Boosting Video-Text Retrieval with Explicit High-Level Semantics

Haoran Wang∗†
wanghaoran09@baidu.com
Department of Computer Vision Technology (VIS), Baidu Inc.
Beijing, China

Di Xu‡
xudi20s@ict.ac.cn
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

Dongliang He
edongliang01@baidu.com
Department of Computer Vision Technology (VIS), Baidu Inc.
Beijing, China

Fu Li
lifu@baidu.com
Department of Computer Vision Technology (VIS), Baidu Inc.
Beijing, China

Zhong Ji
jizhong@tju.edu.cn
School of Electrical and Information Engineering, Tianjin University
Tianjin, China

Jungong Han
jungonghan77@gmail.com
Computer Science Department,
Aberystwyth University
SY23 3FL, UK

Errui Ding
dingerrui@baidu.com
Department of Computer Vision Technology (VIS), Baidu Inc.
Beijing, China

ABSTRACT

Video-text retrieval (VTR) is an attractive yet challenging task for multi-modal understanding, which aims to search for relevant video (text) given a query (video), whilst lacking the awareness of homogeneous high-level semantic information residing in both modalities. To fill this gap, in this work, we propose a novel visual-linguistic aligning model named HiSE for VTR, which improves the cross-modal representation by incorporating explicit high-level semantics. First, we explore the hierarchical property of explicit high-level semantics, and further decompose it into two levels, i.e. discrete semantics and holistic semantics. Specifically, for visual branch, we exploit an off-the-shelf semantic entity predictor to generate discrete high-level semantics. In parallel, a trained video captioning model is employed to output holistic high-level semantics. As for the textual modality, we parse the text into three parts including occurrence, action and entity. In particular, the occurrence corresponds to the holistic high-level semantics, meanwhile both action and entity represent the discrete ones. Then, different graph reasoning techniques are utilized to promote the interaction between holistic and discrete high-level semantics. Extensive experiments demonstrate that, with the aid of explicit high-level semantics, our method achieves the superior performance over state-of-the-art methods on three benchmark datasets, including MSR-VTT, MSVD and DiDeMo.

CCS CONCEPTS

• Information systems → Information retrieval; Retrieval models and ranking.

KEYWORDS

Video-Text Retrieval, High-level Semantics, Vision-language Understanding

1 INTRODUCTION

With the exponentially increasing of on-line videos, video-text retrieval (VTR) is becoming an emerging requirement for people to perceive the world. It refers to searching for a video (text) when given a query text (video), which plays a fundamental role in vision and language understanding [3, 4, 18, 20, 24, 29, 34, 40, 41, 43, 48]. Although remarkable progresses in this area have been made, it is still challenging to precisely match video and text since the raw input multi-modal data exist in heterogeneous spaces.

The main challenge for video-text alignment is how to narrow the gap between the heterogeneous representations from both modalities. To tackle this problem, previous solutions can be divided into two streams of works. One line of studies is built based on single branch representation architecture. These methods [31, 33, 50] typically employ an independent modality to represent video, i.e. appearance and utilize single encoder to represent it, whereas ignoring the that video is actually constituted by multiple constituent modalities. In contrast to these approaches, some multi-expert based
In this work, to bridge the semantic discrepancy between video and text, we propose a high-level semantics aided (HiSE) visual-linguistic embedding model for VTR application, which injects the EHS into cross-modal representation learning. To realize it, we first elaborately design different architectures to collect hierarchical EHS information for video and text, respectively. As for the visual modality, we employ two separate components to generate the discrete and holistic EHS. Specifically, to extract discrete EHS from video, we leverage an off-the-shelf semantic entity predictor [1] to detect semantic entities for all sampled frames. Then, the obtained entities are filtered according to their confidence scores and the retained part of them is taken as discrete EHS. Meanwhile, we employ an video captioning model to generate describing sentence for the given video. The output is taken as visual holistic EHS that contains more coherent information.

On the other side, to extract the hierarchical EHS from text, we utilize a semantic role parser to decompose it into a role graph, which includes three kinds of nodes: occurrence, action and entity. The occurrence refers to the whole sentence, which consists of more context information and is taken as textual holistic EHS. Besides, action is denoted by verbs and entity is represented by noun & noun phrases. Considering action and entity only describe local textual information, we name them as textual discrete EHS.

In addition to EHS acquisition, how to organize and integrate them also plays important role in cross-modal representation learning. For the video part, given the video discrete EHS, we use its textual embeddings in conjunction of a graph convolution operation to obtain the compact vectors, termed video discrete high-level semantic (VDS) representation. As for video holistic EHS representing, we employ another sentence-level text encoder to process it, which outputs the video holistic high-level semantic (VHS) representation. By contrast, for the textual branch, we uniformly employ the corresponding text embeddings to denote both discrete EHS and holistic EHS by compact vectors. Then, based on the role graph structure, we apply multi-graph reasoning to promote the interaction between the output embeddings from both levels. Lastly, we fuse the generated EHS representations with the raw outputs of modality-specific encoders by using their convex combination. Through integrating the EHS knowledge into the cross-modal representation framework, the bidirectional VTR performance can be remarkably boosted. On the whole, our main contributions lie in three-fold.

- We present a High-level SEmantic (HiSE) aided visual-linguistic embedding framework for VTR. To the best of our knowledge, this is the first work integrating explicit high-level semantics into video-text retrieval, which leverages the unified abstract information to strengthen the semantic relationship between two modalities.
- We decompose high-level semantics into two levels: i.e. discrete semantics and holistic semantics, which are responsible to capture local and global information respectively. Moreover, we design elaborated architectures to represent and tie them up to improve the cross-modal representation.
- The extensive experiments on benchmark datasets not only demonstrate the superiority of our approach by outperforming state-of-the-art methods for VTR, but also exhibits the interpretability.

2 RELATED WORK

2.1 Video-Text Retrieval

Accompanied by the renaissance of deep learning, there have been growing interest in research for VTR [5, 14, 22, 27, 31, 34, 53]. A majority of early works tackle this task from perspective of representation architecture. On one side, some works [11, 14, 27, 31] introduce multiple experts to enhance video representation. For instance, MoEE [31] jointly combined three sources of expert components for video encoding, including videos, motion and audio feature. On the other side, several studies [21, 22] focused on text encoding designing. Li et al. [22] introduced multiple sentence encoders and combining similarities from all text encoder-specific joint space. Based on the multi-expert representations, more works...
[8, 45] devoted to performing hierarchical alignments for VTR, and achieved continuous performance improvements.

Recently, since the prevalence of large-scale pre-training technique, a flurry of works [6, 13, 28, 35] leveraging this technique to promote video-text representation have emerged. As a representative study, Luo et al. [28] proposed a CLIP4CLIP method, which transfers the knowledge of the CLIP model to VTR application via an end-to-end fine-tuning. Then, Fang et al. [13] presented to plug a temporal information capturing module in CLIP4CLIP for video representation enhancing. Cao et al. [6] proposed a framework modelling visual consensus, which exploits commonsense knowledge in both vision and language domain to improve VTR. To sum up, all above approaches perform cross-modal alignment based on completely heterogeneous video and text representations. Distinct from them, our HiSE additionally introduces homogeneous high-level semantic clues to narrow the modality gap.

### 2.2 Multi-modal Understanding Using High-level Semantics

High-level semantics plays critical role in multi-modal data understanding [12, 17, 18, 32, 46, 49, 51]. For instance, Wu et al. [46] proposed a method that incorporates semantic concepts into the CNN-RNN architecture to improve image captioning. Similarly, Yao et al. [49] integrated attributes into the CNN-RNN image captioning framework, followed by training in an end-to-end manner. For VQA, Hudson et al. [17] presented to exploit high-level scene-graph knowledge to transform both image and text into semantic concept-based representations. In contrast to previous works, to the best of our knowledge, we make first attempt to simultaneously integrate explicit high-level semantics into video and text representations for VTR. The most related work to ours is SCO [16], which leveraged concepts to enhance image representation for image-text retrieval. On the contrast, we further extend high-level semantics of video to discrete semantics and holistic one, and validate the effectiveness of the hierarchical information on VTR task.

### 3 METHODOLOGY

In this section, we detailedly introduce our proposed high-level semantic (HiSE) aided visual-linguistic embedding model for video-text retrieval (see Figure 2). Firstly, we illustrate the modality-specific encoders to represent video and text along with their corresponding memory banks. Then, we illustrate the concrete architecture of two explicit high-level semantics representation modules, i.e. VSE module and TSE module, respectively. Afterwards, the cross-modal representation aggregating manner, inference method and alignment objectives are introduced sequentially.

#### 3.1 Video and Text Encoder

**3.1.1 Video Encoder.** To obtain the video representation, we first extract the frames from the video clip, and then employ a video encoder to turn them into a sequence of features, followed by fusing them by a feature aggregator. To realize it, we adopt pre-trained CLIP [36] (ViT-B/32) as video encoder, which utilizes the ViT [10] backbone pre-trained based on 400 million image-text pairs. Specifically, we split an image into non-overlapping patches, then use a linear projection to project them into 1-D tokens. The output vectors are taken as input of ViT, which leverages the transformer architecture to model the interaction between image patches. Following CLIP, we adopt the output from the [class] token as the image representation. Consequently, given the input sequence of video frames $O = \{o_1, ..., o_N\}$, the generated video features are denoted as $V = \{v_1, ..., v_N\}$. Afterwards, to further capture the temporal information between frames, we use another Transformer encoder [39] with position embedding to combine frames into one global video representation. Formally, the global video representation is calculated by $V = f_{\text{video}}(O)$, in which $f_{\text{video}}(\cdot)$ represents the video encoder.

**3.1.2 Text Encoder.** For caption encoding, we directly employ the text encoder from the CLIP to extract the textual representation. In particular, it refers to a Transformer with architecture modifications according to [36]. Following CLIP [36], we use the activations of the [EOS] token from the most top layer of the transformer as the global representation of the caption. Given the caption $S$, the global textual representation is computed according to $T = f_{\text{text}}(S)$, where $f_{\text{text}}(\cdot)$ denotes the text encoder.

**3.1.3 Modality-specific Memory Banks Building.** To further enhance the negative interactions for both modalities in contrastive learning, we propose to leverage two dynamic modality-specific memory banks $B_V$ and $B_T$ to store additional video and text representations. Particularly, we follow MoCo [15] to obtain momentum video encoder and text encoder by momentum updating their weights according to the corresponding modality-specific encoders, whose architectures are totally same as $f_{\text{video}}(\cdot)$ and $f_{\text{text}}(\cdot)$, respectively. As such, video or text samples from the latest training iterations are fed to the momentum encoders, which output video and text embeddings and restored them in coupled modality-specific memory banks. This process is implemented via deploying queues.

#### 3.2 Textual High-level Semantics Encoding

In this part, we describe the detailed structure of Textual High-level Semantics Encoding (TSE) module (see Figure 3(a)). We know that, video captioning sentence is naturally constructed based on hierarchical semantic structure, which can be expressed by the
relationships between the actions and entities to describe occurrence (see Figure 3(a)). Such hierarchical structure is beneficial to realize comprehensive understanding for video captions. Considering these relationships can be conveyed by graph structures, we exploit graph convolution to encode the explicit high-level semantics for text. The processing pipeline includes three steps: 1) Initializing semantic role graph node; 2) Graph building; 3) Graph Reasoning.

**Initializing semantic role graph node.** First, the graph nodes are divided into three classes: occurrence nodes, action nodes, and entity nodes. Specifically, the occurrence node representations are initialized by employing the pre-trained CLIP text encoder, which shares the same architecture with the encoder in 3.1.2 (not sharing weights). To initialize action and entity nodes, we use tokenzier of CLIP text encoder to split and map them into token embeddings, then aggregate the token embedding vectors by mean-polling operation to denote phrase-level contents (entity). Finally, the initialized occurrence/action/entity nodes are denoted by $E_o/E_a/E_e$, respectively.

**Role graph building.** Given a video description $S$, we take it as the occurrence node in semantic role graph. Then, an off-the-shelf semantic role parser [38] is leveraged to extract noun phrases and verbs from $S$ along with their semantic relationships. In our role graph, the noun phrases and verbs act as action nodes and entity nodes, respectively. The entity nodes are connected with different action nodes, where the edge type between them is defined by the semantic role of the entity in reference to the action. The action nodes all connect to the occurrence node with direct edges, in which the temporal clues between different actions can be captured.

**Graph Reasoning.** To effectively exploit the multiple roles of relational information contained in our graph, we adopt the relational graph convolutional network (R-GCN) [37] to model interactions in graph between nodes. Specifically, given the initialized nodes $E_i = (E_o, E_a, E_e)$ and role graphs $G = (G^r)$, through one-layer graph convolution with residual connection, the node embedding feature is computed as:

$$E_i^r = \rho(\sum_{r \in R} G^r E_i^{0,0} W^r + E_i^0)$$  \hspace{1cm} (1)

where $G^r \in G \ (r \in R)$ denotes the semantic role matrix under relation of $r$ and $R$ is the number of relation roles. $W^r$ represents the learnable weight matrix under relation $r$, $\rho$ is a ReLU function. To distinguish them, we term $E_i^0$ and the sum vector of $E_i^{r,0}$ as Textual Holistic High-level Semantic (TSH) representation and Discrete High-level Semantic (DTS) representation. Finally, TSH and TDS representations are fused by mean-polling to generate vector $TTS$, namely Textual High-level Semantic (TS) representation, which simultaneously encodes holistic and discrete semantics from three types of nodes.

### 3.3 Visual High-level Semantics Encoding

In this section, we elaborate on the details of visual high-level semantics encoding (VSE) module (see Figure. 3(b)), which consists of two parallel sub-modules to encode discrete and holistic high-level semantics into video representation, respectively.

**3.3.1 Visual Discrete Semantics Encoding.** Video is made up sequence of consecutive frames, in which each frame composes of multiple subjects and background objects. These entities can express most informative semantics from perspective of local information acquiring. Inspired by these observations, we take these entities as discrete semantics of video and propose to incorporate it into the video representation learning.

Given sampled video frames $O = \{o_1, \ldots, o_N\}$, for each sampled frame $o_i$, we adopt the bottom-up attention toolkit [1] to extract visual entities $E_i^v = \{E_i^{v,1}, E_i^{v,2}, \ldots E_i^{v,M}\} (i \in \{1, N\})$. Specifically, the visual entity $E^v$ includes three properties, including concept representation $C^v$, appearance representation $A^v$ and position representation $P^v$. Firstly, the concept property is made up of an object name $C_{o_i}^v$ and its decorated attribute $C_{a_i}^v$, which is completely homogeneous to the textual high-level semantics. Accordingly, we generate concept representation $C^v$ as follows:

$$C^v = MLP(S_o + S_a)$$  \hspace{1cm} (2)

where $MLP(\cdot)$ denotes a fully-connected layer.

Then, we introduce two additional information to be complementary for concept representation, i.e. appearance representation and position representation. As for appearance representation, we employ a fully-connected layer to project ROI (Region of Interest) feature into the joint space, obtaining the appearance representation $A^v$. Similarly, another fully-connected layer is used to produce position representation $P^v$ based on the spatial coordinate of the entity $E^v$. Afterwards, given three types of embeddings $\{C^v, A^v, P^v\}$, similar to TSE module, we also utilize graph convolution operation.
to encode high-level semantics for video, including three steps: 1) Initializing semantic relation graph node; 2) Graph building; 3) Graph Reasoning.

**Initializing semantic relation graph node.** Different with textual semantic role graph node, visual relation graph only contain one type of nodes, i.e. entity nodes. Given the entity set \( E_j \), \( j \in [1, N] \), \( i \in [1, M] \) extracted from video frame sequence, we only select top-K entities \( \{ E_{i1}, E_{i2}, ..., E_{ik} \} \) according to their appearing frequency in video. Then, the entity representation can be defined by aggregating its three types of embeddings as follows:

\[
E_x = \text{MLP} \left( \left[ C_u, (A^p + P^o) \right] \right)
\]

where \([\cdot]\) represents the concatenate operation. The output \( E_x \) serves as the initial node representation in graph.

**Relation graph building.** To further capture the associations between entities, we first build up an entity affinity graph. Concretely, given a set of visual entities \( E_j = \{ E_{i1}, E_{i2}, ..., E_{ik} \}, i \in [1, N] \), we measure the affinity between pairwise entities as follows:

\[
h(E_i, E_j) = \frac{\phi(E_i) \phi(E_j)^T}{\sqrt{D}}
\]

where \( \phi(E_i) = W_{\phi} E_i \) and \( \phi(E_j) = W_{\phi} E_j \) are two node embeddings, and \( \sqrt{D} \) is the dimension of graph nodes. \( W_{\phi} \) and \( W_{\phi} \) both represent embedding matrix. Then, we obtain a fully connected graph \( H = (V, E) \), where \( V \) is graph nodes initialized by visual entities and \( E \) denotes the affinity matrix \( H \) calculated from Eq. 4. \( h(\cdot, \cdot) \) indicates the strength of the affinity between two nodes in graph.

**Graph Reasoning.** Given the relation graph \( H \), we adopt one-layer Graph Convolutional Network (GCN) [19] with residue connection to promote the node embedding learning, which is defined as follows:

\[
E_x = \rho \left( H_0 E_x, W^0 + E_x \right)
\]

where \( W^0 \) is the weight matrix of the GCN layer and \( \rho \) is a ReLU function. Consequently, we can obtain the output node embedding \( E_x \), dubbed Visual Discrete High-level Semantic (VDS) representation.

### 3.3.2 Visual Holistic Semantics Encoding

Distinct from VDS encoding, extracting visual holistic semantics requires more comprehensive and generalized video understanding. Accordingly, the research target of video captioning [40, 55] is consistent with this goal. It refers to automatically generating natural language descriptions of videos, which is tightly associated with VTR task. Inspired by this observation, we attempt to transfer the prior knowledge restored in video captioning model to encode visual holistic semantics. In particular, we directly adopt the off-the-shelf video captioning model [23] to produce the natural language description for video. Given an output captioning sentence \( S^H \), we also use the CLIP text encoder described in section 3.1.2 to encode it. The generated feature vector \( V^{VHS} \) is termed as Visual Holistic Semantic (VHS) representation. Lastly, we aggregate VDS representation and VHS representation by mean-pooling, and thus obtain the final visual high-level semantic (VS) representation \( V^V \).

### 3.4 Cross-Modal Representations Fusion

Given the original video-text representation \( V(T) \) and high-level semantic representation \( V^V(T^T) \), we use a simple convex combination operation to aggregate them as the final cross-modal representations, which can be defined as:

\[
V_F = \alpha V + (1 - \alpha)V^V, \quad T_F = \alpha T + (1 - \alpha)T^T,
\]

where \( \alpha \) is a tuning parameter balancing two types of representations. And \( V_F \) and \( T_F \) respectively denote the fused video and text representations.

### 3.5 Training and Inference

As for training objective, we employ the hubness-aware contrastive loss (HAL) [25] for aligning video and text. Given video representation \( V = \{ V_1, ..., V_Q \} \) and text representation \( T = \{ T_1, ..., T_R \} \), loss function \( L_{HAL}(V, T) \) can be formulated as:

\[
\begin{align*}
L_{HAL} &= \frac{1}{Q} \sum_{q=1}^{Q} \left[ \log \sum_{r \neq q} \exp \left( \frac{(S_{qr} - \gamma)}{\mu} \right) + 1 \right] \\
&= \frac{1}{Q} \sum_{q=1}^{Q} \left[ \log \left( \sum_{r \neq q} \exp \left( \frac{(S_{qr} - 1)}{\mu} \right) + 1 \right) \right];
\end{align*}
\]

where \( \gamma \) is a margin parameter; \( \mu \) is a temperature parameter; \( N \) denotes the number of samples within the mini-batch; \( S_{qr} = \cos(V_q, T_r) \), \( S_{qq} = \cos(T_q, V_q) \), \( S_{qr} = \cos(V_q, T_q) \), and \( S_{rr} = \cos(T_r, T_r) \), where \( \cos(\cdot, \cdot) \) represents similarity function calculating cosine distance.

To enhance cross-modal learning, two types of HAL loss are utilized. First, it is imposed on mini-batch data. Secondly, it is imposed on anchor sample in mini-batch and item from modality-specific memory banks. Formally, the final objective of our HiSE model is defined as:

\[
L = \lambda_1 L_{HAL}(V_F, T_F) + \lambda_2 L_{HAL}(V, B_T) + \lambda_3 L_{HAL}(T, B_V)
\]

where balancing parameters are set to \( \lambda_1 = 10 \) and \( \lambda_2 = 0.1 \).

For inference, we use the cosine distance between fused representations \( V_F \) and \( T_F \) to measure the cross-modal relevance.

### 4 EXPERIMENTS

#### 4.1 Dataset and Settings

**4.1.1 Datasets.** We conduct experiments on three benchmark datasets for bidirectional video-text retrieval, including MSR-VTT [47], MSVD [7] and DiDeMo [2].

- **MSR-VTT** dataset contains 10K videos from YouTube website, with each video annotated with five 20 captions. We follow the 1k-A protocol in [50] and report the experimental results. Specifically, 1k-A protocol adopts 9,000 videos for training and utilizes rest 1,000 video-text pairs for testing.
- **MSVD** dataset includes 1,970 videos with approximately 80,000 captions. It is split into 1,200 training, 100 validation, and 100 testing videos. We report the performance of video-text retrieval on testing set with multiple captions per video.
- **DiDeMo** contains 10,000 videos, with each video annotated by 40 sentences. We follow [27] to conduct video-paragraph
Table 1: Comparisons of Experimental Results on MSR-VTT 1k-A Testing Set. ★: The results of [54] are reported by the model with video encoder of ViT-B/32 for fair comparison.

| Approach                        | Text-to-Video Retrieval | Video-to-Text Retrieval | R@Sum |
|--------------------------------|-------------------------|-------------------------|-------|
|                                | R@1 R@5 R@10 MdR       | R@1 R@5 R@10 MdR       |       |
| Non-CLIP Based                 |                         |                         |       |
| JSFusion[50]                   | 10.2 31.2 43.2 13.0     | - - - -                | -     |
| CE[27]                         | 20.9 48.8 62.4 6.0      | 20.6 50.3 64.0 5.3     | 267.0 |
| MMT[14]                        | 24.6 54.0 67.1 4.0      | 24.4 56.0 67.8 4.0     | 293.9 |
| Support-Set (pretrained)[34]   | 30.1 58.5 69.3 3.0      | 28.5 58.6 71.6 3.0     | 316.6 |
| HIT (pretrained)[26]           | 30.7 60.9 73.2 2.6      | 32.1 62.7 74.1 3.0     | 333.7 |
| FROZEN[5]                      | 31.0 59.5 70.5 3.0      | - - - -                | -     |
| CLIP Based                     |                         |                         |       |
| CLIP[35]                       | 31.2 53.7 64.2 4.0      | 27.2 51.7 62.6 5.0     | 290.6 |
| CLIP4Clip-meanP[28]            | 43.1 70.4 80.8 2.0      | 43.1 70.5 81.2 2.0     | 389.1 |
| CLIP4Clip-seqTransf[28]        | 44.5 71.4 81.6 2.0      | 42.7 70.9 80.6 2.0     | 391.7 |
| VCM[6]                         | 43.8 71.0 80.9 2.0      | 45.1 72.3 82.3 2.0     | 395.4 |
| CenterCLIP [54] ★             | 44.2 71.6 82.1 2.0      | 42.8 71.7 82.2 2.0     | 394.6 |
| HiSE                           | 45.0 72.7 81.3 2.0      | 46.6 73.3 82.3 2.0     | 401.2 |

Table 2: Comparisons of Experimental Results on MSVD Testing Set.

| Approach                        | Text-to-Video Retrieval | Video-to-Text Retrieval | R@Sum |
|--------------------------------|-------------------------|-------------------------|-------|
|                                | R@1 R@5 R@10 MdR       | R@1 R@5 R@10 MdR       |       |
| Non-CLIP Based                 |                         |                         |       |
| VSE[33]                        | 12.3 30.1 42.3 14       | 34.7 59.9 70.0 33      | 249.3 |
| CE[27]                         | 19.8 49.0 63.8 -        | - - - -                | -     |
| MoEE[31]                       | 21.1 52.0 66.7 5.0      | 27.3 55.1 65.0 4.3     | 287.2 |
| TT-CE+[9]                      | 25.4 56.9 71.3 4.0      | 27.1 55.3 67.1 4.0     | 303.1 |
| Support-Set (pretrained)[34]   | 28.4 60.0 72.9 4.0      | 34.7 59.9 70.0 3.0     | 325.9 |
| FROZEN[5]                      | 31.0 59.5 70.5 3.0      | - - - -                | -     |
| CLIP Based                     |                         |                         |       |
| CLIP[35]                       | 37.0 64.1 73.8 3.0      | 59.9 85.2 90.7 1.0     | 410.7 |
| CLIP4Clip-meanP[28]            | 46.2 76.1 84.6 2.0      | 56.6 79.7 84.3 1.0     | 427.5 |
| CLIP4Clip-seqTransf[28]        | 45.2 75.5 84.3 2.0      | 62.0 87.3 92.6 1.0     | 446.9 |
| HiSE                           | 45.9 76.2 84.6 2.0      | 66.3 90.7 95.2 1.0     | 458.9 |

retrieval, where all the descriptions of one video are combined into a single text.

4.1.2 Evaluation Metrics. For evaluation, we employ three standard retrieval criteria: recall at rank $K$ ($R@K$, higher is better), median rank (MdR, lower is better) and sum of recall ($R@sum$, higher is better). In particular, $R@K$ measures the fraction of queries for which the matched item is found among the top $K$ retrieved results. The MdR denotes the median rank of correct items in the retrieved ranking list. Besides, $R@sum$ criterion is calculated by summing all metrics of $R@K$, which can better reflect the overall performance.

4.2 Implementation Details

For text encoding, the basic text transformer and occurrence node encoder in THS module are both initialized by CLIP text transformer. For video encoding, the spatial transformer (ViT) is initialized with CLIP (ViT/B-32). The dimension of the joint embedding space is set to 512. The caption token length is 32 and frame length is 12. Note that on DiDeMo dataset, the captioning sentences are contacted as one paragraph for video-paragraph retrieval, where the caption token length is et to 64. The size of memory banks is set to 4096 and the momentum coefficient is equal to 0.995. The fusion ratio parameter $\alpha$ in Eq.6 is empirically set to 0.9. In alignment objective, we follow [25] to set $\gamma = 0.3$ and $\mu = 0.1$ in Eq.7. Our model is fine-tuned by Adam optimizer with mini-batch size of 256. As for the learning rate, we follow the CLIP [36] to decay it with a cosine schedule. The initial learning rate is 1e-7 for basic text encoder and video encoder and 1e-4 for other modules. All our experiments are implemented on 8 NVIDIA Tesla P40 GPUs.

4.3 Comparison to State-of-the-art Methods

The experimental results on MSR-VTT and MSVD datasets are listed in Table 4 and Table 5, respectively. Note that the results on
Table 3: Performance of our HiSE method with different representation components on MSR-VTT 1k-A test set. Visual Discrete and Holistic High-level Semantics are abbreviated as “VDS” and “VHS”, respectively. Textual Discrete and Holistic Semantics are abbreviated as “TDS” and “THS”, respectively.

| High-level Semantics Encoding Module | Text-to-Video Retrieval | Video-to-Text Retrieval |
|-------------------------------------|-------------------------|-------------------------|
|                                     | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VDS                                | 43.6 | 72.1 | 80.9 | 44.5 | 72.2 | 82.0 |
| VHS                                |     |     |     |     |     |     |
| TDS                                | ✓   |     |     | 44.5 | 72.3 | 81.3 |
| THS                                | ✓   | ✓   | ✓   | 44.6 | 72.2 | 81.3 |
| VDS                                | ✓   | ✓   | ✓   | 44.7 | 72.3 | 81.3 |
| VHS                                |     |     |     |     |     |     |
| TDS                                | ✓   | ✓   | ✓   | 45.4 | 72.5 | 81.4 |
| THS                                | ✓   | ✓   | ✓   | 46.0 | 72.8 | 81.9 |
| VDS                                | ✓   | ✓   | ✓   | 44.9 | 72.4 | 81.2 |
| VHS                                |     |     |     |     |     |     |
| TDS                                | ✓   | ✓   | ✓   | 45.4 | 72.5 | 81.4 |
| THS                                | ✓   | ✓   | ✓   | 46.4 | 73.1 | 82.2 |
| VDS                                | ✓   | ✓   | ✓   | 45.0 | 72.8 | 81.3 |
| VHS                                |     |     |     |     |     |     |
| TDS                                | ✓   | ✓   | ✓   | 46.6 | 73.3 | 82.3 |
| THS                                | ✓   | ✓   | ✓   | 46.6 | 73.3 | 82.3 |

Figure 4: Visualization results of video discrete and holistic high-level semantics extracted on MSR-VTT dataset. The listed ground truth sentences are randomly selected from MSR-VTT captions.

DiDeMo dataset is placed in the supplementary materials due to limited space.

On MSR-VTT dataset, as shown in Table 4, we can see our HiSE outperforms the competitors in most evaluation metrics. For text retrieval, compared with the second best method, we achieve absolute boost (1.5%, 1.0%) on (R@1, R@5). For video retrieval, although R@1 of our method is slightly lower than that of [13], the summation of all our three criteria still outperforms it by 0.8%. Moreover, as the most comprehensive criteria, the R@sum of our model obviously surpasses other algorithms, which achieves 5.8% improvement in comparison to the best competitor.

The results on the MSVD dataset are presented in Table 5. It can be seen that our HiSE arrives at 458.9% on the criteria of “R@Sum”, which outperforms the second best method by 15.5%. Especially for text retrieval, the HiSE model surpasses the previous best method by (4.3%, 3.4%, 2.6%) on (R@1, R@5, R@10), respectively. These results substantially demonstrate the advance of our method. Moreover, it can be observed that the text retrieval performance is better than video retrieval, we conjecture the possible reason is that the high-level semantics can provide scarce and complementary information for visual modality. By contrast, it is naturally homogeneous to the textual modality, thus the improvement is not that striking.

4.4 Ablation Studies

In this section, we conduct a series of ablation experiments to explore the impacts of different components in our HiSE model. Note that all results are reported on MSR-VTT 1k-A test set.

4.4.1 Impact of Discrete High-level Semantics.

To begin with, we investigate in how the incorporation of discrete high-level semantics affects the performance of HiSE model. From Table 3, we can see that when TDS representation is employed, our HiSE can obtain 0.2% and 0.2% performance gain on R@1 for video retrieval and text retrieval, respectively. Besides, comparing line #3 with line #6, adopting VDS representation will further bring in 0.9%
The holistic high-level semantics really contributes to improving video-text alignment. 2) The complementarity existing between ground truth (GT) caption of video in first line is "a man walking down street in front of an official building". As for discrete semantics, the predicting phrases, such as "walking man" and "white building", can be properly aligned with the GT sentence. Moreover, the generated holistic semantics "a man walking on the street" is also very consistent with the GT one. These results can reflect additional interpretability conveyed by our model.

4.4.2 Impact of Holistic High-level Semantics. Here, we explore the impact of holistic high-level semantics extraction. From Table 3, comparing line #3 with line #4, we can see the VHS module collaborates well with THS module, which can bring about 0.2% performance boost on R@1 for video retrieval and 0.3% performance boost on R@1 for text retrieval, respectively. Moreover, comparing line #6 with line #8, it can be observed that the combination of all four types of high-level semantic representations will lead to the best performance. These results can verify two main points: 1) The holistic high-level semantics really contributes to improving video-text alignment. 2) The complementarity existing between discrete and holistic high-level semantics also matters for acquiring performance improvements.

4.4.3 Bidirectional Cross-Modal Retrieval Results. We compare the bidirectional VTR results obtained by different models. From Figure 5, it can be seen that the retrieval results listed in the left column is superior than those in the right column. These results further validate the adoption of explicit high-level semantics really contributes to improving the visual-linguistic embeddings, returning more reasonable retrieval results.

5 CONCLUSIONS

The ambiguous understanding of videos and texts impedes the ability of machine to build accurate cross-modal association. In this work, we proposed a explicit high-level semantics (HiSE) aided visual-linguistic embedding model for improving video-text retrieval. Particularly, we study how to mine explicit high-level semantics from both texts and videos, and incorporate them into cross-modal representations learning. Doing so allows us to disentangle explainable information from raw data that supplies complementary knowledge for the traditional visual-linguistic aligning framework. The experiments conducted on three benchmark datasets validate our method achieves superior performance over the state-of-the-art solutions.
[51] Alireza Zareian, S. Karaman, and Shih-Fu Chang. 2020. Bridging Knowledge Graphs to Generate Scene Graphs. In European Conference on Computer Vision. 606–623.

[52] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. 2019. Lookahead optimizer: k steps forward, 1 step back. In Advances in neural information processing systems.

[53] Rui Zhao, Kecheng Zheng, Zheng-Jun Zha, Hongtao Xie, and Jiebo Luo. 2021. Memory Enhanced Embedding Learning for Cross-Modal Video-Text Retrieval. arXiv preprint arXiv:2103.15686 (2021).

[54] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. CenterCLIP: Token Clustering for Efficient Text-Video Retrieval. In SIGIR.

[55] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. 2018. End-to-End Dense Video Captioning with Masked Transformer. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8739–8748.
A OVERVIEW OF APPENDIX

In this appendix, we give some details which were omitted in the main body of our manuscript owing to the limited space. Concerning the experiments, we report the experimental results on DiDeMo dataset, impact of representation fusion hyper-parameter $\alpha$, impact of different alignment objectives, and data distribution visualization result in joint embedding space.

B EXPERIMENTS

B.1 Experimental Results on DiDeMo dataset

The experimental results on DiDeMo dataset are presented in Table 4. It can be seen that, in comparison to the method with the same raw video encoder as ours, i.e., CLIP4Clip-seqTransf, the overall retrieval quality reflected by the R@Sum metric is increased by a large margin (+ 8.5 %). Moreover, compared to the best competitor CLIP4Clip-meanP, our HiSE obtains 0.7 and 1.3 % performance gains on R@1 metric for video retrieval and text retrieval, respectively. We believe the major improvement derives from the exploitation of the introducing of explicit high-level semantics, which supplies more complementary information to narrow the modality gap.

B.2 Ablation Studies

Consistent with the main body of our paper, all our ablation experiments are conducted on MSR-VTT 1k-A test set.

B.2.1 Impact of Graph Reasoning in High-level Semantics Representation. In this part, we explore the effect of graph reasoning modules in high-level semantics representation components. As shown in Table 5, for visual semantics representation, when we replace mean-pooling with visual graph reasoning module, our model can obtain 0.1% and 0.4% performance gain on R@1 for video retrieval and text retrieval, respectively. As for textual branch, compared to adopting mean-pooling, employing textual graph reasoning to aggregate hierarchical information can result in 0.3% boost for video retrieval and 0.1% boost for text retrieval. Furthermore, the deployment of both visual and textual graph reasoning can lead to the best performance. These results are consistent with our designing aim of these modules, which leverage the graph reasoning technique to promote the interaction between hierarchies of high-level semantics.

B.2.2 Impact of Representation Fusion Parameter. To explore the impact of parameter $\alpha$ of fusing raw video-text representation $V(T)$ and high-level semantic representation $V^{VS}(V^{TS})$. In Figure 6, we can see that the bidirectional retrieval performances both decrease when $\alpha$ varies from 0.9 to 1. Considering $\alpha=1$ indicates the high-level semantics is removed from the video-text representation, these results validate the effectiveness of introducing high-level semantics to improve cross-modal discrimination.

B.2.3 Impact of Different Alignment Objectives. In this part, we analyze the impact of different alignment objectives in our HiSE method. Specifically, in Table 6, we present the retrieval result of HiSE replacing the M-HAL loss with the prevailing Bi-directional InfoNCE (B-InfoNCE) loss [28, 36] in recent comparison studies. From Table 6, it can be seen that although using the same objective, our model still outperforms the second best competitor by 0.6% improvement for video retrieval and 1.4% improvement for text retrieval, respectively. It further validates the advance of our proposed explicit high-level semantics encoding modules. Besides, the two items in M-HAL loss can both bring about respective performance gain. These results confirm the effectiveness and rationality of our employed M-HAL loss for video-text retrieval.

B.3 Qualitative Analysis

B.3.1 T-SNE Visualization of Video-Text Representations. To further investigate how the explicit high-level semantics affects the learned joint embedding space, we adopt t-SNE [30] to visualize the learned cross-modal representations from MSR-VTT 1k-A test set, containing 1000 images and 1000 texts. In particular, the data distribution of the baseline model without utilizing explicit semantics and our full HiSE model are depicted in Figure 7(a) and Figure 7(b). In comparison to the former, it can be observed the distributions of videos and texts are further mixed by our proposed HiSE method. These results indicate that the proposed explicit semantics incorporating method contributes to reducing the distribution difference between two modalities.
Table 4: Comparisons of Experimental Results on DiDeMo Testing Set.

| Approach          | Text-to-Video Retrieval | Video-to-Text Retrieval | R@Sum |
|-------------------|-------------------------|-------------------------|-------|
|                   | R@1 | R@5 | R@10 | MdR | R@1 | R@5 | R@10 | MdR |       |
| Non-CLIP Based    |     |     |      |     |     |     |      |     |       |
| S2VT[44]          | 11.9 | 33.6 | -    | 13.0 | 13.2 | 33.6 | -    | 15.0 |       |
| FSE[52]           | 13.9 | 36.0 | -    | 11.0 | 13.1 | 33.9 | -    | 12.0 |       |
| CE[27]            | 16.1 | 41.1 | -    | 8.3  | 15.6 | 40.9 | -    | 8.2  |       |
| TT-CE+[9]         | 21.6 | 48.6 | 62.9 | 6.0  | 21.1 | 47.3 | 61.1 | 6.3  | 262.6 |
| MoEE[32]          | 16.1 | 41.2 | 55.2 | 8.3  | 16.0 | 41.7 | 54.6 | 8.7  | 224.8 |
| Frozen[5]         | 34.6 | 65.0 | 74.7 | 3.0  | -    | -    | -    | -    |       |
| CLIP Based        |     |     |      |     |     |     |      |     |       |
| CLIP4Clip-seqTransf[28] | 42.8 | 68.5 | 79.2 | 2.0  | 41.4 | 68.2 | 79.1 | 2.0  | 379.2 |
| CLIP4Clip-meanP[28] | 43.4 | 70.2 | 80.6 | 2.0  | 42.5 | 70.6 | 80.2 | 2.0  | 387.5 |
| HiSE              | 44.1 | 69.9 | 80.3 | 2.0  | 43.8 | 70.4 | 79.2 | 2.0  | 387.7 |

Table 5: Performance of our HiSE method with different high-level semantics aggregating modules on MSR-VTT 1k-A test set. Visual Graph Reasoning and Textual Graph Reasoning are abbreviated as “VGR” and “TGR”, respectively. Mean-pooling operation for visual and textual modality are abbreviated as “VMP” and “TMP”, respectively.

| High-level Semantics Aggregating Module | Text-to-Video Retrieval | Video-to-Text Retrieval |
|----------------------------------------|-------------------------|-------------------------|
|                                        | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VGR                                    |     |     |      |     |     |      |
| TGR                                    | -   | ✓   | ✓    | 44.5 | 72.2 | 81.0  | 46.0 | 73.0 | 82.1  |
| VMP                                    | ✓   | -   | ✓    | 44.6 | 72.3 | 81.1  | 46.4 | 73.2 | 82.2  |
| TMP                                    | -   | ✓   | ✓    | 44.8 | 72.5 | 81.3  | 46.1 | 73.1 | 82.1  |
| VGR                                    | ✓   | ✓   | -    | 45.0 | 72.7 | 81.3  | 46.6 | 73.3 | 82.3  |

Table 6: Performance of our HiSE with different alignment objectives on MSR-VTT 1k-A test set. Coupled Memory Banks are abbreviated as “CMB”.

| Approach | Alignment Objective | Components | Text-to-Video Retrieval | Video-to-Text Retrieval |
|----------|---------------------|------------|-------------------------|-------------------------|
|          |         |           | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| CLIP[35] | B-InfoNCE          | CMB HAL   | 31.2 | 53.7 | 64.2 | 27.2 | 51.7 | 62.6 |
| CLIP4Clip-meanP[28] | B-InfoNCE          | - -       | 43.1 | 70.4 | 80.8 | 43.1 | 70.5 | 81.2 |
| CLIP4Clip-seqTransf[28] | B-InfoNCE          | - -       | 44.5 | 71.4 | 81.6 | 42.7 | 70.9 | 80.6 |
| HiSE     | B-InfoNCE          | - -       | 45.1 | 72.1 | 80.9 | 44.1 | 72.8 | 81.5 |
| HiSE     | M-HAL              | - ✓        | 44.6 | 72.5 | 81.1 | 46.3 | 73.0 | 82.2 |
| HiSE     | M-HAL              | ✓ -        | 45.0 | 72.7 | 81.3 | 46.6 | 73.3 | 82.3 |