m}_j^{\mathrm{V}}$ prioritizes the arrow–backdrop contact and collapse dynamics at time j.

4 Identification Theory

We establish the theoretical guarantees that motivate the modular design in Section 5. We show that a suitable objective recovers the latent semantics shared

by video and text up to block-wise equivalence, even though captions provide only weak, temporally sparse supervision.

Definition 4.1 (Block-wise Identifiability). *A latent vector* $\mathbf{v}$ *is block-wise identifiable if it is related to its estimate* $\hat{\mathbf{v}}$ *through an invertible map on every block selected by the associated mask* $\mathbf{m}$.

Learning objective. Let $\hat{\mathbf{z}}_t^{\mathrm{V}} = \hat{g}^{\mathrm{V}}(\mathbf{X}_t)$ and $\hat{\mathbf{z}}^{\mathrm{T}} = \hat{g}^{\mathrm{T}}(\mathbf{T})$ denote the encoder outputs defined in Section 5. We estimate frame-wise masks $\hat{\mathbf{m}}_t^{\mathrm{V}}, \hat{\mathbf{m}}_t^{\mathrm{T}} \in \mathcal{M}$ by solving

$$\begin{aligned} \min_{\hat{g}^{\mathrm{V}}, \hat{g}^{\mathrm{T}}, \{\hat{\mathbf{m}}_t^{\mathrm{V}}, \hat{\mathbf{m}}_t^{\mathrm{T}}\}} \quad & \sum_t \left(\|\hat{\mathbf{m}}_t^{\mathrm{V}}\|_0 + \|\hat{\mathbf{m}}_t^{\mathrm{T}}\|_0 \right) \\ \text{s.t.} \quad & \sum_t \left\| \hat{\mathbf{z}}_t^{\mathrm{V}} \odot \hat{\mathbf{m}}_t^{\mathrm{V}} - \hat{\mathbf{z}}^{\mathrm{T}} \odot \hat{\mathbf{m}}_t^{\mathrm{T}} \right\|^2 < \eta. \end{aligned} \tag{3}$$

The constraint captures the limiting form of the modular contrastive loss in Section 5, forcing the dual-mask relation between $\hat{\mathbf{m}}_t^{\mathrm{T}} \odot \hat{\mathbf{z}}^{\mathrm{T}}$ and $\hat{\mathbf{z}}_t^{\mathrm{V}} \odot \hat{\mathbf{m}}_t^{\mathrm{V}}$ to be close for every frame. The sparsity objective encourages minimal supports on both sides so that only frame-relevant coordinates remain active.

Condition 4.2 (Identification Conditions).

(i) Smoothness & invertibility. The generators $(g^{\mathrm{V}}, g^{\mathrm{T}})$ *in Section 5 are smooth and admit smooth inverses, so no semantic information is lost in* $(\mathbf{X}_t, \mathbf{T})$.

(ii) Full joint support. Every pair $(\mathbf{z}, \mathbf{m})$ *that satisfies the dual-mask constraint in Section 5 occurs with positive density, ensuring that concepts are observed under all admissible temporal spans.*

(iii) Temporal stability. Each video admits segments $\{\mathcal{S}_k\}$ *such that the true text-side mask is constant on* $\mathcal{S}_k$ *while visual masks may vary as long as the dual-mask constraint holds.*

(iv) View diversity. Conditioned on $(\mathbf{z}, \mathbf{m})$*, the nuisance variables* $(\boldsymbol{\epsilon}_t^{\mathrm{V}}, \boldsymbol{\epsilon}^{\mathrm{T}})$ *vary in a neighborhood, yielding multiple conditionally independent realizations of every semantic block.*

Theorem 4.3 (Identification Theorem). *Assume the data-generating process in Section 3 and let* $(\hat{g}^{\mathrm{V}}, \hat{g}^{\mathrm{T}}, \{\hat{\mathbf{m}}_t^{\mathrm{V}}, \hat{\mathbf{m}}_t^{\mathrm{T}}\})$ *be an optimum of* (3)*. Under Condition 4.2, the true representation block* $[\mathbf{z}]_{\tilde{\mathcal{B}}}$ *is block-wise identifiable for any index set* $\tilde{\mathcal{B}}$ *that can be written as either* $\cup_{\mathbf{m}\in\mathcal{V}} \mathcal{B}(\mathbf{m})$ *or* $\cap_{\mathbf{m}\in\mathcal{V}} \mathcal{B}(\mathbf{m})$ *over any subset of masks* $\mathcal{V} \subset \mathcal{M}$.

Concept preservation. Theorem 4.3 ensures that the concept block associated with a temporal span $\mathbf{m}$ is retained in the learned representation. Hence, even when a caption highlights only a short portion of a long video, MoVA preserves the frame-level semantics selected by $\mathbf{m}$, preventing unrelated frames from overwriting them.

Concept disentanglement. The intersection operation in Theorem 4.3 allows us to recover atomic concepts that recur across temporally disjoint captions.

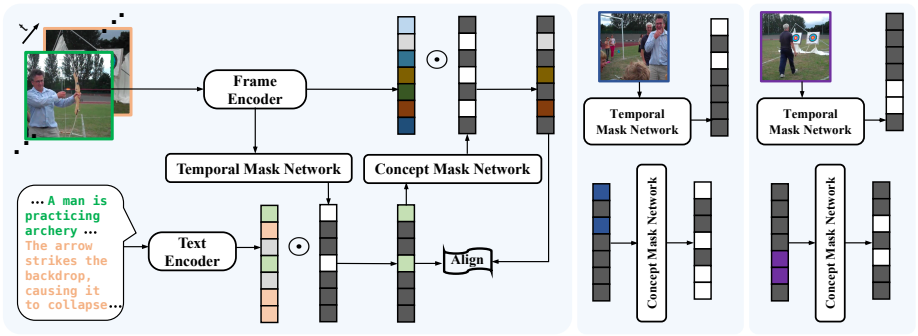

Fig. 3: MoVA overview. We integrate dual asymmetric projections: the Temporal Mask Network selects which subset of the global text representation to be used, while the Concept Mask Network selects the parts of the frame representation that align with the selected, correlated text.

For example, repeated mentions of the same actor or action can be isolated by intersecting the corresponding masks, which explains the compositional behavior observed in Section 6.

Theoretical contribution. Theorem 4.3 extends the multi-view identification frameworks of [47, 68] to temporally indexed video–text data without requiring explicit segment labels. Whereas prior work presumes the view group of each caption is known, our objective infers the grouping automatically through the dual-mask constraint, providing the first temporal identification guarantee for CLIP-style video–text models. For the proof of the identification theory, please refer to the supplementary material.

5 MoVA: Modular Video-Text Alignment

In this section, we present our empirical approach to modular video-text alignment based on the identification theory in Section 4, detailing the asymmetric dual projections, model architecture and training objectives.

Text-to-Video Projection Learning. From the global text representation to each local frame, the text-to-video projection aims to account for frames' differing focal points on the text, pairing every frame with the portion of the text most relevant to it. $\mathbf{m}^{\mathrm{T}}(\cdot)$ denotes the masking projection of $\mathbf{X}_t$. The Temporal Mask Network (TMN) consists of a two-layer Transformer block. It takes the frame and its learnable positional embedding as the query, allows every token to interact with that frame and with all other tokens in the sequence, and outputs a binary vector $\mathbf{m}^{\mathrm{T}}(\hat{\mathbf{z}}_t^{\mathrm{V}})$ via a straight-through estimator [4], i.e., the subset of text most associated with that frame. Starting from each frame, we (i) pull the frame's visual embedding $\hat{\mathbf{z}}_t^{\mathrm{V}}$ closer to its corresponding masked text embedding $\hat{\mathbf{z}}_t^{\mathrm{T}} \odot \hat{\mathbf{m}}_t^{\mathrm{T}}$ and (ii) enforce that for the current frame t, its similarity to its own masked text exceeds its similarity to the masked texts of other frames (e.g.,

$t-1$). Frames farther away in the current video (beyond the temporal window) and frames from other video samples are regarded as harder examples. We define the contrastive frame-level loss as follows:

$$\begin{aligned}\ell_{\text{ctrf}}(t) = &\left[1 - \langle \hat{\mathbf{z}}_t^{\text{V}}, \hat{\mathbf{z}}_t^{\text{T}} \odot \hat{\mathbf{m}}_t^{\text{T}} \rangle\right] + \\ &\left[\max(0, \Delta - \langle \hat{\mathbf{z}}_t^{\text{V}}, \hat{\mathbf{z}}_t^{\text{T}} \odot \hat{\mathbf{m}}_t^{\text{T}} \rangle + \langle \hat{\mathbf{z}}_t^{\text{V}}, \hat{\mathbf{z}}_k^{\text{T}} \odot \hat{\mathbf{m}}_t^{\text{T}} \rangle)\right],\end{aligned} \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity; Δ is the contrastive margin; and $\hat{\mathbf{z}}_k^{\text{T}}$ denotes the k-th negative sample. For brevity, we omit the explicit sample index i in (4). The aggregate loss is given by:

$$\mathcal{L}_{\text{align}}^{\text{t}\to\text{v}} = \frac{1}{L} \sum_{i=1}^{N} \sum_{t=1}^{L_i} \ell_{\text{ctrf}}(t), \tag{5}$$

where L_i denotes the valid frames per sample, $L = \sum_{i=1}^{N} \sum_{t=1}^{L_i}$ This is consistent with our goal of addressing temporal misalignment and enforces alignment between the global text and the local frames.

Video-to-Text Projection Learning. Each frame in the video preserves, in its entirety, the cross-modal semantic information contained in the temporally combined text subset obtained from text-to-video projection learning. We shift the control from text to frames: each frame selects the portion most related to its own visual information and aligns it with the correlated text subset. Following [61], we introduce temporal modeling and extend it to video–text alignment. We build the Concept Mask Network (CMN), $\hat{\mathbf{m}}_t^{\text{V}}(\cdot)$, which learns a masking projection of text subset $\hat{\mathbf{z}}_t^{\text{T}} \odot \hat{\mathbf{m}}_t^{\text{T}}$, which we denote as $\hat{\mathbf{s}}^{\text{T}}$ for brevity. It consists of a single-layer Transformer block and an attention-pooling layer that adaptively down-samples the output to match the dimensionality of the CLIP representation.

To disentangle frame-specific factors and ensure identifiability, we learn the projection with two modular contrastive terms. Same-Frame Different-Mask (sfdm) fixes a frame representation $\hat{\mathbf{z}}_t^{\text{V}}$ (encoder output, normalized) and applies caption-conditioned visual-dimension masks produced from different captions; letting $\hat{\mathbf{m}}_{j,t}$ denote the selector induced by caption j fot t-th frame and using temperature τ,

$$\mathcal{L}_{\text{sfdm}} = -\frac{1}{L} \sum_{i=1}^{N} \sum_{t=1}^{L_i} \log \frac{\exp\Big(\langle \hat{\mathbf{m}}_{i,t}^{\text{V}} \odot \hat{\mathbf{z}}_{i,t}^{\text{V}},\, \hat{\mathbf{s}}_i^{\text{T}} \rangle / \tau\Big)}{\sum_{j=1}^{N'} \exp\Big(\langle \hat{\mathbf{m}}_{j,t}^{\text{V}} \odot \hat{\mathbf{z}}_{i,t}^{\text{V}},\, \hat{\mathbf{s}}_j^{\text{T}} \rangle / \tau\Big)}\,. \tag{6}$$

Different-Frame Same-Mask (dfsm) contrasts the text subset $\hat{\mathbf{s}}^{\text{T}}$ in the positive pair with randomly sampled frame representations (from both intra-video and inter-video):

$$\mathcal{L}_{\text{dfsm}} = -\frac{1}{L} \sum_{i=1}^{N} \sum_{t=1}^{L_i} \log \frac{\exp\Big(\langle \hat{\mathbf{m}}_{i,t}^{\text{V}} \odot \hat{\mathbf{z}}_{i,t}^{\text{V}},\, \hat{\mathbf{s}}_i^{\text{T}} \rangle / \tau\Big)}{\sum_{j=1}^{N'} \exp\Big(\langle \hat{\mathbf{m}}_{i,t}^{\text{V}} \odot \hat{\mathbf{z}}_{j,t}^{\text{V}},\, \hat{\mathbf{s}}_i^{\text{T}} \rangle / \tau\Big)}\,. \tag{7}$$

Table 1: Retrieval performance on ActivityNet. "↑" denotes that higher is better. "↓" denotes that lower is better. **Bold** and underlined values denote the best and second-best results, respectively. † denotes that the method uses DSL [13] as post-processing operations.

Method	Text → Video					Video → Text				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip [30]	43.8	74.9	86.6	2.0	6.4	42.9	75.3	86.6	2.0	6.4
DRL [51]	44.2	73.9	84.1	2.0	7.9	42.7	73.8	84.5	2.0	7.7
X-CLIP [31]	42.9	73.7	84.7	2.0	7.4	42.2	74.6	85.5	2.0	6.9
ProST [29]	44.5	72.5	83.6	2.0	8.5	43.7	73.9	84.7	2.0	7.0
InternVideo [56]	42.2	73.1	84.7	2.0	7.7	41.6	73.8	85.0	2.0	7.1
DiCoSA [22]	43.7	73.8	84.4	2.0	8.0	40.6	71.8	83.8	2.0	7.7
DGL [67]	43.5	74.2	84.8	2.0	8.2	43.4	73.6	85.1	2.0	7.9
VideoCLIP-XL [49]	46.9	75.1	86.3	2.0	6.6	37.7	68.2	81.1	2.0	10.1
MoVA (Ours)	47.8	77.4	87.0	2.0	7.0	46.7	76.6	86.8	2.0	6.3
MoVA† (Ours)	**53.6**	**79.1**	**88.1**	**1.0**	**6.3**	**54.4**	**79.9**	**88.6**	**1.0**	**5.7**

Thus, we have the alignment objective from frames to text:

$$\mathcal{L}_{\text{align}}^{\text{v}\to\text{t}} = \mathcal{L}_{\text{sfdm}} + \mathcal{L}_{\text{dfsm}} \tag{8}$$

MoVA training objective. We utilize a global symmetric retrieval loss $\mathcal{L}_{\text{g}}$ following [30] as grounding, where the similarity matrix is given by cosine similarities between global video and text representations. We impose sparsity constraints for both $\hat{\mathbf{m}}^{\text{T}}$ and $\hat{\mathbf{m}}^{\text{V}}$, encouraging textual and visual concepts to be encoded in a minimal set of latent dimensions, thereby facilitating the disentanglement of distinct concepts. The training objective of MoVA is a weighted sum of aforementioned loss terms:

$$\mathcal{L} = \lambda_{\text{g}}\mathcal{L}_{\text{g}} + \lambda_{\text{tv}}\mathcal{L}_{\text{align}}^{\text{t}\to\text{v}} + \lambda_{\text{vt}}\mathcal{L}_{\text{align}}^{\text{v}\to\text{t}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} \tag{9}$$

6 Experiments

6.1 Setup

Datasets. MoVA is evaluated on both long and short video–text benchmarks. Classic video-text retrieval datasets ActivityNet [26], MSVD [5] (YouTube2Text), and DiDeMo [1] are utilized for evaluation, following the standard train/val/test splits [30]. To test scalability to longer descriptions, we adopt the recently released VideoUFO [53] and UltraVideo [66].

Implementation details. To facilitate the transfer of sparse mapping information from image-text alignment to video-text alignment, we first train on the image-caption dataset ShareGPT4v [8] following [61], and use the resulting weights to initialize our text encoder, frame encoder, and Concept Mask

Network. We adopt a positional encoding capable of handling 248 tokens, overcoming the 77-token limit in the original CLIP. For mask modeling, we apply sigmoid to restrict the output to the range (0, 1) and employ straight through estimation (STE) [4] to binarize the outputs. The initial learning rate for text encoder and frame encoder is 10^{-7}, and the initial learning rate for other modules is 10^{-4}. Unless stated otherwise, all methods use ViT-B/16 initialization during training with batch size 256. For VideoCLIP-XL [49], we evaluate the publicly released checkpoint, which was trained with ViT-L/14 initialization. We set the max token length, max frame length, and number of training epochs to 64, 64, and 20 for ActivityNet, DiDeMo, VideoUFO, and UltraVideo, and to 32, 12, and 3 for MSVD; for VideoCLIP-XL and our method, the maximum token length is 248. DSL [13] post-processing is only applied when explicitly noted.

Metrics. We use retrieval metrics that capture both precision and ranking stability. Recall at K (R@1/5/10) quantifies whether the correct item appears within the top-K retrieved results, while Median Rank (MdR) and Mean Rank (MnR) diagnose the heavy-tail behavior introduced by long ambiguous descriptions. All metrics are reported for both text→video and video→text scenarios following prior work [30, 31].

6.2 Results

Video-text retrieval. In this work, we focus on methods that adapt image-pretrained CLIP models to video without large-scale video post-pretraining corpora. Accordingly, our controlled comparisons emphasize approaches whose initialization and supervision mainly come from image-based CLIP pretraining. Methods that leverage extra post-pretraining video data, e.g., CLIP-ViP trained with HD-VILA-100M [64,65] and VidLA trained with YT-VidLA-800M [39], are outside the main controlled setting. We include VideoCLIP-XL [49] as a reference for long-caption video-text retrieval. We compare our model with several state-of-the-art works on the video-text retrieval task. Our model achieves the best results on all datasets even without post-processing like DSL, as shown in Tables 1 and 2. We find that for nearly all methods, R@1 in text-to-video retrieval is higher than in video-to-text, which is the opposite of the pattern in image–text retrieval (where image-to-text R@1 typically exceeds text-to-image). This highlights the greater complexity of video–text alignment relative to the near one-to-one mapping in image–text alignment, underscores the guiding role of the global text representation in video–text alignment, and shows the benefit of leveraging it to help disentangle the representations. The gains are consistent across all caption–video length regimes—from short-caption/short-video (e.g., MSVD) to long-caption/long-video (e.g., VideoUFO)—demonstrating strong bidirectional vision–language correspondence and favorable scaling behavior.

Long-text-to-video generation. Figure 5 highlights MoVA's ability to keep long textual descriptions intact when serving as the language interface for text-to-video generators such as VideoCrafter2 [7]. Compared with CLIP and SmartCLIP—which focus exclusively on image–text alignment—our method generates

Fig. 4: A test example on ActivityNet illustrating the most relevant text subsets and weights assigned by the Temporal Mask Network to each frame (top-3 only).

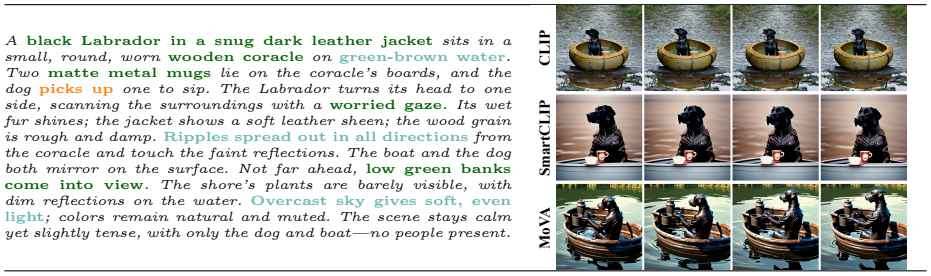

Fig. 5: Example of Long-text-to-video generation. We replace the CLIP text encoder in VideoCrafter2 [7]. Compared with existing text encoders trained through image–text alignment when used for text-to-video generation, our method MoVA comprehensively captures modular temporal-level and object-level concepts. MoVA can generate details such as the *matte metal mug* and the *pick up* motion.

videos that are more vivid and more faithfully aligned with the text. TMN dynamically selects span-level concepts and CMN aligns them with temporally evolving frames, which allows the generator to render subtle objects (e.g., the *matte metal mug*) and actions (the dog *picking up* the mug) that would otherwise vanish. For quantitative evaluation, we use VBench [19] to benchmark VideoCrafter2 in a zero-shot setting by replacing only its CLIP text encoder, and evaluate on 2,048 prompts randomly sampled from VidProM [52]. The results are reported in Table 3.

Temporal mask analysis. Figure 4 visualizes the top-3 textual spans selected by TMN across an ActivityNet video. It shows our model can select the aligned portion of the global text for various frames. This provides a perspective on addressing temporal misalignment, and it avoids using identical masks for adjacent frames, which reflects the role of the per-frame contrastive loss in Eq. (4).

Concept-level visualization. Concept GradCAMs in Figure 6 reveal that CMN isolates object-centric and motion-centric bases: when the caption em-

Table 2: Text-to-video retrieval performance on various datasets. "↑" denotes that higher is better. "↓" denotes that lower is better.

Method	MSVD					DiDeMo				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip [30]	47.4	77.8	85.6	2.0	10.3	44.8	73.4	81.6	2.0	13.5
DRL [51]	49.8	81.2	89.5	2.0	9.5	49.0	76.5	84.5	2.0	12.0
X-CLIP [31]	50.4	80.6	89.8	**1.0**	8.4	47.8	79.4	82.3	2.0	12.5
ProST [29]	46.4	74.4	83.8	2.0	12.1	47.5	75.2	84.6	2.0	12.3
InternVideo [56]	44.2	74.5	84.1	2.0	10.9	50.8	78.8	86.6	**1.0**	6.9
VideoCLIP-XL [49]	48.6	81.0	86.6	**1.0**	10.2	38.6	63.6	73.0	3.0	49.2
MoVA (Ours)	**52.6**	**83.0**	**90.6**	**1.0**	**7.8**	**57.5**	**83.2**	**91.6**	**1.0**	**4.8**
Method	**VideoUFO**					**UltraVideo**				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CLIP4Clip [30]	34.5	62.1	72.4	3.0	45.1	43.3	74.6	84.9	2.0	9.2
DRL [51]	24.2	45.7	54.9	7.0	227.5	35.9	65.9	76.5	3.0	16.6
X-CLIP [31]	37.7	65.6	75.6	2.0	35.4	41.9	73.5	84.3	2.0	7.7
ProST [29]	48.2	74.5	82.7	2.0	28.0	51.8	83.5	90.3	**1.0**	5.7
InternVideo [56]	42.8	70.2	76.4	2.0	33.6	35.4	65.9	76.3	3.0	14.0
VideoCLIP-XL [49]	57.4	82.2	88.4	**1.0**	23.6	42.6	71.8	82.2	2.0	8.4
MoVA (Ours)	**62.4**	**87.0**	**92.3**	**1.0**	**6.9**	**58.5**	**87.8**	**94.4**	**1.0**	**3.4**

Table 3: Quantitative evaluation results on VBench.

Method	Subject consistency	Background consistency	Motion smoothness	Aesthetic quality	Imaging quality	Avg.
CLIP	0.9700	0.9690	0.9799	0.6271	0.6822	0.8456
SmartCLIP	0.9671	0.9722	**0.9856**	0.7064	0.6980	0.8659
MoVA (Ours)	**0.9776**	**0.9766**	0.9849	**0.7126**	**0.7023**	**0.8708**

phasizes "some pigs," activations gather around the animals despite distracting background motion. This supports our claim that semantic asymmetry must be modeled bidirectionally—text modules decide which frames to attend to, and frame-conditioned visual masks decide which text concepts remain active. Together with the retrieval gains, these qualitative trends provide high-level evidence that MoVA preserves global semantics while remaining compositional.

6.3 Ablation Study

Scalability. Panels (a) and (b) of Figure 7 compare video-text retrieval when training on ActivityNet and UltraVideo. MoVA exhibits steadily improving retrieval performance as training epochs increase. Panel (c) further shows consistent improvements when swapping ViT-B/16, ViT-L/14, and ViT-H/14 backbones [37], indicating that TMN/CMN act as architecture-agnostic plugs rather than overfitting to a specific capacity regime. Collectively, these curves demonstrate that the dual asymmetric projections scale gracefully with both data vol-

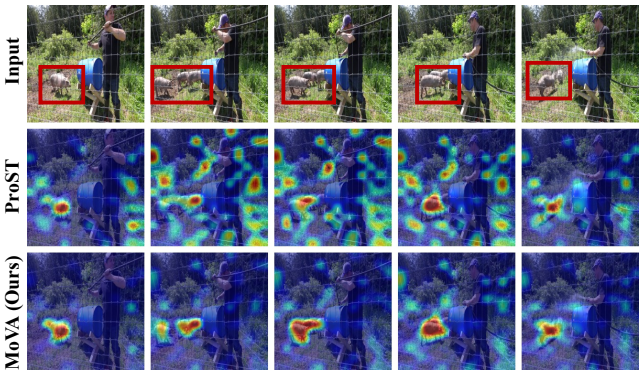

Fig. 6: Visualization of learned concepts. We perform representation visualization by formulating a proxy classification task. The cosine similarity between frame embeddings across time and a generated caption (e.g., "some pigs") serves as the classification score for GradCAM [41] attribution. Red boxes mark the temporally active regions.

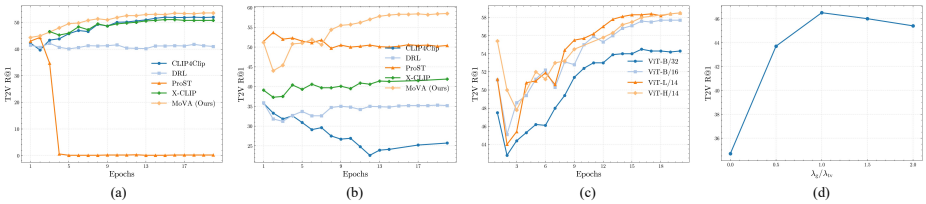

Fig. 7: (a) Scaling with epochs on ActivityNet; (b) Scaling with epochs on UltraVideo; (c) Scaling trends across ViT backbones; (d) Ablation on loss-coefficient ratio λ_g/λ_{tv}.

ume and model size, which is essential for the long-form scenarios highlighted in the Introduction.

Loss weighting. Figure 7(d) studies the interaction between the global retrieval loss $\mathcal{L}_{\text{g}}$ and the modular alignment loss $\mathcal{L}_{\text{align}}^{\text{t}\to\text{v}}, \mathcal{L}_{\text{align}}^{\text{v}\to\text{t}}$. We fix $\lambda_{\text{tv}} = \lambda_{\text{vt}} = 0.5$ and sweep λ_{g}, plotting performance against $\lambda_{\text{g}}/\lambda_{\text{tv}}$. Extremely small ratios (e.g., 0.2) under-emphasize the global constraint and slightly hurt MnR, whereas overly large ratios (> 1.2) collapse the per-frame masks, validating the sparsity/coverage trade-off. The sweet spot around 1.0 balances whole-video grounding with frame-level disentanglement.

Effectiveness of dual asymmetric projections. Table 4 summarizes three intermediate variants derived from the SmartCLIP initialization. (i) Fine-tuning with only the global retrieval loss $\mathcal{L}_{\text{g}}$ removes all modular objectives, corresponding to the image-text set-

Table 4: Ablation studies on the ActivityNet dataset. "↑" denotes that higher is better. "↓" denotes that lower is better.

Ablation Items	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
i	43.7	73.5	84.0	7.9	42.5	72.9	84.2	7.7
ii	35.6	65.1	77.6	11.7	33.7	62.4	75.8	11.6
iii	42.0	70.0	81.5	9.7	41.9	70.8	80.8	9.0
iv	**47.6**	**77.4**	**87.0**	**7.0**	**46.7**	**76.6**	**86.8**	**6.3**

ting and yielding the weakest R@1. (ii) We retain SmartCLIP's single-direction visual→text masking by learning a 3D mask from video frames to text tokens; this ignores the text-side global guidance emphasized in Section 1 and is unable to resolve temporal misalignment, leading to the sharpest degradation (R@1 35.6). (iii) We keep our TMN but apply temporal masks directly on raw text tokens rather than on the global text representation. (iv) The full MoVA stacks both TMN and CMN so that text can guide frame selection while frames can reweight textual concepts. These targeted ablations corroborate that learning dual asymmetric projections is effective for video-text alignment.

7 Conclusion

In this paper, we address temporal misalignment and semantic asymmetry in video–text alignment by formulating identification conditions that explicitly connect textual descriptions to their atomic visual counterparts. Building on these insights, MoVA introduces dual asymmetric projections that maintain global semantics while isolating frame-specific concepts, leading to disentangled and compositional cross-modal representations. Quantitative evaluations and qualitative studies jointly validate the theoretical claims, confirming that principled structure can translate into practical gains for multimodal video-text alignment. Looking ahead, MoVA will be integrated with next-generation multimodal foundation models to better handle long-horizon narratives, event compositionality, and fine-grained temporal grounding.

Acknowledgments

We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, MBZUAI-WIS Joint Program, and the Al Deira Causal Education project.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)
4. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)

5. 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)
6. Chen, G., Deng, Y., Zhu, P., Li, Y., Shen, Y., Li, Z., Zhang, K.: Causalverse: Benchmarking causal representation learning with configurable high-fidelity simulations. Advances in Neural Information Processing Systems **38** (2026)
7. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024). https://doi.org/10.1109/CVPR52733.2024.00698
8. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. In: European Conference on Computer Vision. pp. 370–387. Springer (2024). https://doi.org/10.1007/978-3-031-72643-9_22
9. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., Liu, J.: Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. Advances in Neural Information Processing Systems **36**, 72842–72866 (2023)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
11. Chen, Y., Chen, D., Liu, R., Zhou, S., Xue, W., Peng, W.: Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18688–18698 (2024)
12. Chen, Y., Wang, J., Lin, L., Qi, Z., Ma, J., Shan, Y.: Tagging before alignment: Integrating multi-modal tags for video-text retrieval. Proceedings of the AAAI Conference on Artificial Intelligence **37**(1), 396–404 (2023). https://doi.org/10.1609/aaai.v37i1.25113
13. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)
14. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11583–11593 (2021)
15. Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., Vogt, J.E.: Identifiability results for multimodal contrastive learning. arXiv preprint arXiv:2303.09166 (2023)
16. Deng, C., Chen, Q., Qin, P., Chen, D., Wu, Q.: Prompt switch: Efficient clip adaptation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15648–15658 (2023)
17. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. Advances in Neural Information Processing Systems **36**, 35544–35575 (2023)
18. Gresele, L., Rubenstein, P.K., Mehrjou, A., Locatello, F., Schölkopf, B.: The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In: Uncertainty in Artificial Intelligence. pp. 217–227. PMLR (2020)
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)

20. Hyvarinen, A., Morioka, H.: Nonlinear ica of temporally dependent stationary sources. In: Artificial intelligence and statistics. pp. 460–469. PMLR (2017)
21. Jiang, W., Guan, W., Li, H., Li, Z., Fang, Y., Peng, Y., Zhao, X., Liu, Y.: Text-conditional visual-language alignment for video captioning. IEEE Transactions on Circuits and Systems for Video Technology **36**(3), 3185–3200 (2026). https://doi.org/10.1109/TCSVT.2025.3616201
22. Jin, P., Li, H., Cheng, Z., Huang, J., Wang, Z., Yuan, L., Liu, C., Chen, J.: Text-video retrieval with disentangled conceptualization and set-to-set alignment. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. pp. 938–946 (2023). https://doi.org/10.24963/ijcai.2023/104
23. Kim, D., Zhang, J., Jin, W., Cho, S., Dai, Q., Park, J., Luo, C.: Subject-driven video generation via disentangled identity and motion. arXiv preprint arXiv:2504.17816 (2025)
24. Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., Paiton, D.: Towards nonlinear disentanglement in natural data with temporal sparse coding. arXiv preprint arXiv:2007.10930 (2020)
25. Ko, D., Choi, J., Ko, J., Noh, S., On, K.W., Kim, E.S., Kim, H.J.: Video-text representation learning via differentiable weak temporal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5016–5025 (2022)
26. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
27. Lai, Z., Zhang, H., Zhang, B., Wu, W., Bai, H., Timofeev, A., Du, X., Gan, Z., Shan, J., Chuah, C.N., et al.: Veclip: Improving clip training via visual-enriched captions. In: European Conference on Computer Vision. pp. 111–127. Springer (2024)
28. Lewis, M., Nayak, N., Yu, P., Merullo, J., Yu, Q., Bach, S., Pavlick, E.: Does CLIP bind concepts? probing compositionality in large image models. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 1487–1500. Association for Computational Linguistics, St. Julian's, Malta (2024). https://doi.org/10.18653/v1/2024.findings-eacl.101
29. Li, P., Xie, C.W., Zhao, L., Xie, H., Ge, J., Zheng, Y., Zhao, D., Zhang, Y.: Progressive spatio-temporal prototype matching for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4100–4110 (2023)
30. 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)
31. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., Ji, R.: X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In: Proceedings of the 30th ACM international conference on multimedia. pp. 638–647 (2022)
32. Materzyńska, J., Torralba, A., Bau, D.: Disentangling visual and written concepts in clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16410–16419 (2022)
33. Morioka, H., Hyvärinen, A.: Causal representation learning made identifiable by grouping of observational variables. arXiv preprint arXiv:2310.15709 (2023)
34. Morioka, H., Hyvarinen, A.: Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In: International conference on artificial intelligence and statistics. pp. 3399–3426. PMLR (2023)

35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
36. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: International Conference on Learning Representations (2021)
37. 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)
38. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6545–6554 (2023)
39. Rizve, M.N., Fei, F., Unnikrishnan, J., Tran, S., Yao, B.Z., Zeng, B., Shah, M., Chilimbi, T.: Vidla: Video-language alignment at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14043–14055 (2024)
40. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (1978). https://doi.org/10.1109/TASSP.1978.1163055
41. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
42. Shen, Y., Zhu, P., Li, Z., Xie, S., Deka, N., Liu, Z., Tang, Z., Chen, G., Zhang, K.: Controllable video generation with provable disentanglement. In: The Fourteenth International Conference on Learning Representations (2026)
43. Shi, Y., Xu, H., Yuan, C., Li, B., Hu, W., Zha, Z.J.: Learning video-text aligned representations for video captioning. ACM Transactions on Multimedia Computing, Communications and Applications **19**(2), 1–21 (2023)
44. Sun, Y., Kong, L., Chen, G., Li, L., Luo, G., Li, Z., Zhang, Y., Zheng, Y., Yang, M., Stojanov, P., Segal, E., Xing, E.P., Zhang, K.: Causal representation learning from multimodal biomedical observations. arXiv preprint arXiv:2411.06518 (2025)
45. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C.: Winoground: Probing vision and language models for visio-linguistic compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5238–5248 (2022)
46. Tian, K., Cheng, Y., Liu, Y., Hou, X., Chen, Q., Li, H.: Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 5207–5214 (2024)
47. Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F.: Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems **34**, 16451–16467 (2021)
48. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li,

Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
49. Wang, J., Wang, C., Huang, K., Huang, J., Jin, L.: Videoclip-xl: Advancing long description understanding for video clip models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 16061–16075 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.898
50. Wang, J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3313–3322 (2022)
51. Wang, Q., Zhang, Y., Zheng, Y., Pan, P., Hua, X.S.: Disentangled representation learning for text-video retrieval. arXiv preprint arXiv:2203.07111 (2022)
52. Wang, W., Yang, Y.: Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. Advances in Neural Information Processing Systems **37**, 65618–65642 (2024). https://doi.org/10.52202/079017-2096
53. Wang, W., Yang, Y.: Videoufo: A million-scale user-focused dataset for text-to-video generation. arXiv preprint arXiv:2503.01739 (2025)
54. Wang, X., Zhu, L., Yang, Y.: T2vlad: Global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5079–5088 (2021)
55. Wang, X., Zhu, L., Zheng, Z., Xu, M., Yang, Y.: Align and tell: Boosting text-video retrieval with local alignment and fine-grained supervision. IEEE Transactions on Multimedia **25**, 6079–6089 (2023). https://doi.org/10.1109/TMM.2022.3204444
56. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: The Twelfth International Conference on Learning Representations (2024)
57. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024)
58. 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)
59. Wu, P., He, X., Tang, M., Lv, Y., Liu, J.: Hanet: Hierarchical alignment networks for video-text retrieval. In: Proceedings of the 29th ACM international conference on Multimedia. pp. 3518–3527 (2021)
60. Xie, S., Kong, L., Zheng, Y., Tang, Z., Xing, E.P., Chen, G., Zhang, K.: Learning vision and language concepts for controllable image generation. In: Forty-second International Conference on Machine Learning (2025)
61. Xie, S., Kong, L., Zheng, Y., Yao, Y., Tang, Z., Xing, E.P., Chen, G., Zhang, K.: Smartclip: Modular vision-language alignment with identification guarantees. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29780–29790 (2025). https://doi.org/10.1109/CVPR52734.2025.02772
62. Xiong, T., Wang, Y., Zhou, D., Lin, Z., Feng, J., Liu, X.: Lvd-2m: A long-take video dataset with temporally dense captions. Advances in Neural Information Processing Systems **37**, 16623–16644 (2024)
63. Xu, D., Yao, D., Lachapelle, S., Taslakian, P., Von Kügelgen, J., Locatello, F., Magliacane, S.: A sparsity principle for partially observable causal representation learning. arXiv preprint arXiv:2403.08335 (2024)

64. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16456–16466 (2022)
65. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. In: The Eleventh International Conference on Learning Representations (2023)
66. Xue, Z., Zhang, J., Hu, T., He, H., Chen, Y., Cai, Y., Wang, Y., Wang, C., Liu, Y., Li, X., et al.: Ultravideo: High-quality uhd video dataset with comprehensive captions. arXiv preprint arXiv:2506.13691 (2025)
67. Yang, X., Zhu, L., Wang, X., Yang, Y.: Dgl: Dynamic global-local prompt tuning for text-video retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6540–6548 (2024)
68. Yao, D., Xu, D., Lachapelle, S., Magliacane, S., Taslakian, P., Martius, G., von Kügelgen, J., Locatello, F.: Multi-view causal representation learning with partial observability. arXiv preprint arXiv:2311.04056 (2023)
69. Yao, W., Chen, G., Zhang, K.: Learning latent causal dynamics. arXiv preprint arXiv:2202.04828 (2022)
70. Yao, W., Sun, Y., Ho, A., Sun, C., Zhang, K.: Learning temporally causal latent processes from general temporal data. arXiv preprint arXiv:2110.05428 (2021)
71. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? In: The Eleventh International Conference on Learning Representations (2023)
72. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: European conference on computer vision. pp. 310–325. Springer (2024)
73. Zhang, W., Wan, C., Liu, T., Tian, X., Shen, X., Ye, J.: Enhanced motion-text alignment for image-to-video transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18504–18515 (2024)
74. Zheng, Y., Ng, I., Zhang, K.: On the identifiability of nonlinear ica: Sparsity and beyond. Advances in neural information processing systems **35**, 16411–16422 (2022)
75. Zheng, Y., Zhang, K.: Generalizing nonlinear ica beyond structural sparsity. Advances in Neural Information Processing Systems **36**, 13326–13355 (2023)

Supplementary Material

A Proof

Our identification analysis shows that the selective, mask-based alignment objective in (3) recovers caption-conditioned semantic blocks up to blockwise invertible maps and extends to unions/intersections of such blocks. We study the constrained limit of the loss (alignment tolerance $\eta \to 0$), which enforces exact masked equality at optimum

$$\hat{\mathbf{z}}_t^{\mathrm{V}} \odot \hat{\mathbf{m}}_t^{\mathrm{V}} \;=\; \hat{\mathbf{z}}^{\mathrm{T}} \odot \hat{\mathbf{m}}_t^{\mathrm{T}}, \qquad t = 1, \ldots, T,$$

mirroring the dual-mask generative relation (2). Because similarities are computed *after masking*, the loss constrains only the active coordinates; the sparsity term selects minimal supports that still achieve strong positive alignment and negative separation.
In this setting, Lemma A.1 proves that masked alignment forces nuisance invariance on the active coordinates and yields blockwise invertible reparameterizations of the text and visual blocks, reducing identification to choosing the correct coordinates. Leveraging mask recurrence and view diversity, Lemma A.3 shows that the learned text masks are class-wise consistent and support-correct at the optimum (no extras, no misses), a consequence of the alignment geometry plus the sparsity/contrastive trade-off. With these blocks fixed, Theorem A.4 identifies any union or intersection of recurring blocks up to blockwise invertible maps, and the corresponding claims hold on the visual side via the dual-mask equalities.

Lemma A.1 (Nuisance invariance on the active block and blockwise invertibility). *Let $r_{\mathrm{T}} := \hat{g}^{\mathrm{T}} \circ g^{\mathrm{T}}$ and $r_{\mathrm{I}} := \hat{g}^{\mathrm{V}} \circ g^{\mathrm{V}}$. Fix a frame t and assume Condition 4.2(i) (smoothness and local invertibility of the modality generators and encoders). Assume further that in a small neighborhood of the data point the learned mask supports are stable, i.e.,* $\mathrm{supp}(\hat{\mathbf{T}}_t)$ *and* $\mathrm{supp}(\hat{\mathbf{V}}_t)$ *do not change. Define the binary diagonal projectors*

$$P_T := \mathrm{Diag}(\hat{\mathbf{T}}_t), \qquad P_V := \mathrm{Diag}(\hat{\mathbf{V}}_t).$$

At an optimum of the selective-alignment objective (the constrained limit of (3)), the masked alignment holds

$$P_T\, \hat{\mathbf{z}}^{\mathrm{T}} \;=\; P_V\, \hat{\mathbf{z}}_t^{\mathrm{V}}, \quad \hat{\mathbf{z}}^{\mathrm{T}} = r_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}}), \quad \hat{\mathbf{z}}_t^{\mathrm{V}} = r_{\mathrm{I}}(\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}}). \tag{10}$$

Then:

1. *(*Nuisance invariance on the text active block*) For every coordinate i with $[P_T]_{ii} = 1$ and every text-nuisance coordinate j, $\frac{\partial [\hat{\mathbf{z}}^{\mathrm{T}}]_i}{\partial [\boldsymbol{\epsilon}^{\mathrm{T}}]_j} = 0$. Equivalently, $P_T\, \hat{\mathbf{z}}^{\mathrm{T}}$ is independent of $\boldsymbol{\epsilon}^{\mathrm{T}}$.*

2. *(*Blockwise invertible reparameterizations*) There exist smooth, locally invertible maps h_{T} and h_{V} (defined on the respective active blocks) such that*

$$P_T\,\hat{\mathbf{z}}^{\mathrm{T}} \;=\; P_T\,h_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}), \qquad P_V\,\hat{\mathbf{z}}_t^{\mathrm{V}} \;=\; P_V\,h_{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}}).$$

Proof. **Step A: Fix projectors and use alignment.** Because mask supports are locally constant, P_T and P_V are fixed binary projectors in the neighborhood. By optimality of the selective-alignment loss, (10) holds. Importantly, the right-hand side $P_V\,\hat{\mathbf{z}}_t^{\mathrm{V}} = P_V\,r_{\mathrm{I}}(\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}})$ does not contain the text nuisance $\boldsymbol{\epsilon}^{\mathrm{T}}$.
Step B: Nuisance invariance on the active block (contradiction via partial derivatives). Fix any active coordinate i with $[P_T]_{ii} = 1$ and any nuisance coordinate j of $\boldsymbol{\epsilon}^{\mathrm{T}}$. Suppose, towards a contradiction, that $\partial[r_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}})]_i/\partial[\boldsymbol{\epsilon}^{\mathrm{T}}]_j \neq 0$ at some $(\mathbf{z}_0^{\mathrm{T}}, \boldsymbol{\epsilon}_0^{\mathrm{T}})$. By smoothness, this partial derivative keeps a nonzero sign in a small interval of $[\boldsymbol{\epsilon}^{\mathrm{T}}]_j$ around $[\boldsymbol{\epsilon}_0^{\mathrm{T}}]_j$. Holding $(\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}})$ fixed and varying only $[\boldsymbol{\epsilon}^{\mathrm{T}}]_j$ along that interval makes the left-hand side component $[P_T\,r_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}^{\mathrm{T}})]_i$ change strictly, while the right-hand side component $[P_V\,r_{\mathrm{I}}(\mathbf{z}_t^{\mathrm{V}}, \boldsymbol{\epsilon}_t^{\mathrm{V}})]_i$ remains constant (it does not depend on $\boldsymbol{\epsilon}^{\mathrm{T}}$), contradicting (10). Hence $\partial[\hat{\mathbf{z}}^{\mathrm{T}}]_i/\partial[\boldsymbol{\epsilon}^{\mathrm{T}}]_j = 0$ for all active i and all j, i.e., $P_T\,\hat{\mathbf{z}}^{\mathrm{T}}$ is invariant to $\boldsymbol{\epsilon}^{\mathrm{T}}$.
Step C: Blockwise invertible reparameterizations (local diffeomorphisms on the active block). Define $\phi_T(\mathbf{z}^{\mathrm{T}}) := P_T\,r_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}}, \boldsymbol{\epsilon}_0^{\mathrm{T}})$ for any fixed $\boldsymbol{\epsilon}_0^{\mathrm{T}}$; this is well-defined because $P_T\,\hat{\mathbf{z}}^{\mathrm{T}}$ is independent of $\boldsymbol{\epsilon}^{\mathrm{T}}$ by Step B. By Condition 4.2(i), the composite encoder r_{T} is smooth and locally invertible with respect to the semantic coordinates, and its restriction to the active coordinates selected by P_T has full (block) rank. Therefore, by the inverse function theorem (or constant-rank theorem), ϕ_T defines a local diffeomorphism between the true text latent block and its encoded image on that block. We denote the induced reparameterization by h_{T} and obtain $P_T\,\hat{\mathbf{z}}^{\mathrm{T}} = P_T\,h_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}})$. The same argument on the visual side yields $P_V\,\hat{\mathbf{z}}_t^{\mathrm{V}} = P_V\,h_{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}})$.

Assumption A.2 (Mask Recurrence (class-based)). *There is a finite family of text-side masks $\{\mathbf{T}^{(k)}\}_{k\in\mathcal{K}}$ with $\Pr(\mathbf{T} = \mathbf{T}^{(k)}) > 0$. For each k, the set $\mathcal{S}_k := \{t : \mathbf{T}_t = \mathbf{T}^{(k)}\}$ contains multiple samples (not necessarily adjacent in time), and conditioned on $\mathbf{T}_t = \mathbf{T}^{(k)}$ the visual latents/nuisances vary in a neighborhood (view diversity).*

Lemma A.3 (Consistency and support-correctness of learned text masks). *Let $r_{\mathrm{T}} := \hat{g}^{\mathrm{T}} \circ g^{\mathrm{T}}$ and $r_{\mathrm{I}} := \hat{g}^{\mathrm{V}} \circ g^{\mathrm{V}}$ and adopt the notation of Lemma A.1. Assume Condition 4.2(i)–(iv) and Assumption A.2. Fix a class k and write $H_T := h_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}})$ and $H_{V,t} := h_{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}})$ from Lemma A.1. At an optimum of the selective-alignment objective (3), the following hold:*

1. ***Class-wise consistency.*** *For all $t_1, t_2 \in \mathcal{S}_k$,*

$$\hat{\mathbf{m}}_{t_1}^{\mathrm{T}} = \hat{\mathbf{m}}_{t_2}^{\mathrm{T}}\,.$$

2. ***Support-correctness.*** *Let $\mathbf{T}^{(k)}$ be the true text mask of class k. Then $\mathcal{B}(\hat{\mathbf{m}}_t^{\mathrm{T}}) = \mathcal{B}(\mathbf{T}^{(k)})$ for all $t \in \mathcal{S}_k$ (up to a block permutation).*

Proof. **Setup.** By Lemma A.1, on a support-stable neighborhood the masked alignment reads, for each t,

$$H_T \odot \hat{\mathbf{m}}_t^{\mathrm{T}} \;=\; H_{V,t} \odot \hat{\mathbf{m}}_t^{\mathrm{V}}, \tag{11}$$

where $H_T = h_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}})$ depends only on $\mathbf{z}^{\mathrm{T}}$ (no $\boldsymbol{\epsilon}^{\mathrm{T}}$), while $H_{V,t} = h_{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}})$ can vary across $t \in \mathcal{S}_k$ by view diversity.

(1) Class-wise consistency. Fix k and suppose, by contradiction, there exist $t_1, t_2 \in \mathcal{S}_k$ with $\hat{\mathbf{m}}_{t_1}^{\mathrm{T}} \neq \hat{\mathbf{m}}_{t_2}^{\mathrm{T}}$. Let $\Delta := \mathcal{B}(\hat{\mathbf{m}}_{t_1}^{\mathrm{T}}) \triangle \mathcal{B}(\hat{\mathbf{m}}_{t_2}^{\mathrm{T}})$ be the symmetric difference. Pick any $i \in \Delta$; without loss of generality assume $i \in \mathcal{B}(\hat{\mathbf{m}}_{t_1}^{\mathrm{T}})$ and $i \notin \mathcal{B}(\hat{\mathbf{m}}_{t_2}^{\mathrm{T}})$. Taking coordinate i in (11), we get

$$[H_T]_i \;=\; [H_{V,t_1}]_i \, [\hat{\mathbf{m}}_{t_1}^{\mathrm{V}}]_i, \qquad 0 \;=\; [H_{V,t_2}]_i \, [\hat{\mathbf{m}}_{t_2}^{\mathrm{V}}]_i.$$

By Assumption A.2 (view diversity in the class), the right factors $[H_{V,t}]_i$ vary in a neighborhood across $t \in \mathcal{S}_k$. Thus, to satisfy the second equality for (almost) all such t_2, the only robust possibility is that $[\hat{\mathbf{m}}_{t_2}^{\mathrm{V}}]_i = 0$ on a full neighborhood; in particular the right-hand side is forced to zero generically. But the first equality requires reproducing the *same* nonzero value $[H_T]_i$ (which is independent of t) on some t_1. Because $(H_{V,t})_{t \in \mathcal{S}_k}$ explore a neighborhood, the pair of requirements is generically incompatible unless i is *never* selected on the text side within the class. Consequently Δ must be empty. Hence $\hat{\mathbf{m}}_{t_1}^{\mathrm{T}} = \hat{\mathbf{m}}_{t_2}^{\mathrm{T}}$ for all $t_1, t_2 \in \mathcal{S}_k$.

(2) Support-correctness (no extras, no misses). Write $\hat{\mathbf{m}}^{\mathrm{T}}$ for the common class-wise text mask established above.

No extra coordinates. Suppose $\mathcal{B}(\hat{\mathbf{m}}^{\mathrm{T}})$ strictly contains $\mathcal{B}(\mathbf{T}^{(k)})$. Let $E := \mathcal{B}(\hat{\mathbf{m}}^{\mathrm{T}}) \setminus \mathcal{B}(\mathbf{T}^{(k)})$ be the set of spurious text coordinates. For any $i \in E$, the text side reveals $[H_T]_i$ in the positive pairs for class k, but the true cross-modal semantics in class k do not require i. In contrastive training, these exposed coordinates $i \in E$ (a) increase similarity to negatives that coincidentally activate i, and (b) are penalized by the sparsity term $\sum_t \|\hat{\mathbf{m}}_t^{\mathrm{T}}\|_0$. Turning them off strictly improves the objective without harming the positive alignment (since class-k semantics do not need them). Therefore the optimum cannot have $E \neq \emptyset$.

No missing coordinates. Suppose $\mathcal{B}(\hat{\mathbf{m}}^{\mathrm{T}})$ is a strict subset of $\mathcal{B}(\mathbf{T}^{(k)})$. Let $M := \mathcal{B}(\mathbf{T}^{(k)}) \setminus \mathcal{B}(\hat{\mathbf{m}}^{\mathrm{T}})$ be the set of missing true coordinates. For $i \in M$, the positive alignment in (11) discards genuinely informative text dimensions $[H_T]_i$, reducing positive similarity and making hard negatives harder to separate. Activating i increases positive alignment while the sparsity cost grows only additively by $|M|$. By the usual margin trade-off in the contrastive objective, the optimum cannot omit any $i \in \mathcal{B}(\mathbf{T}^{(k)})$. Thus $M = \emptyset$.

Combining the two parts gives $\mathcal{B}(\hat{\mathbf{m}}^{\mathrm{T}}) = \mathcal{B}(\mathbf{T}^{(k)})$ (up to a block permutation).

Theorem A.4 (Block-wise identifiability for unions and intersections). *Let $\{\mathbf{T}^{(k)}\}_{k \in \mathcal{K}}$ be the (true) text-side masks for the recurring classes, and let $\hat{\mathbf{T}}^{(k)}$ be the corresponding learned masks established in Lemma A.3 (class-wise*

consistent and support-correct, up to a block permutation). For any finite subfamily $\mathcal{V} \subseteq \mathcal{K}$, define the union and intersection index sets

$$\tilde{\mathcal{B}}_{\cup} := \bigcup_{k \in \mathcal{V}} \mathcal{B}(\mathbf{T}^{(k)}), \qquad \tilde{\mathcal{B}}_{\cap} := \bigcap_{k \in \mathcal{V}} \mathcal{B}(\mathbf{T}^{(k)}).$$

Then the true latent sub-vectors $[\mathbf{z}]_{\tilde{\mathcal{B}}_{\cup}}$ and $[\mathbf{z}]_{\tilde{\mathcal{B}}_{\cap}}$ are identifiable up to block-wise invertible maps; concretely, there exist smooth local reparameterizations $h_{\cup}, h_{\cap}$ such that for the text representation,

$$[\hat{\mathbf{z}}^{\mathrm{T}}]_{\tilde{\mathcal{B}}_{\cup}} = [h_{\cup}(\mathbf{z}^{\mathrm{T}})]_{\tilde{\mathcal{B}}_{\cup}}, \qquad [\hat{\mathbf{z}}^{\mathrm{T}}]_{\tilde{\mathcal{B}}_{\cap}} = [h_{\cap}(\mathbf{z}^{\mathrm{T}})]_{\tilde{\mathcal{B}}_{\cap}},$$

and the same holds for the aligned visual blocks via the dual-mask equalities.

Proof. Let $H_T := h_{\mathrm{T}}(\mathbf{z}^{\mathrm{T}})$ and $H_{V,t} := h_{\mathrm{V}}(\mathbf{z}_t^{\mathrm{V}})$ be the block-wise reparameterizations from Lemma A.1. By Lemma A.3, for each class k there exists a learned text mask $\hat{\mathbf{T}}^{(k)}$ whose support equals $\mathcal{B}(\mathbf{T}^{(k)})$ (up to a block permutation), and for any frame t belonging to class k we have the masked alignment

$$\mathrm{Diag}(\hat{\mathbf{T}}^{(k)})\, \hat{\mathbf{z}}^{\mathrm{T}} \;=\; \mathrm{Diag}(\hat{\mathbf{V}}_t)\, \hat{\mathbf{z}}_t^{\mathrm{V}} \;=\; \mathrm{Diag}(\hat{\mathbf{T}}^{(k)})\, H_T, \tag{12}$$

where the last equality is Lemma A.1's invariance on the text active block.
(A) Unions. Define the union mask (projector) $P_{\cup} := \mathrm{Diag}(\mathbf{1}_{\tilde{\mathcal{B}}_{\cup}})$. Take any coordinate $i \in \tilde{\mathcal{B}}_{\cup}$. Then $i \in \mathcal{B}(\mathbf{T}^{(k)})$ for some $k \in \mathcal{V}$, hence by (12) we have $[\hat{\mathbf{z}}^{\mathrm{T}}]_i = [H_T]_i$ (up to the fixed within-block permutation inherited from $\hat{\mathbf{T}}^{(k)}$). Since this holds for every $i \in \tilde{\mathcal{B}}_{\cup}$,

$$P_{\cup}\, \hat{\mathbf{z}}^{\mathrm{T}} \;=\; P_{\cup}\, H_T.$$

By Lemma A.1, the restriction of H_T to any active coordinates is a local diffeomorphism of the corresponding true latent block; in particular, the Jacobian of $\mathbf{z}^{\mathrm{T}} \mapsto [H_T]_{\tilde{\mathcal{B}}_{\cup}}$ has full (block) rank in a neighborhood. Therefore $[H_T]_{\tilde{\mathcal{B}}_{\cup}}$ is an invertible (block-wise) reparameterization of $[\mathbf{z}^{\mathrm{T}}]_{\tilde{\mathcal{B}}_{\cup}}$, which we denote by $[h_{\cup}(\mathbf{z}^{\mathrm{T}})]_{\tilde{\mathcal{B}}_{\cup}}$. Hence $[\hat{\mathbf{z}}^{\mathrm{T}}]_{\tilde{\mathcal{B}}_{\cup}} = [h_{\cup}(\mathbf{z}^{\mathrm{T}})]_{\tilde{\mathcal{B}}_{\cup}}$ and the union block is identifiable up to a block-wise invertible map.
(B) Intersections. For intersections, take any two classes $k_1, k_2 \in \mathcal{V}$ and consider the common index set $\mathcal{B}^{(1,2)} := \mathcal{B}(\mathbf{T}^{(k_1)}) \cap \mathcal{B}(\mathbf{T}^{(k_2)})$. Restricting (12) to $\mathcal{B}^{(1,2)}$ for frames from class k_1 and from class k_2 yields

$$\begin{aligned} [H_T]_{\mathcal{B}^{(1,2)}} &= [\hat{\mathbf{z}}^{\mathrm{T}}]_{\mathcal{B}^{(1,2)}} \\ &= \big[\mathrm{Diag}(\hat{\mathbf{V}}_t)\, \hat{\mathbf{z}}_t^{\mathrm{V}}\big]_{\mathcal{B}^{(1,2)}} \quad \text{for both } k_1 \text{ and } k_2. \end{aligned} \tag{13}$$

Thus the same sub-vector $[H_T]_{\mathcal{B}^{(1,2)}}$ is compelled by alignment across multiple classes/frames, pinning it to the text coordinates in the intersection. By Lemma A.1, the restriction $\mathbf{z}^{\mathrm{T}} \mapsto [H_T]_{\mathcal{B}^{(1,2)}}$ is locally invertible on that block, so we obtain identifiability of $[\mathbf{z}^{\mathrm{T}}]_{\mathcal{B}^{(1,2)}}$ up to a block-wise invertible map. Since

finite intersections can be built iteratively (pairwise intersections are associative/commutative at the index-set level), this extends to $\tilde{\mathcal{B}}_\cap = \bigcap_{k\in\mathcal{V}} \mathcal{B}(\mathbf{T}^{(k)})$:

$$[\hat{\mathbf{z}}^{\mathrm{T}}]_{\tilde{\mathcal{B}}_\cap} = [H_T]_{\tilde{\mathcal{B}}_\cap} = [h_\cap(\mathbf{z}^{\mathrm{T}})]_{\tilde{\mathcal{B}}_\cap}.$$

(C) Visual side. For any block selected on the text side, (12) implies the same block is realized on the visual side via $\mathrm{Diag}(\hat{\mathbf{V}}_t)\,\hat{\mathbf{z}}_t^{\mathrm{V}}$, so the corresponding visual sub-vectors are also identifiable up to block-wise invertible maps by Lemma A.1's h_{V}.

Combining (A)–(C) establishes the claim for both unions and intersections.

B Dataset Details

We evaluate MoVA on five video-text retrieval datasets: ActivityNet [26], MSVD [5], DiDeMo [1], VideoUFO [53], and UltraVideo [66].

ActivityNet [26] contains 20,000 YouTube videos and 100,000 captions (849 hours in total) covering 200 activity types, with an average length of 153 seconds. Following [30], we concatenate the multiple descriptions per video into a single paragraph for paragraph-to-video retrieval. We train on 10,000 videos and evaluate on the `val1` split with 5,000 videos.

MSVD [5] includes 1,970 short clips (1–62 seconds) and about 40 captions per video. We adopt the standard split of 1,200/100/670 videos for train/val/test and report retrieval under the multi-caption evaluation protocol [30].

DiDeMo [1] consists of 10,000 videos and 40,000 captions (average length $\sim$ 30 seconds). Similar to ActivityNet, we perform paragraph-to-video retrieval. The train/val/test sets contain 8,395/1,065/1,004 videos, and results are reported on the test split.

VideoUFO [53] is a million-scale user-focused benchmark with 1.09M clips across 1,291 topics at 720p resolution. Captions average 155.5 words, while videos average 12.6 seconds (about 3.5K hours in total), posing retrieval challenges due to both scale and description richness. We randomly sample 95% (1,037,126) clips for training and 5% (54,586) for testing.

UltraVideo [66] targets high-quality text-to-video research with 4K/8K footage and structured captions. It contains 17K clips (143 hours) with 30.9-second videos on average and very long descriptions (average 850.3 words). We use the UltraVideo-Long subset and refer to it as UltraVideo for brevity, taking the first 75% (12,447) clips from the original caption annotation file for training and the remaining 25% (4,150) for testing.

Taken together, these datasets span short clips with concise captions to ultra-high-resolution videos paired with hundred-word paragraphs, offering a comprehensive testbed for scaling MoVA's alignment ability across length, resolution, and topic breadth. VideoUFO and UltraVideo (introduced to video-text retrieval for the first time in this work) deliberately stress long, information-dense captions: short captions can retrieve many near-duplicate results, whereas long captions challenge the model to extract effective signals while avoiding loss of key

concepts. The rich annotations also encourage multi-task reuse of the learned video–text representations, such as long-text-to-video generation.

C Additional Experiments

Additional quantitative results. Table 5 reports DiDeMo retrieval when every method uses a weaker ViT-B/32 backbone. MoVA still outperforms the strongest baseline (CLIP-VIP) by +3.0 R@1 and halves MnR (7.8 vs. 13.5) without leveraging auxiliary video–subtitle or video–caption corpora—we only fine-tune on the given benchmark. In Table 2, MoVA likewise surpasses high-resolution models such as VideoCLIP-XL [49], showing that the modular alignment remains effective without resorting to larger encoders.

Table 5: Text-to-video retrieval performance on the DiDeMo dataset (ViT-B/32 as backbone). "↑" denotes that higher is better. "↓" denotes that lower is better.

Method	R@1↑	R@5↑	MdR↓	MnR↓
CLIP4Clip [30]	42.8	68.5	2.0	18.9
X-CLIP [31]	45.2	74.0	2.0	14.6
CLIP-VIP [65]	47.4	75.2	2.0	13.5
UCOFIA [58]	46.5	74.8	2.0	13.1
ProST [29]	44.9	72.7	2.0	13.7
MoVA (Ours)	**50.4**	**82.1**	**1.0**	**7.8**

Table 6: Quantitative comparison of total parameter counts (in Millions).

Method	**Params (M)**
CLIP [37]	149.6
SmartCLIP [61]	153.0
CLIP4Clip [30]	162.3
DRL [51]	162.8
X-CLIP [31]	162.8
ProST [29]	177.9
DiCoSA [22]	162.3
VideoCLIP-XL [49]	427.9
MoVA (Ours)	174.7

Parameters and training cost. Table 6 reports total parameters. The Concept Mask Network adds 11,099,650 (≈ 11.1M) parameters and the Temporal Mask Network adds 7,946,753 (≈ 7.9M), bringing MoVA to 174.7M—about a 7.6%

increase over CLIP4Clip (162.3M) yet far below VideoCLIP-XL (427.9M, from its released ViT-L/14 checkpoint). Compared with ProST (177.9M), MoVA uses roughly 1.8% fewer parameters while delivering stronger retrieval and concept disentanglement. On an 8×MI210 node, training ActivityNet for one epoch takes 76.7 minutes for MoVA versus 92.9 for CLIP4Clip, showing the added modules keep memory and compute overhead acceptable relative to the gains.

Ablations. Tables 7 and 8 study frame counts and TMN block numbers. Performance is stable across frame budgets (32–160), peaking near 96 frames, indicating the model does not rely on blindly scaling frames. Increasing TMN blocks beyond a shallow depth does not yield further gains, so MoVA remains efficient without over-deep temporal masking. In addition, both $\mathcal{L}_{\text{sfdm}}$ and $\mathcal{L}_{\text{dfsm}}$ are necessary, as removing $\mathcal{L}_{\text{dfsm}}$ alone reduces ActivityNet R@1 by 3.6. Furthermore, soft mask variant (same L_1 sparsity) is comparable (ActivityNet R@1/5/10 47.4/77.3/86.6; DiDeMo R@1 57.2; UltraVideo R@1 58.3). Hard masks remain default because of better interpretability.

Table 7: Ablation study on the number of input frames. Evaluated on the ActivityNet dataset.

Frames	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
32	46.7	75.6	86.3	7.1	45.3	75.0	86.9	7.4
64	47.2	76.4	86.2	6.6	45.1	75.4	87.1	6.6
96	47.9	77.3	86.9	6.9	46.7	76.2	86.9	6.3
128	47.8	77.4	87.0	7.0	46.7	76.6	86.8	6.3
160	46.6	76.1	87.2	6.9	45.2	76.1	86.9	6.4

Table 8: Ablation study on the number of TMN blocks. Evaluated on the ActivityNet dataset. The Temporal Mask Network depth is varied while keeping other settings fixed.

Blocks	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MnR↓	R@1↑	R@5↑	R@10↑	MnR↓
1	47.8	77.0	87.0	6.7	46.8	77.0	87.6	6.2
2	47.8	77.4	87.0	7.0	46.7	76.6	86.8	6.3
3	46.6	76.2	86.4	6.5	45.3	76.2	87.0	6.5
4	47.8	76.9	87.1	6.3	46.8	76.8	87.1	6.2
8	47.0	76.6	87.0	6.1	45.7	76.6	87.5	5.9

More visualization examples. Figure 8 shows two long-prompt generations. The first captures a stylized meerkat in a dim, curtain-lined set with a patchworked coat, magenta collar, and subtle swaying while singing. The second depicts an industrial welding scene with a rusted mask, blue-black jacket, glowing orange material, and gritty metal surroundings. Beyond static fidelity, MoVA preserves temporal cues (e.g., rhythmic swaying, welding interaction) without drifting or collapsing motion, and avoids artifacts like the frame-level watermarks

often seen when vanilla CLIP encoders are fine-tuned on video-text data. Similar to Figure 5, we omit other video-text fine-tuned baselines (e.g., CLIP4Clip, DGL) because after the same number of iterations they lose structured text-to-video guidance, often producing near-identical outputs for different texts or misaligned videos.

A highly stylized, **slender meerkat** *stands upright, serving as the central figure in a* **peculiar, dimly lit environment**. *She wears an elaborate* **long coat patchworked with tattered fabrics** *in various colors, finished with a striking* **magenta collar and lapels**. *The meerkat stands on a rough floor beneath a focused spotlight. Behind her, a backdrop of massive, abstract, vividly colored* **curtains**—*yellow, blue, and burgundy—overlaid with* **stylized graffiti**, *suggests a fantastical, tribal-inspired aesthetic. She* **sways in place**, *rocking from heel to toe with small shoulder rolls, tilting her head toward the light as she* **delivers a clear, lilting melody**; *the tune rises and falls in long, sustained phrases, tender and affecting.*

A highly detailed, **anthropomorphic humanoid robot** *on the right is engaged in* **metalworking and assembly**. *It wears a* **blue-black textured jacket** *spattered with rust-colored orange and white paint or grime, giving it a worn, workmanlike aesthetic. Its head is enclosed in a* **metal welding mask**—*heavily rusted, with complex wiring or hoses trailing from the back. The robot* **grips** *a* **glowing, bright-orange, intensely hot piece of material** *while interacting with the angular metallic arm of a second, partially visible robot on the left. The scene is* **industrial and gritty**, *marked by* **heavy metal surfaces**, *exposed wires, and a focus on mechanical labor.*

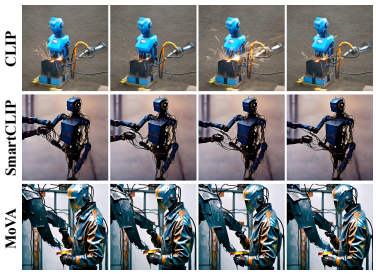

Fig. 8: Additional Comparison Examples of Long-text-to-video generation. These examples further demonstrate MoVA's capability in handling complex descriptive prompts. **Top:** The model accurately captures the stylized aesthetic, rendering the *patchworked coat* and the specific motion of *swaying* while singing. **Bottom:** MoVA successfully generates the fine-grained details of the *rusted welding mask* and the *glowing material* within an *industrial* atmosphere.