Hierarchical Banzhaf Interaction for General Video-Language Representation Learning

Peng Jin ⓘ, Hao Li ⓘ, Li Yuan ⓘ, Shuicheng Yan ⓘ, *Fellow, IEEE*, and Jie Chen ⓘ

Abstract—Multimodal representation learning, with contrastive learning, plays an important role in the artificial intelligence domain. As an important subfield, video-language representation learning focuses on learning representations using global semantic interactions between pre-defined video-text pairs. However, to enhance and refine such coarse-grained global interactions, more detailed interactions are necessary for fine-grained multimodal learning. In this study, we introduce a new approach that models video-text as game players using multivariate cooperative game theory to handle uncertainty during fine-grained semantic interactions with diverse granularity, flexible combination, and vague intensity. Specifically, we design the Hierarchical Banzhaf Interaction to simulate the fine-grained correspondence between video clips and textual words from hierarchical perspectives. Furthermore, to mitigate the bias in calculations within Banzhaf Interaction, we propose reconstructing the representation through a fusion of single-modal and cross-modal components. This reconstructed representation ensures fine granularity comparable to that of the single-modal representation, while also preserving the adaptive encoding characteristics of cross-modal representation. Additionally, we extend our original structure into a flexible encoder-decoder framework, enabling the model to adapt to various downstream tasks. Extensive experiments on commonly used text-video retrieval, video-question answering, and video captioning benchmarks, with superior performance, validate the effectiveness and generalization of our method. The code is available at https://github.com/jpthu17/HBI.

Index Terms—Video-Language Representation Learning, Text-Video Retrieval, Video Question Answering, Video Captioning.

✦

1 INTRODUCTION

Multimodal representation learning, which aims to narrow the gap among different modalities, is crucial for making the most of multimodal data [1]. As an essential subfield, video-language representation learning aims to understand the relationship between videos and their associated textual descriptions. It is beneficial for various downstream tasks including text-video retrieval, video-question answering (VideoQA), and video captioning. Recently, an increasing number of video-language representation learning models [2]–[4] apply contrastive learning to project the video and text features into a common latent space based on the semantic similarities of video-text pairs. In this manner, multimodal contrastive learning enables networks to learn discriminative video-language representations.

Lacking fine-grained alignment annotations, the video-language contrastive learning models [5]–[8] typically perform coarse-grained feature alignment based on the global similarity between the video and the text. As shown in Fig. 1(a), previous contrastive models only exploit the coarse-grained labels of video-text pairs to learn a global semantic interaction. However, for multimodal downstream tasks like text-video retrieval and VideoQA, models must

- *Peng Jin and Hao Li contributed equally to this work.*
- *Peng Jin and Hao Li are with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China, and also with the Pengcheng Laboratory, Shenzhen, China. E-mail: jp21@stu.pku.edu.cn, lihao1984@pku.edu.cn.*
- *Li Yuan and Jie Chen are with the School of Electronic and Computer Engineering, Peking University, Shenzhen, China, and also with the Pengcheng Laboratory, Shenzhen, China. E-mail: yuanli-ece@pku.edu.cn, chenj@pcl.ac.cn.*
- *Shuicheng Yan is with Skywork AI and Nanyang Technological University. E-mail: shuicheng.yan@gmail.com.*
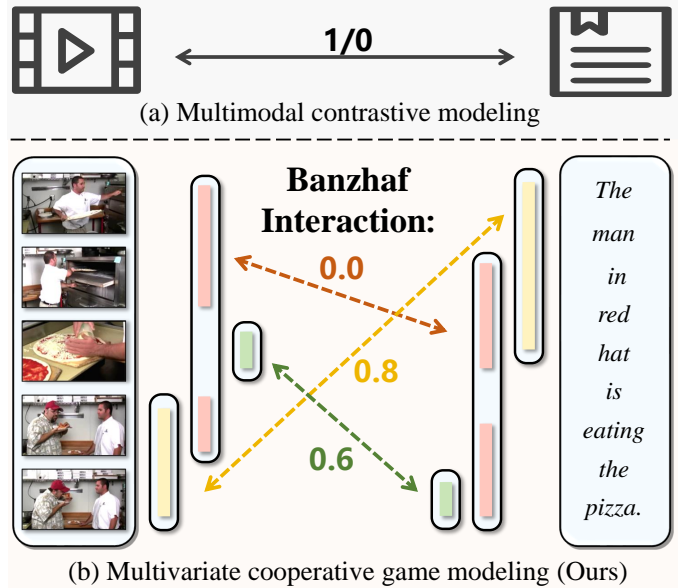- *Corresponding authors: Li Yuan, Jie Chen.*

Fig. 1: (a) Previous methods only learn a global semantic interaction from the coarse-grained labels of video-text pairs. (b) We model multimodal alignment as a cooperative game process, utilizing Banzhaf Interaction to evaluate possible correspondence between video frames and text words.

possess the capability to capture detailed and interpretable features. The coarse-grained text-video alignment hinders the interaction process between the visual entity and the textual phrase. Establishing fine-grained alignment through naive methods would require high-quality annotated data to explore clear video-text relationships, which is currently
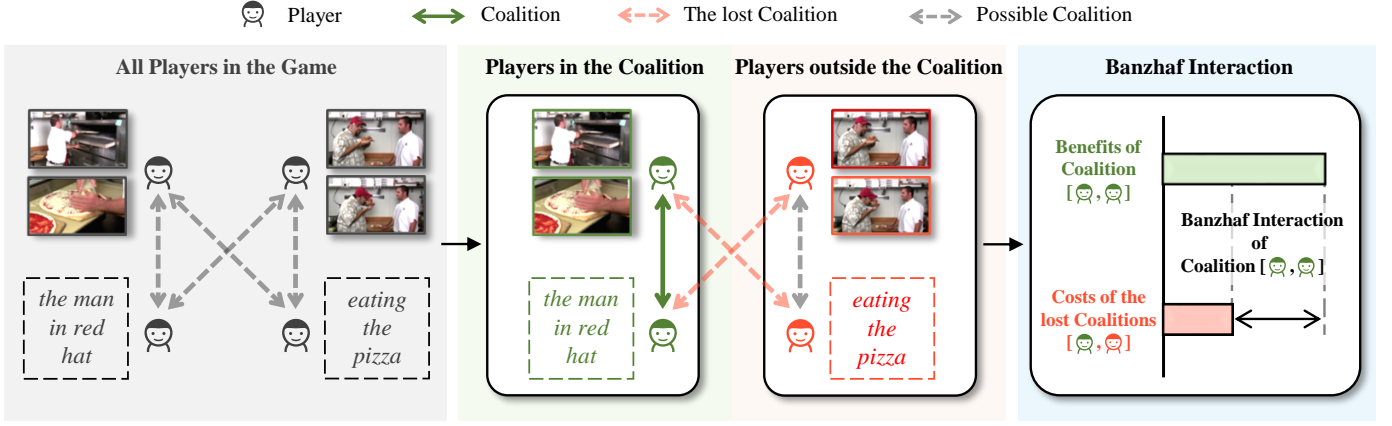
Fig. 2: **The intuition of employing Banzhaf Interaction in video-language representation learning.** When certain players (frames and words) form a coalition, it entails the exclusion of these players from potential coalitions with others, rendering them mutually exclusive from the target coalition. Banzhaf Interaction quantifies the disparity between the benefits derived from the coalition and the costs incurred due to the lost coalitions. Therefore, Banzhaf Interaction effectively captures the incremental benefits conferred by the coalition. We refer the reader to Eq. 1 for the detailed formula.

unavailable, particularly on large-scale vision-language datasets. This underscores the possibility of other learning signals to improve typical contrastive learning.

To address the above fine-grained alignment problem, we propose an innovative approach to model video-language representation learning as a multivariate cooperative game by formulating video and text as players in a cooperative game, as shown in Fig. 1(b). Intuitively, when visual and textual representations exhibit high semantic similarity, they are more likely to collaborate and enhance the overall multimodal similarity score. Guided by this insight, we regard the collection of multiple representations as a coalition and suggest using the game-theoretic interaction index, *i.e.,* Banzhaf Interaction [9], to measure the degree of cooperation within a coalition. The Banzhaf Interaction [10] is a well-known concept in cooperative game theory. As shown in Fig. 2, it quantifies the additional benefits accruing to a coalition compared to the costs of lost coalitions with other players. A high Banzhaf Interaction of a coalition indicates its greater contribution to the semantic similarity. Hence, it can help evaluate the potential correspondence between video frames and text words, making cross-modal contrastive learning sensitive and explainable.

In our preliminary conference paper [11], we propose Hierarchical Banzhaf Interaction (HBI), which treats video frames and text words as players, leveraging cross-modality similarity measurement as the characteristic function in the cooperative game. Specifically, HBI utilizes Banzhaf Interaction to delineate the trend of cooperation among any set of features. Moreover, to efficiently establish coalitions among game players, HBI introduces a token merge module to cluster the original frames (words) and reduce the number of players. By stacking token merge modules, HBI achieves hierarchical interaction, encompassing entity-level interactions among frames and words, action-level interactions among clips and phrases, and event-level interactions among segments and paragraphs.

Although our HBI framework has shown noticeable improvements, its current approach to game interactions solely relies on independently encoded video and text represen-

tations. This makes it hard to accurately calculate Banzhaf Interaction. Specifically, videos may contain redundant information, as illustrated in Fig. 2, which can have a bias in the calculations for Banzhaf Interaction. Employing cross-modal joint encoding helps select important video content based on the query text, reducing the bias in Banzhaf Interaction. However, since the cross-modal representation can only encode the entire video and text, it cannot be directly applied to our fine-grained game interaction modeling. What is worse, when semantic alignment between text and video is lacking, this text-conditioned representation might not only fail to enhance single-modal representations but may also perform worse than single-modal representations.

To this end, we significantly extend our previously proposed HBI framework to introduce HBI V2, which combines the strengths of both single-modal and cross-modal representations. In HBI V2, we reconstruct the representation by integrating single-modal and cross-modal components. This reconstructed representation dynamically adjusts the weights of single-modal and cross-modal components, ensuring fine granularity comparable to that of the single-modal representation to appropriate our aim of fine-grained game interaction modeling while preserving the adaptability of cross-modal representation.

Furthermore, our previously proposed HBI framework primarily addresses one task, *i.e.,* text-video retrieval. To enhance the versatility of HBI V2 across various downstream tasks, including VideoQA and video captioning, we expand our original structure into a flexible encoder-decoder framework comprising an encoder and a task-specific decoder. We conduct comprehensive experiments on multiple widely used datasets, including three text-video retrieval datasets (*MSRVTT* [12], *ActivityNet Captions* [13], *DiDeMo* [14]), three VideoQA datasets (*MSRVTT-QA* [15], *MSVD-QA* [15], *ActivityNet-QA* [16]), and one video captioning dataset (*MSRVTT* [12]). As shown in Fig. 3, our proposed HBI V2 consistently surpasses both the previous HBI and existing methods across all downstream tasks. These results demonstrate the effectiveness of our proposed HBI V2 framework. Our main contributions are as follows:

Text-to-Video R@1 on the MSRVTT

49.4	HBI V2 (Ours)
48.6	HBI *CVPR 2023*
47.8	Prompt Switch *ICCV 2023*
47.5	UATVR *ICCV 2023*
47.0	DGL *AAAI 2024*

Video-to-Text R@1 on the MSRVTT

47.7	HBI V2 (Ours)
46.9	UATVR *ICCV 2023*
46.8	HBI *CVPR 2023*
46.0	Prompt Switch *ICCV 2023*
44.9	DGL *AAAI 2024*

Answer Accuracy on the MSVD-QA

53.4	HBI V2 (Ours)
52.5	TG-VQA *IJCAI 2023*
52.1	HBI *CVPR 2023*
48.6	Co-Tok *ECCV 2022*
48.3	All in One *CVPR 2023*

CIDEr on the MSRVTT

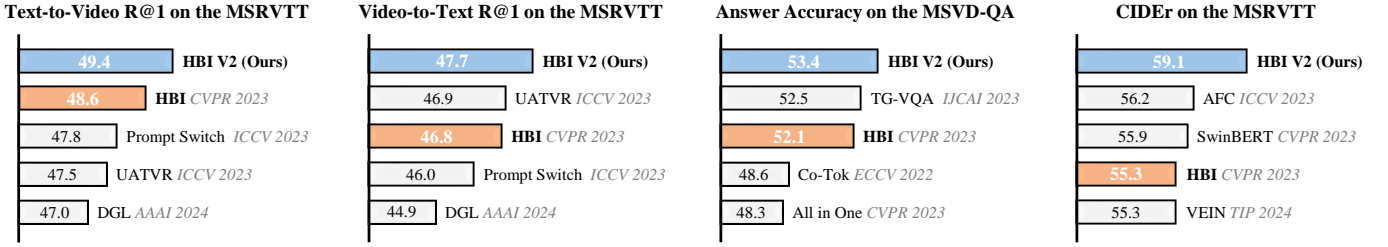|	
59.1	HBI V2 (Ours)
56.2	AFC *ICCV 2023*
55.9	SwinBERT *CVPR 2023*
55.3	HBI *CVPR 2023*
55.3	VEIN *TIP 2024*

Fig. 3: **Performance comparisons on text-video retrieval, video-question answering, and video captioning.** Our proposed framework, HBI V2, designed for general video-language representation learning, demonstrates superior performance consistently. Notably, HBI V2 not only surpasses the previous HBI, but also outperforms existing task-specific methods.

- To the best of our knowledge, we are the first to introduce the multivariate cooperative game process into fine-grained video-language learning.
- To mitigate the bias in Banzhaf Interaction, we propose to reconstruct the representation as a fusion of single-modal and cross-modal components. This reconstructed representation retains the advantages of both fine-grained and adaptive encoding.
- With task-specific prediction heads, HBI V2 achieves impressive performance on various tasks of video-language learning, including text-video retrieval, video question answering, and video captioning.

2 RELATED WORK

2.1 Visual-Language Learning

Recently, contrastive learning methods show great success in cross-modal tasks [17]–[24], such as text-video retrieval, VideoQA, and video captioning. The main challenge of cross-modal learning is to use vision-language pairs to learn common representations shared between modalities.

Text-Video Retrieval. Text-video retrieval requires models to establish the correct match between texts and videos. Most works [8], [25], [26] of text-video retrieval map text and video to the same semantic space, where the similarity between text features and video features can be directly calculated. Recently, contrastive learning methods have shown great success in advancing the state-of-the-art performance of the retrieval task. Contrastive learning methods [3], [27], [28] try to learn data representations from positive and negative pairs, making the representations of positive pairs have high similarity, and negative pairs have low similarity. Due to manually labeling the fine-grained relationships being unavailable, contrastive learning cannot capture fine-grained information in a supervised manner. To this end, we model video-text as game players with cooperative game theory and propose to combine Banzhaf Interaction with multimodal representation learning.

Video Question Answering. VideoQA [29]–[31] requires models to analyze the complex semantic correlation between the video and the question. Recently, several contrastive learning-based VideoQA models [32], [33] use the contrastive loss for cross-modality explicit alignment and fusion. With the contrastive loss, VideoQA models successfully map the different modalities into the same latent space. However, VideoQA requires element-level matching between text entities and video clips to predict the correct answer. Lacking the element-level fine-grained matching

annotations, existing contrastive-learning-based VideoQA models suffer from slow convergence, requiring massive data [34], [35] for pretraining. In contrast to prior works, we explicitly capture the fine-grained semantic relationships between video and text via Banzhaf Interaction.

Video Captioning. Video captioning requires models to automatically describe videos using natural language sentences. Existing methods [36], [37] adopt the encoder-decoder architecture to generate captions flexibly from the video features. Recently, contrastive-learning-based video captioning models [38], [39] align the video clips and textual entities to enhance the quality of the generated video descriptions. However, similar to retrieval and VideoQA, existing contrastive-learning-based models fail to achieve element-level concept matching, restricting the caption quality. In this work, we demonstrate that the proposed HBI V2 also achieves competitive performance for video captioning.

2.2 Cooperative Game Theory

The game-theoretic interaction [40] consists of a set of players with a characteristic function. The characteristic function maps each team of players to a real number which indicates the payoff obtained by all players working together to complete the task. The core of game-theoretic interaction is to allocate different payoffs to game individuals fairly and reasonably. There are several interaction strategies including core interaction [41], Shapley interaction [42], and Banzhaf interaction [10]. The game-theoretic interaction has multiple applications in different fields [43], [44].

Toward the goal of achieving fine-grained alignment between video and text, we propose to leverage multi-player game theory to construct an alignment label generator. We aim to obtain the semantic relationship between visual tokens and text tokens. And the game theory targets generating an appropriate coalition construction strategy for multiple players. Thus, we propose to introduce game theory to help the alignment label generation by considering the video clips and the text entities as players.

The cooperative game theory consists of a set $\mathcal{N} = \{1, 2, ..., n\}$ of players with a characteristic function ϕ. The characteristic function ϕ maps each team of players to a real number. This number indicates the payoff obtained by all players working together to complete the task. The core of the cooperative game theory is calculating how much gain is obtained and how to distribute the total gain fairly [42].

In a cooperative game, some players tend to form a coalition: it may happen that $\phi(\{i\})$ and $\phi(\{j\})$ are small,
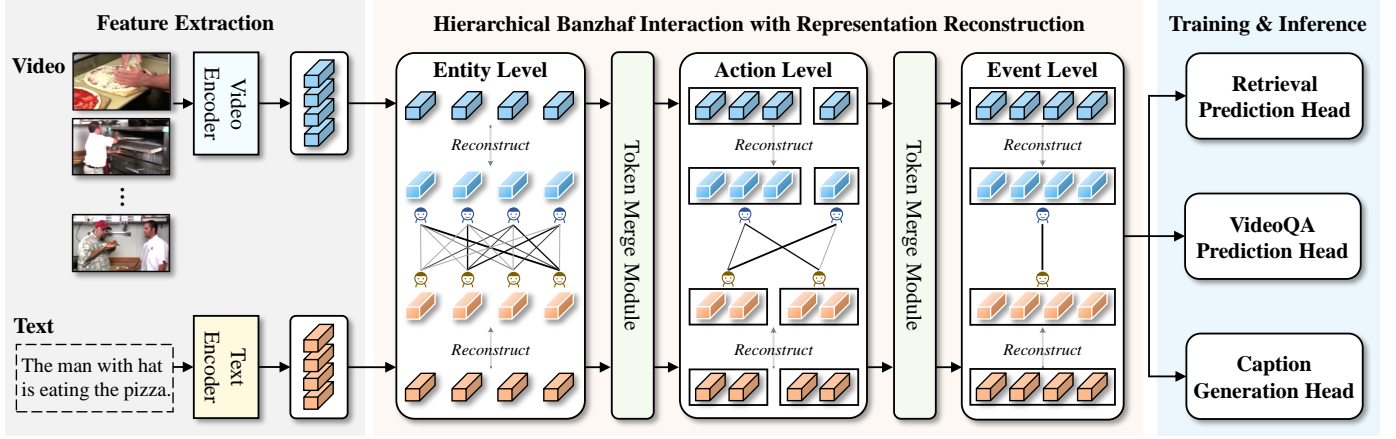
Fig. 4: **Overview of our proposed HBI V2 framework.** We employ a dual-stream encoder to extract features for video tokens and text tokens. Subsequently, we reconstruct the original representation by merging single-modal and cross-modal components. We propose a novel proxy training objective, which uses Banzhaf Interaction to evaluate possible correspondence between video tokens and text tokens from various levels. Furthermore, we customize different task-specific prediction heads for various downstream tasks.

and at the same time $\phi(\{i,j\})$ is large. The Banzhaf Interaction [9] measures the additional benefits brought by the target coalition compared with the costs of the lost coalitions of these players with others. The costs of the lost coalitions can be estimated by each player in the target coalition working individually. For a coalition $\{i,j\}$, we consider $[\{i,j\}]$ as a single hypothetical player, which is the union of the players in $\{i,j\}$. Subsequently, the reduced cooperative game is formed by removing the individual players in $\{i,j\}$ from the game and adding $[\{i,j\}]$ to the game.

Definition 1. Banzhaf Interaction. Given a coalition $\{i,j\} \subseteq \mathcal{N}$, the Banzhaf Interaction $\mathcal{I}([\{i,j\}])$ for the player $[\{i,j\}]$ is defined as:

$$\mathcal{I}([\{i,j\}]) = \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i,j\}} p(\mathcal{C}) \Big[\phi(\mathcal{C} \cup \{[\{i,j\}]\}) +$$
$$\phi(\mathcal{C}) - \phi(\mathcal{C} \cup \{i\}) - \phi(\mathcal{C} \cup \{j\}) \Big], \quad (1)$$

where $p(\mathcal{C}) = \frac{1}{2^{n-2}}$ is the likelihood of $\mathcal{C}$ being sampled. "$\mathcal{N} \setminus \{i,j\}$" denotes removing $\{i,j\}$ from $\mathcal{N}$.

Intuitively, $\mathcal{I}([\{i,j\}])$ reflects the tendency of interactions inside $\{i,j\}$. The higher value of $\mathcal{I}([\{i,j\}])$ indicates that player i and player j cooperate closely with each other.

3 METHODOLOGY

Our model consists of three submodules: (1) **Representation Reconstruction**, which merges single-modal and cross-modal representations. (2) **Hierarchical Banzhaf Interaction Module**, establishing multi-level Banzhaf Interaction between video and text by modeling them as game players. (3) **Task-Specific Prediction Heads**, designed for various tasks. Fig. 4 shows the overview of our proposed HBI V2.

3.1 Representation Reconstruction

Following previous works [8], [45], we employ CLIP [27] as the backbone for generating both visual and textural representations. Specifically, for video representation, we evenly sample frames from the video to form the input frame

sequence. Subsequently, we leverage ViT [46] to encode the frame sequence and adapt the output from the [CLS] token as the frame embedding. Following this, we employ a 4-layer transformer to aggregate the embedding of all frames and obtain the frame representation $\boldsymbol{V}_f^s = \{\hat{v}_f^i\}_{i=1}^{N_v}$, where $\hat{v}_f^i$ signifies a frame feature and N_v denotes the total number of frames. For text representation, we utilize the text encoder of CLIP to obtain the text representation $\boldsymbol{T}_w^s = \{\hat{t}_w^j\}_{j=1}^{N_t}$, where N_t indicates the length of text tokens.

Dynamically encoding video based on query text helps reduce bias in Banzhaf Interaction. However, since the cross-modal representation can only encode the entire video and text, it cannot be directly applied to our fine-grained game interaction modeling. When the semantic content of the text and video diverge greatly, this text-conditioned representation might not only fail to enhance single-modal representations but also might potentially perform worse than original single-modal representations. To this end, we propose to reconstruct the representation as a fusion of single-modal and cross-modal components. This reconstructed representation offers the combined benefits of the granularity inherent in single-modal representations and the adaptability of cross-modal representations. As shown in Fig. 5, we apply this representation reconstruction method to both video and text, enabling the enhanced encoding of both modalities.

Starting with the video representation, we first aggregate frame representation with the text-frame attention encoder. Concretely, we feed the [CLS] token t_c from the text representation as a query and the single-modal representation $\boldsymbol{V}_f^s$ as keys and values into an attention module. The resulting cross-modal video representation $\boldsymbol{V}_f^c$ is defined as:

$$\boldsymbol{V}_f^c = \{v^c\}^{N_v}, \quad v^c = \text{Softmax}(t_c \boldsymbol{V}_f^{s\top}) \boldsymbol{V}_f^s. \quad (2)$$

Subsequently, the video representation $\boldsymbol{V}_f$ is defined as:

$$\boldsymbol{V}_f = \gamma \boldsymbol{V}_f^s + (1-\gamma) \boldsymbol{V}_f^c, \quad (3)$$

where $\gamma = \text{MLP}(\boldsymbol{V}_f^s - \boldsymbol{V}_f^c)$ serves as a factor used to adjust the balance between the single-modal and cross-modal components. When $\gamma = 0$, the video representation is equivalent
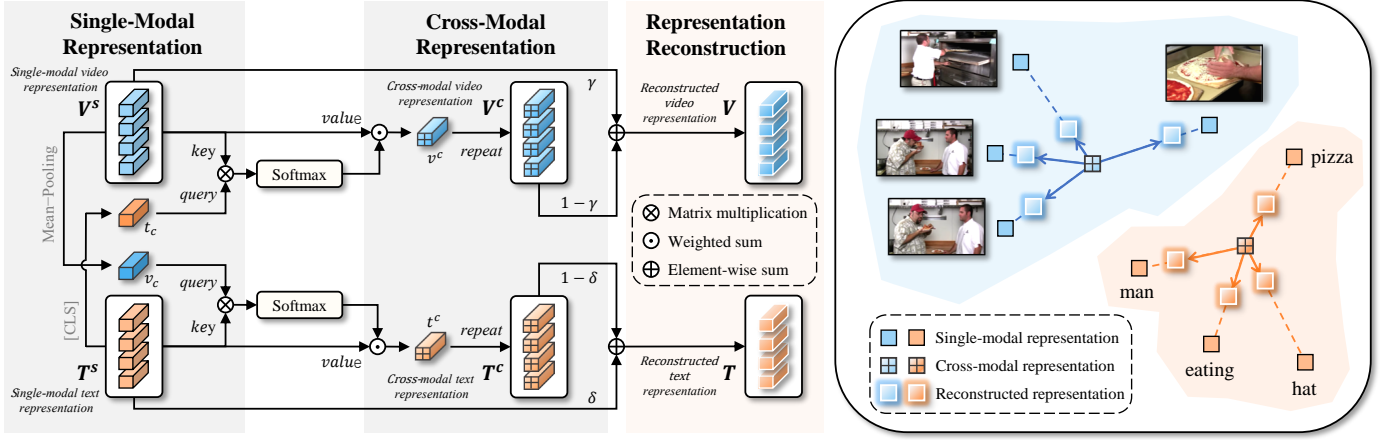
Fig. 5: **The representation reconstruction module.** To address the bias in calculations within Banzhaf Interaction, we reconstruct both video and text representation as a fusion of single-modal and cross-modal components. The representation reconstruction module maintains the granularity inherent in single-modal representations while preserving the adaptive encoding capabilities of cross-modal representations.

to the cross-modal video representation, *i.e.*, $V_f = V_f^c$. Conversely, when $\gamma = 1$, the video representation corresponds to the single-modal video representation, *i.e.*, $V_f = V_f^s$.

To reconstruct the text representation, we employ the mean of the video frame representation, denoted as $v_c = \text{MeanPool}(V_f^s)$, as a query. Subsequently, we use this query along with the single-modal representation T_w^s as keys and values to obtain the cross-modal text representation T_w^c:

$$T_w^c = \{t^c\}^{N_t}, \quad t^c = \text{Softmax}(v_c T_w^{s\top}) T_w^s. \quad (4)$$

Finally, the text representation T_w is defined as:

$$T_w = \delta T_w^s + (1 - \delta) T_w^c, \quad (5)$$

where $\delta = \text{MLP}(T_w^s - T_w^c)$ serves as a factor to adjust the proportion of single-modal and cross-modal components. When $\delta = 0$, the text representation is equivalent to the cross-modal text representation, *i.e.*, $T_w = T_w^c$. Conversely, when $\delta = 1$, the text representation corresponds to the single-modal text representation, *i.e.*, $T_w = T_w^s$.

3.2 Hierarchical Banzhaf Interaction

Generally, given a corpus of video-text pairs (v, t), cross-modal representation learning aims to learn a video encoder and a text encoder. The problem is formulated as a cross-modality similarity measurement $S_{v,t}$ by cross-modal contrastive learning, where the matched video-text pairs are close and the mismatched pairs are away from each other.

As the fine-grained relationships are unavailable, prior works typically directly apply the cross-modal contrastive loss to optimize the similarity scores $S_{v,t}$. To move a step further, we model video-text as game players with multivariate cooperative game theory to handle the uncertainty during fine-grained semantic interaction with diverse granularity, flexible combination, and vague intensity.

3.2.1 *Video-Text as Game Players*
Given features $V_f = \{v_f^i\}_{i=1}^{N_v}$ and $T_w = \{t_w^j\}_{j=1}^{N_t}$, fine-grained cross-modal learning aims to find semantically matched video-text feature pairs. Specifically, if a video

frame and a text word have strong semantic correspondence, then they tend to cooperate with each other and contribute to the fine-grained similarity score. Thus, we can consider $\mathcal{N} = \{v_f^i\}_{i=1}^{N_v} \cup \{t_w^j\}_{j=1}^{N_t}$ as players in the game.

To achieve the goal of the cooperative game and cross-modal learning to be completely consistent, the characteristic function ϕ should meet all the following criteria:

- **(a)** the final score benefits from strongly corresponding semantic pairs $\{v_f^+, t_w^+\}$, *i.e.*, $\phi(\mathcal{N}) - \phi(\mathcal{N} \setminus \{v_f^+, t_w^+\} \cup \{[\{v_f^+, t_w^+\}]\}) < 0$;
- **(b)** the final score is compromised by semantically irrelevant pairs $\{v_f^-, t_w^-\}$, *i.e.*, $\phi(\mathcal{N}) - \phi(\mathcal{N} \setminus \{v_f^-, t_w^-\} \cup \{[\{v_f^-, t_w^-\}]\}) > 0$;
- **(c)** when there are no players to cooperate, the final score is zero, *i.e.*, $\phi(\{v_f^i\}_{i=1}^{N_v}) = \phi(\{t_w^j\}_{j=1}^{N_t}) = \phi(\emptyset) = 0$, where $\emptyset$ denotes the empty set.

Note that any function satisfying the above conditions can be used as the characteristic function ϕ. For simplicity, we use cross-modality similarity measurement S as ϕ. Subsequently, we leverage Banzhaf Interaction to value possible correspondence between video frames and text words and to enhance cross-modal representation learning.

3.2.2 *Multivariate Cooperative Game Modeling*
In the following, we first introduce the basic two-player interaction between video frames and word tokens. Subsequently, we expand the two-player interaction to the multivariate interaction using the token merge module. To make our module clear, we provide pseudo-code in Alg. 1.

For a coalition $\{v_f^i, t_w^j\}$, referring to Eq. 1, we can calculate the Banzhaf Interaction $\mathcal{I}([\{v_f^i, t_w^j\}])$. Due to the disparity in semantic similarity and interaction index, we design a prediction header to predict the fine-grained relationship $\mathcal{R}_{i,j}$ between the i_{th} video frame and the j_{th} text word. The prediction header consists of a convolutional layer for encoding, a self-attention module for capturing global interaction, and a convolutional layer for decoding. In Table 5, we provide the experiment results of the prediction header with different structures.

Algorithm 1: Multivariate Cooperative Game

Require: $\boldsymbol{V}_f^s = \{\hat{v}_f^i\}_{i=1}^{N_v}$ (original video representation),
$\boldsymbol{T}_w^s = \{\hat{t}_w^j\}_{j=1}^{N_t}$ (original text representation)
1: # Entity level
2: Calculate $\boldsymbol{V}_f^c$, then combine $\boldsymbol{V}_f^s$ and $\boldsymbol{V}_f^c$ to obtain $\boldsymbol{V}_f$
3: Calculate $\boldsymbol{T}_w^c$, then combine $\boldsymbol{T}_w^s$ and $\boldsymbol{T}_w^c$ to obtain $\boldsymbol{T}_w$
4: Use $\boldsymbol{V}_f$ and $\boldsymbol{T}_w$ to predict Banzhaf Interaction and calculate Banzhaf Interaction loss $\mathcal{L}_I^e$
5:
6: # Action level
7: Obtain coalitions $\boldsymbol{C}_v$ from $\boldsymbol{V}_f^s$ using DPC-KNN
8: Obtain coalitions $\boldsymbol{C}_t$ from $\boldsymbol{T}_w^s$ using DPC-KNN
9: Utilize $\boldsymbol{C}_v$ as queries, with $\boldsymbol{V}_f^s$ serving as keys and values, to obtain $\boldsymbol{V}_a^s$ through Attention
10: Utilize $\boldsymbol{C}_t$ as queries, with $\boldsymbol{T}_w^s$ serving as keys and values, to obtain $\boldsymbol{T}_a^s$ through Attention
11: Calculate $\boldsymbol{V}_a^c$, then combine $\boldsymbol{V}_a^s$ and $\boldsymbol{V}_a^c$ to obtain $\boldsymbol{V}_a$
12: Calculate $\boldsymbol{T}_a^c$, then combine $\boldsymbol{T}_a^s$ and $\boldsymbol{T}_a^c$ to obtain $\boldsymbol{T}_a$
13: Use $\boldsymbol{V}_a$ and $\boldsymbol{T}_a$ to predict Banzhaf Interaction and calculate Banzhaf Interaction loss $\mathcal{L}_I^a$
14:
15: # Event level
16: Obtain coalitions $\boldsymbol{C}_v'$ from $\boldsymbol{V}_a^s$ using DPC-KNN
17: Obtain coalitions $\boldsymbol{C}_t'$ from $\boldsymbol{T}_a^s$ using DPC-KNN
18: Utilize $\boldsymbol{C}_v'$ as queries, with $\boldsymbol{V}_a^s$ serving as keys and values, to obtain $\boldsymbol{V}_v^s$ through Attention
19: Utilize $\boldsymbol{C}_t'$ as queries, with $\boldsymbol{T}_a^s$ serving as keys and values, to obtain $\boldsymbol{T}_v^s$ through Attention
20: Calculate $\boldsymbol{V}_v^c$, then combine $\boldsymbol{V}_v^s$ and $\boldsymbol{V}_v^c$ to obtain $\boldsymbol{V}_v$
21: Calculate $\boldsymbol{T}_v^c$, then combine $\boldsymbol{T}_v^s$ and $\boldsymbol{T}_v^c$ to obtain $\boldsymbol{T}_v$
22: Use $\boldsymbol{V}_v$ and $\boldsymbol{T}_v$ to predict Banzhaf Interaction and calculate Banzhaf Interaction loss $\mathcal{L}_I^v$
23: **return** $\mathcal{L}_I^e, \mathcal{L}_I^a, \mathcal{L}_I^v$

Subsequently, we optimize the Kullback-Leibler (KL) divergence [47] between the $\mathcal{I}([\{v_f^i, t_w^j\}])$ and $\mathcal{R}_{i,j}$. Concretely, we define the probability distribution of the video-to-text task and the text-to-video task as:

$$\mathcal{D}_{v2t}^{\mathcal{I}} = [p_{i,1}^{\mathcal{I}}, p_{i,2}^{\mathcal{I}}, ..., p_{i,N_t}^{\mathcal{I}}], \ \mathcal{D}_{t2v}^{\mathcal{I}} = [\hat{p}_{1,j}^{\mathcal{I}}, \hat{p}_{2,j}^{\mathcal{I}}, ..., \hat{p}_{N_v,j}^{\mathcal{I}}],$$

$$p_{i,j}^{\mathcal{I}} = \frac{\exp\Big(\mathcal{I}([\{v_f^i, t_w^j\}])\Big)}{\sum_{k=1}^{N_t} \exp\Big(\mathcal{I}([\{v_f^i, t_w^k\}])\Big)},$$

$$\hat{p}_{i,j}^{\mathcal{I}} = \frac{\exp\Big(\mathcal{I}([\{v_f^i, t_w^j\}])\Big)}{\sum_{k=1}^{N_v} \exp\Big(\mathcal{I}([\{v_f^k, t_w^j\}])\Big)}. \tag{6}$$

Similarly, the probability distribution $\mathcal{D}_{v2t}^{\mathcal{R}}$ and $\mathcal{D}_{t2v}^{\mathcal{R}}$ are calculated in the same way using $\mathcal{R}^{i,j}$:

$$\mathcal{D}_{v2t}^{\mathcal{R}} = [p_{i,1}^{\mathcal{R}}, p_{i,2}^{\mathcal{R}}, ..., p_{i,N_t}^{\mathcal{R}}], \ \mathcal{D}_{t2v}^{\mathcal{R}} = [\hat{p}_{1,j}^{\mathcal{R}}, \hat{p}_{2,j}^{\mathcal{R}}, ..., \hat{p}_{N_v,j}^{\mathcal{R}}],$$

$$p_{i,j}^{\mathcal{R}} = \frac{\exp(\mathcal{R}_{i,j})}{\sum_{k=1}^{N_t} \exp(\mathcal{R}_{i,k})},$$

$$\hat{p}_{i,j}^{\mathcal{R}} = \frac{\exp(\mathcal{R}_{i,j})}{\sum_{k=1}^{N_v} \exp(\mathcal{R}_{k,j})}. \tag{7}$$

Finally, the Banzhaf Interaction loss $\mathcal{L}_I$ is defined as:

$$\mathcal{L}_I = \mathbb{E}_{\boldsymbol{v},\boldsymbol{t}}\Big[\text{KL}(\mathcal{D}_{v2t}^{\mathcal{R}}\|\mathcal{D}_{v2t}^{\mathcal{I}}) + \text{KL}(\mathcal{D}_{t2v}^{\mathcal{R}}\|\mathcal{D}_{t2v}^{\mathcal{I}})\Big]. \tag{8}$$

The Banzhaf Interaction loss $\mathcal{L}_I$ brings the probability distributions of the output $\mathcal{R}$ of the prediction header and Banzhaf Interaction $\mathcal{I}$ close together to establish fine-grained semantic alignment between video frames and text words. In particular, it can be directly removed during inference, rendering an efficient and semantics-sensitive model.
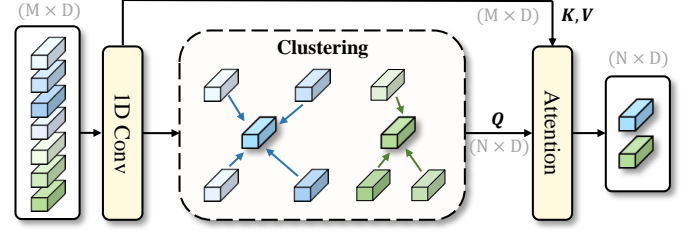

Fig. 6: **The token merge module.** "1D-Conv" denotes the one-dimensional convolutional layer. M input tokens with D channels are first clustered into N clusters. Subsequently, we feed the merged tokens as queries Q and the original tokens as keys K and values V into an attention module.

For multivariate interaction, an intuitive method is to compute Banzhaf Interaction on any candidate set of visual frames and text words directly. However, the number of candidate sets is too large, *i.e.*, $2^{N_v+N_t}$. To reduce the number of candidate sets, we cluster the original visual (textual) tokens and compute the Banzhaf Interaction between the merged tokens. By stacking token merge modules, we get cross-modal interaction efficiently at different semantic levels, *i.e.*, entity-level interactions on the frames and words, action-level interactions on the clips and phrases, and event-level interactions on the segments and paragraphs. Fig. 6 illustrates the framework of the token merge module.

Specifically, we utilize DPC-KNN [48], a k-nearest neighbor-based density peaks clustering algorithm, to cluster the visual (textual) tokens. Starting with the frame-level tokens $\boldsymbol{V}_f^s = \{\hat{v}_f^i\}_{i=1}^{N_v}$, we first use a one-dimensional convolutional layer to enhance the temporal information between tokens. Subsequently, we compute the local density ρ_i of each token $\hat{v}_f^i$ according to its K-nearest neighbors:

$$\rho_i = \exp\Big(-\frac{1}{K}\sum_{\hat{v}_f^k \in \text{KNN}(\hat{v}_f^i)} \|\hat{v}_f^k - \hat{v}_f^i\|^2\Big), \tag{9}$$

where $\text{KNN}(\hat{v}_f^i)$ is the K-nearest neighbors of $\hat{v}_f^i$. After that, we compute the distance index ξ_i of each token $\hat{v}_f^i$:

$$\xi_i = \begin{cases} \min\limits_{j:\rho_j>\rho_i} \|\hat{v}_f^j - \hat{v}_f^i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i. \\ \max\limits_{j} \|\hat{v}_f^j - \hat{v}_f^i\|^2, & \text{otherwise}. \end{cases} \tag{10}$$

Intuitively, ρ denotes the local density of tokens, and ξ represents the distance from other high-density tokens.

We consider those tokens with relatively high $\rho_i \times \xi_i$ as cluster centers and then assign other tokens to the nearest cluster center according to the Euclidean distances. Inspired by [49], [50], we use the weighted average tokens of each cluster to represent the corresponding cluster, where the weight $W = \text{MLP}_w(\boldsymbol{V}_f^s)$. Subsequently, we feed the weighted average tokens as queries Q and the original tokens as keys K and values V into an attention module.

To make the token with high weight have a high contribution to the output, we incorporate the weight W into the attention module as follows:

$$\text{Attention}(Q, K, V, W) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d_k}} + W\Big)V, \tag{11}$$

where d_k is the channel number of the queries and is used as the scaling factor. We treat the output of the attention
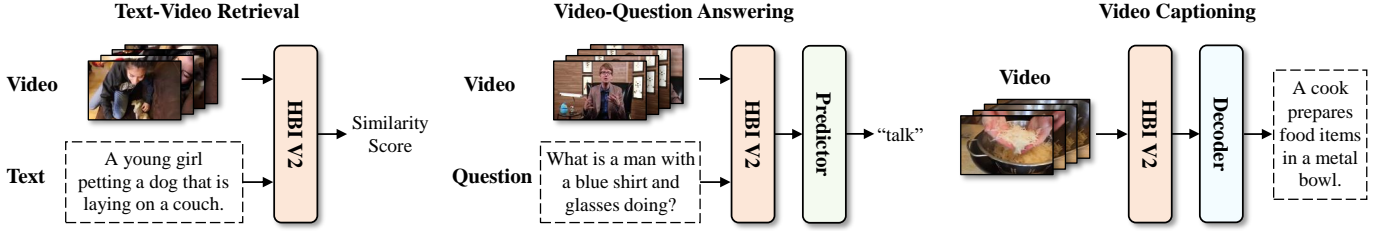
Fig. 7: **The task-specific prediction heads.** To enhance the versatility of HBI V2 across various downstream tasks, we expand our original structure into a flexible encoder-decoder framework comprising an encoder and a task-specific decoder.

module as features at a higher semantic level than the entity level, that is, the action-level visual tokens. Similarly, we merge the action-level tokens again to get the event-level tokens. The action-level textual tokens and event-level textual tokens are calculated in the same way.

3.3 Task-Specific Prediction Heads

As shown in Fig. 7, to enhance the versatility of HBI V2 across various downstream tasks, including VideoQA and video captioning, we expand our original HBI structure into a flexible encoder-decoder framework. This framework consists of an encoder and a task-specific decoder, allowing for efficient adaptation to diverse tasks.

In text-to-video retrieval, given a text query alongside a gallery of videos, the objective is to rank all videos so that the video corresponding to the text query is ranked as high as possible. Similarly, in video-to-text retrieval, the goal is to rank all text candidates based on the video query. In our HBI V2 framework, we can directly rank candidates by leveraging the similarity scores between video and text, eliminating the necessity for an additional prediction head.

For video-question answering, due to the established video-text fine-grained alignment from the hierarchical Banzhaf Interaction module, we can adopt a simplified answer prediction head, without the need for sophisticated multi-modal fusion/reasoning stages like many previous VideoQA models. For the question-answering prediction head, we concatenate the video representation and text representation followed by an MLP to predict the answer.

For video captioning, we utilize a single-layer transformer decoder as the generator of the caption. The decoder sequentially produces hidden features. Subsequently, we employ a linear projection layer to map these hidden features to the vocabulary dictionary. Notably, given the absence of text input in the video captioning task, we rely on single-modal representations instead of reconstructed representations during inference. However, it is worth noting that the reconstructed representations are still utilized for video-language alignment purposes during training.

3.4 Training Objective

To achieve fine-grained semantic alignment, we first employ cross-modal contrastive loss. Given video representation $\boldsymbol{V}_f = \{v_f^i\}_{i=1}^{N_v}$ and text representation $\boldsymbol{T}_w = \{t_w^j\}_{j=1}^{N_t}$, the alignment matrix is defined as: $A = [a_{ij}]^{N_v \times N_t}$, where $a_{ij} = \frac{(v_f^i)^\top t_w^j}{\|v_f^i\|\|t_w^j\|}$ represents the alignment score between the i_{th} video frame and the j_{th} text word. For the i_{th} video frame, we calculate its maximum alignment score

as $\max_j a_{ij}$. Subsequently, we use the weighted average maximum alignment score over all video frames as the video-to-text similarity. Similarly, we can obtain the text-to-video similarity. The total similarity score is defined as:

$$S_{\boldsymbol{v},\boldsymbol{t}} = \frac{1}{2}\Big(\underbrace{\sum_{i=1}^{N_v} \omega_v^i \max_j a_{ij}}_{\text{video-to-text similarity}} + \underbrace{\sum_{j=1}^{N_t} \omega_t^j \max_i a_{ij}}_{\text{text-to-video similarity}} \Big), \quad (12)$$

where $[\omega_v^0, \omega_v^1, ..., \omega_v^{N_v}] = \text{Softmax}\big(\text{MLP}_v(\boldsymbol{V}_f)\big)$ and $[\omega_t^0, \omega_t^1, ..., \omega_t^{N_t}] = \text{Softmax}\big(\text{MLP}_t(\boldsymbol{T}_w)\big)$ are the weights of the video frames and text words, respectively. Then the cross-modal contrastive loss [62] is formulated as:

$$\mathcal{L}_C = -\frac{1}{2}\Big[\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(S_{\boldsymbol{v}_k,\boldsymbol{t}_k}/\tau)}{\sum_l^B \exp(S_{\boldsymbol{v}_k,\boldsymbol{t}_l}/\tau)}+ \\ \frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(S_{\boldsymbol{v}_k,\boldsymbol{t}_k}/\tau)}{\sum_l^B \exp(S_{\boldsymbol{v}_l,\boldsymbol{t}_k}/\tau)}\Big], \quad (13)$$

where B is the batch size and τ is the temperature hyper-parameter. This loss function maximizes the similarity of positive pairs and minimizes the similarity of negative pairs.

Combining the cross-modal contrastive loss $\mathcal{L}_C$ and Banzhaf Interaction loss $\mathcal{L}_I$, the full objective of semantic alignment is formulated as $\mathcal{L} = \mathcal{L}_C + \alpha\mathcal{L}_I$, where α is the trade-off hyper-parameter. We train the network at three semantic levels, which are shown as follows:

$$\mathcal{L}^e = \mathcal{L}_C^e + \alpha\mathcal{L}_I^e, \quad \mathcal{L}^a = \mathcal{L}_C^a + \alpha\mathcal{L}_I^a, \quad \mathcal{L}^v = \mathcal{L}_C^v + \alpha\mathcal{L}_I^v, \quad (14)$$

where $\mathcal{L}^e$, $\mathcal{L}^a$, and $\mathcal{L}^v$ represent the semantic alignment loss at the entity level, action level, and event level, respectively.

To further improve the generalization ability, we optimize the additional KL divergence between the distribution among different semantic levels. We find that the entity-level similarity $S_{\boldsymbol{v},\boldsymbol{t}}^e$ converges first in the training process, so we distill the entity-level similarity to the other two semantic levels. Starting with entity-level similarity $S_{\boldsymbol{v},\boldsymbol{t}}^e$ distilling to action-level similarity $S_{\boldsymbol{v},\boldsymbol{t}}^a$, we first calculate the distribution $\mathcal{D}_{v2t}^e$ and $\mathcal{D}_{t2v}^e$ by replacing $\mathcal{I}([\{v,t\}])$ with $S_{\boldsymbol{v},\boldsymbol{t}}^e$ in Eq. 6. The distribution $\mathcal{D}_{v2t}^a$ and $\mathcal{D}_{t2v}^a$ are calculated using $S_{\boldsymbol{v},\boldsymbol{t}}^a$ in the same way. The $\mathcal{L}_D^{e2a}$ loss is defined as:

$$\mathcal{L}_D^{e2a} = \mathbb{E}_{\boldsymbol{v},\boldsymbol{t}}\Big[\text{KL}(\mathcal{D}_{v2t}^a\|\mathcal{D}_{v2t}^e) + \text{KL}(\mathcal{D}_{t2v}^a\|\mathcal{D}_{t2v}^e)\Big]. \quad (15)$$

The $\mathcal{L}_D^{e2v}$ loss from entity-level similarity to event-level similarity is calculated in the same way.

Furthermore, depending on the specific task, we may require the inclusion of task-specific loss. Particularly, in the

TABLE 1: **Comparisons to current text-video retrieval methods on the MSRVTT, ActivityNet Captions, and DiDeMo datasets.** "↑" denotes that higher is better. "↓" denotes that lower is better. All results do not use inverted softmax [51].

(a) Text-to-video retrieval and video-to-text retrieval performance on the MSRVTT dataset.

Methods	Text-to-Video						Video-to-Text					
	R@1 ↑	R@5 ↑	R@10 ↑	Rsum ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	Rsum ↑	MdR ↓	MnR ↓
MMT [52] ECCV 2020	26.6	57.1	69.6	153.3	4.0	24.0	27.0	57.5	69.7	154.2	3.7	21.3
T2VLAD [53] CVPR 2021	29.5	59.0	70.1	158.6	4.0	-	31.8	60.0	71.1	162.9	3.0	-
TT-CE [54] ICCV 2021	29.6	61.6	74.2	165.4	3.0	-	32.1	62.7	75.0	169.8	3.0	-
Support-Set [55] ICLR 2021	30.1	58.5	69.3	157.9	3.0	-	28.5	58.6	71.6	158.7	3.0	-
CLIP4Clip [56] Neurocomputing 2022	44.5	71.4	81.6	197.5	**2.0**	15.3	42.7	70.9	80.6	194.2	**2.0**	11.6
EMCL [3] NeurIPS 2022	46.8	73.1	83.1	203.0	**2.0**	12.8	46.5	73.5	83.5	203.5	**2.0**	8.8
X-Pool [57] CVPR 2022	46.9	72.8	82.2	201.9	**2.0**	14.3	44.4	73.3	84.0	201.7	**2.0**	9.0
TS2-Net [58] ECCV 2022	47.0	74.5	**83.8**	205.3	**2.0**	13.0	45.3	74.1	83.7	203.1	**2.0**	9.2
UATVR [59] ICCV 2023	47.5	73.9	83.5	204.9	**2.0**	12.3	46.9	73.8	83.7	204.5	**2.0**	8.6
Prompt Switch [60] ICCV 2023	47.8	73.9	82.2	203.9	**2.0**	14.1	46.0	**74.3**	**84.8**	205.1	**2.0**	**8.5**
HBI	48.6	74.6	83.4	206.6	**2.0**	**12.0**	46.8	**74.3**	84.3	205.4	**2.0**	8.9
HBI V2	**49.4**	**74.6**	**83.8**	**207.8**	2.0	12.2	**47.7**	74.2	84.4	**206.3**	2.0	8.9

(b) Text-to-video retrieval performance on the ActivityNet Captions and DiDeMo datasets.

Method	ActivityNet Captions						DiDeMo					
	R@1 ↑	R@5 ↑	R@10 ↑	Rsum ↑	MdR ↓	MnR ↓	R@1 ↑	R@5 ↑	R@10 ↑	Rsum ↑	MdR ↓	MnR ↓
ClipBERT [32] CVPR 2021	21.3	49.0	63.5	133.8	6.0	-	20.4	48.0	60.8	129.2	6.0	-
Frozen [34] ICCV 2021	-	-	-	-	-	-	34.6	65.0	74.7	174.3	3.0	-
CLIP4Clip [56] Neurocomputing 2022	40.5	72.4	83.6	196.5	**2.0**	7.5	42.8	68.5	79.2	190.5	**2.0**	18.9
TS2-Net [58] ECCV 2022	41.0	73.6	84.5	199.1	**2.0**	8.4	41.8	71.6	82.0	195.4	**2.0**	14.8
UATVR [59] ICCV 2023	-	-	-	-	-	-	43.1	71.8	82.3	197.2	**2.0**	15.1
EMCL [3] NeurIPS 2022	41.2	72.7	83.6	197.5	**2.0**	8.6	45.3	74.2	82.3	201.8	**2.0**	12.3
CenterCLIP [61] SIGIR 2022	43.9	74.6	85.8	204.3	**2.0**	6.7	-	-	-	-	-	-
HBI	42.2	73.0	84.6	199.8	**2.0**	6.6	46.9	**74.9**	82.7	204.5	**2.0**	**12.1**
HBI V2	**45.5**	**75.3**	**86.0**	**206.8**	**2.0**	**6.5**	**47.4**	74.6	**84.5**	**206.5**	**2.0**	12.7

case of text-video retrieval, there is no necessity to introduce a task-specific loss. However, for tasks such as VideoQA and video captioning, an additional cross-entropy loss $\mathcal{L}_{task}$ needs to be incorporated. The overall loss is the combination of semantically alignment losses, self-distillation losses, and task-specific loss, which is defined as:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}^e + \mathcal{L}^a + \mathcal{L}^v}_{\text{deep supervision}} + \beta \underbrace{(\mathcal{L}_D^{e2a} + \mathcal{L}_D^{e2v})}_{\text{self-distillation}} + \lambda \mathcal{L}_{task}, \quad (16)$$

where β is the trade-off hyper-parameter. For VideoQA and video captioning, we apply λ to balance the feature alignment penalty and the task-specific penalty.

4 EXPERIMENTS

4.1 Experimental Settings

Text-Video Retrieval Datasets. **MSRVTT** [12] contains 10K YouTube videos, each video with 20 text descriptions. We follow the training protocol in [52], [63], [64] and evaluate text-to-video and video-to-text retrieval tasks on the 1K-A testing split with 1K video or text candidates defined by [65]. **ActivityNet Captions** [13] consists of densely annotated temporal segments of 20K YouTube videos. Following [52], [53], [55], we concatenate descriptions of segments in a video to construct "video-paragraph" for retrieval. We use the 10K training split to finetune the model and report the performance on the 5K "val1" split. **DiDeMo** [14] contains 10K videos annotated with 40K text descriptions. We concatenate descriptions of segments in a video to construct a "video paragraph" for retrieval. We follow the training and evaluation protocol in CLIP4Clip [56].

Video-Question Answering Datasets. **MSRVTT-QA** [15] comprises 10K videos and 243K question-answer pairs. The video has a longer duration of 10-30 seconds. **MSVD-QA** [15] comprises 1,970 short clips (10 seconds on average) and 50,505 question-answer pairs. **ActivityNet-QA** [16] consists of 58,000 QA pairs on 5,800 complex long web videos. The average video length of ActivityNet-QA is 180 seconds, longer than MSRVTT-QA and MSVD-QA.

Video Captioning Dataset. The **MSRVTT-Caption** [12] in the video captioning task is the same as the MSRVTT dataset in the text-video retrieval task.

Evaluation Metrics. We choose Recall at rank K (R@K), Median Rank (MdR), and mean rank (MnR) to evaluate the retrieval performance. We select the answer accuracy to evaluate the video-question answering performance. We apply four metrics for the video caption task, including BLEU-4 [66], ROUGE-L [67], METEOR [68], and CIDEr [69].

Implementation Details. We utilize the CLIP [27] as the pre-trained model. We use the Adam optimizer [70] and set the temperature τ to 0.01. Since the calculation of the exact Banzhaf Interaction is an NP-hard problem [71], existing methods mainly use sampling-based methods [72] to obtain unbiased estimates. To speed up the computation of Banzhaf Interaction for many data instances, we pre-train a tiny model to learn a mapping from a set of input features to a result using MSE loss. The tiny model consists of 2 CNN layers and a self-attention layer. The input is the similarity matrix of video frames and text tokens, and the output is the estimation of Banzhaf Interaction. For text-video retrieval, the frame length and caption length are 12 and 32 for the MSRVTT dataset. We set the frame length and caption length to 64 and 64 for the ActivityNet Captions
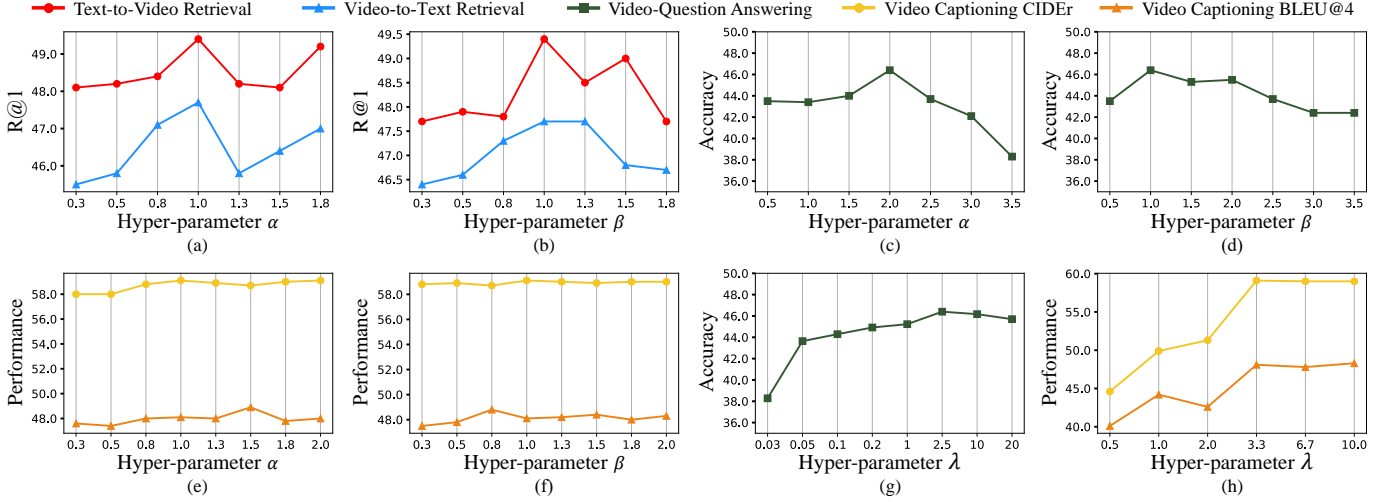
Fig. 8: **Parameter sensitivity.** α, β, and λ are the trade-off hyper-parameters in Eq. 14 and Eq. 15. (a) and (b) are the effects of hyper-parameters α and β on the MSRVTT dataset for the text-video retrieval task. (c) and (d) are the effects of hyper-parameters α and β on the MSRVTT-QA dataset for the video-question answering task. (e) and (f) are the effects of hyper-parameters α and β on the MSRVTT dataset for the video captioning task. (g) and (h) are the effects of hyper-parameter λ on the MSRVTT-QA and MSRVTT datasets for the video-question answering and video captioning tasks.

dataset. We set the frame length and caption length to 32 and 64 for the DiDeMo dataset. The initial learning rate is 1e-7 for text encoder and video encoder and 1e-3 for other modules. The network is optimized with a batch size of 128 in 5 epochs. For video-question answering, we use the target vocabulary and train a fully connected layer on top of the final language features to classify the answer. The initial learning rate is 1e-7 for text encoder and video encoder and 1e-3 for other modules. We set the task balance factor λ to 2.5. The network is optimized with a batch size of 32 in 5 epochs. For video captioning, we initialize the learning rate at 5e-4 and freeze the CLIP encoder. The network is optimized with a batch size of 64 in 50 epochs.

4.2 Comparison with State-of-the-Art

We conduct extensive experiments on three datasets for text-video retrieval, three datasets for video-question answering, and one dataset for video captioning. In Table 1, we show the retrieval results of our method on MSRVTT, ActivityNet Captions, and DiDeMo datasets. Additionally, Table 2 showcases the video-question answering results on the MSRVTT-QA, MSVD-QA, and ActivityNet-QA datasets. Furthermore, in Table 3, we exhibit the video captioning results on the MSRVTT dataset. Our proposed HBI V2 consistently outperforms both the previous HBI and existing methods across all downstream tasks. These results demonstrate the effectiveness of our proposed HBI V2 framework. Additionally, We observe that HBI V2 exhibits notable advancements over HBI on the MSVD-QA and ActivityNet-QA, the improvement on the MSRVTT-QA appears less pronounced. We consider this may be because our framework focuses on video-language alignment, whereas the MSRVTT-QA dataset presents a greater challenge in video-language reasoning, resulting in comparatively marginal enhancements.

4.3 Ablation Study

Ablation about Components. To illustrate the importance of each part of our method including the Banzhaf Inter-

action, the deep supervision structure, the self-distillation, and the representation reconstruction, we conduct ablation experiments on both MSRVTT and MSRVTT-QA datasets in Table 4. The Banzhaf Interaction boosts the baseline with an improvement up to 0.8% at R@1 and 1.3% at answer accuracy. Furthermore, deep supervision and self-distillation significantly improve the generalization ability. Additionally, the representation reconstruction further improves the performance of the model. Our full model attains superior performance, surpassing the baseline by 2.8% at R@1 for text-to-video retrieval and 1.9% at answer accuracy for video-question answering. This shows that the four parts are beneficial for aligning videos and texts.

The parameter α serves as the hyper-parameter that balances the cross-modality contrastive loss ($\mathcal{L}_C$) and the Banzhaf Interaction loss ($\mathcal{L}_I$). Meanwhile, the parameter β acts as the hyper-parameter regulating the trade-off between the loss from deep supervision and the loss from self-distillation. λ is the hyper-parameter used to balance the feature alignment penalty and the task-specific penalty. To determine the optimal settings for α and β for the text-video retrieval task, we evaluate the scale range setting $\alpha \in [0.3, 1.8]$ as shown in Fig. 8(a). Our findings indicate that enhancing α from 0.5 to 0.8 leads to an improvement in R@1 from 48.2% to 48.4%, with saturation observed at $\alpha = 1.0$ for text-to-video retrieval. Consequently, we adopt $\alpha = 1.0$ to achieve optimal performance. In Fig. 8(b), we demonstrate the impact of the hyper-parameter β, evaluating the scale range setting $\beta \in [0.3, 1.8]$. We find that the model achieves the best performance at $\beta = 1.0$. For the video-question answering task, we assess the scale range settings $\alpha \in [0.5, 3.5]$ (Fig. 8(c)) and $\beta \in [0.5, 3.5]$ (Fig. 8(d)). We find that the model achieves the best performance at $\alpha = 2.0$ and $\beta = 1.0$. Therefore, we establish α as 2.0 and β as 1.0 for video-question answering. For the video captioning task, as shown in Fig. 8(e) and (f), we find that the video captioning task is robust to hyper-parameters α and β. We consider that this is because we trained the video

TABLE 2: **Comparisons to VideoQA methods on the MSRVTT-QA, MSVD-QA, and ActivityNet-QA datasets.** We gray out pretrained models to ensure a fair comparison.

Methods	MSRVTT-QA	MSVD-QA	ActivityNet-QA
	Accuracy ↑	Accuracy ↑	Accuracy ↑
Co-Tok [33]	45.7	48.6	-
All-in-One [73]	46.8	48.3	-
FrozenBiLM [74]	47.0	54.8	43.2
MGIN [75]	38.2	39.7	-
IGV [76]	38.3	40.8	-
CLIP-QA [27]	39.0	38.5	-
VQA-T [77]	39.6	41.2	36.8
DRV [78]	40.0	42.2	-
CLIP4Clip [56]	40.9	39.3	-
SViTT [79]	43.2	-	43.0
TG-VQA [80]	46.3	52.5	48.3
HBI	46.2	52.1	48.2
HBI V2	**46.4**	**53.4**	**49.3**

TABLE 3: **Comparisons to current video captioning methods on the MSRVTT dataset.** We gray out pretrained models to ensure a fair comparison.

Methods	MSRVTT			
	BLEU@4 ↑	ROUGE ↑	METEOR ↑	CIDEr ↑
SwinBERT [81]	45.4	30.6	64.1	55.9
MV-GPT [82]	48.9	64.0	38.7	60.0
MGCMP [83]	41.7	62.1	28.9	51.4
OpenBook [84]	42.8	61.7	29.3	52.9
AVSSN [85]	42.8	61.7	28.8	46.9
GSEN [86]	42.9	61.7	28.4	51.0
AFC [87]	43.1	62.7	29.8	56.2
VEIN [88]	44.1	62.6	30.0	55.3
SemSynAN [89]	44.3	62.5	28.8	50.1
EMCL [3]	45.3	63.2	30.2	54.6
ClipCaption [90]	46.1	63.7	30.7	57.7
HBI	45.2	63.0	30.1	55.3
HBI V2	**48.1**	**64.9**	**31.4**	**59.1**

TABLE 4: **Ablation study about the importance of each part of our method on the MSRVTT dataset.** "QA" denotes the answer accuracy on MSRVTT-QA.

$\mathcal{L}_I$ Banzhaf Interaction	Deep Supervision	$\mathcal{L}_D$ Self Distillation	Representation Reconstruction	Text-to-Video				QA↑
				R@1↑	R@5↑	R@10↑	MnR↓	
				46.6	73.1	83.0	13.3	44.5
✓				47.4	74.2	82.8	12.1	45.8
	✓			47.2	74.1	82.6	12.0	46.0
	✓	✓		47.6	73.8	83.2	**11.9**	46.0
✓	✓			48.2	73.0	83.1	12.0	46.1
✓	✓	✓		48.6	**74.6**	83.4	12.0	46.2
✓	✓	✓	✓	**49.4**	**74.6**	**83.8**	12.2	**46.4**

TABLE 5: **Effect of Prediction heads on the MSRVTT.** "None" denotes no header. "SA" denotes the attention module. "CNN+SA" performs the best.

Method	Text-to-Video			
	R@1↑	R@5↑	R@10↑	MnR↓
Baseline	46.6	73.1	83.0	13.3
None	47.7	73.7	83.2	**12.2**
MLP	48.0	74.1	82.6	12.5
CNN	48.5	73.9	82.6	12.6
MLP+SA	48.4	71.8	81.7	12.7
CNN+SA	**49.4**	**74.6**	**83.8**	**12.2**

captioning task for 50 epochs, significantly more than the 5 epochs used for the retrieval and video-question answering tasks. The additional training epochs help mitigate the impact of hyper-parameters α and β on model performance. We set α to 1.0 and β to 1.0 for video captioning. For the hyper-parameter λ, as shown in Fig. 8(g) and (h), we adopt $\lambda = 2.5$ for video-question answering and $\lambda = 3.3$ for video captioning to achieve optimal performance.

Effect of the Prediction Header of $\mathcal{R}$. Due to the disparity in semantic similarity and interaction index, we design a prediction header to predict the fine-grained relationship $\mathcal{R}_{i,j}$ between the i_{th} video frame and the j_{th} text word. To explore the impact of the structure of the prediction header on our method, we compare four popular structures, *i.e.*, "MLP", "CNN", "MLP+SA", and "CNN+SA". As shown in Table 5, we find that the combination of CNN and attention ("CNN+SA") captures both local and global interactions, so it is beneficial for predicting the fine-grained relationship between video and text. As a result, we adopt "CNN+SA" to achieve optimal performance in practice.

Effect of the Number of Clusters. The cluster module shown in Fig. 6 merges the visual tokens and textual tokens, thereby reducing computational overhead. We investigate the impact of merging clusters, as summarized in Table 6, where N_v^a and N_v^o represent the number of visual clusters at the action and event levels, respectively. N_t^a and N_t^o denote the number of text clusters at the action and event levels, respectively. We find the cluster number is a trade-off hyper-parameter: (a) Greater numbers of clusters may lead to semantically similar tokens being classified into different clusters, while (b) fewer clusters may result in dissimilar tokens being grouped together. Upon analysis of Table 6, we select $\{N_v^a, N_v^o, N_t^a, N_t^o\}$ as $\{6, 2, 16, 4\}$ to achieve optimal performance in practice.

Necessity of the Self Distillation. To illustrate the impact of the self-distillation of our method, we conduct ablation experiments on the MSRVTT dataset in Table 7. "E", "A", "O" denote entity level, action level, and event level, respectively. "$->$" indicates the distillation direction. For example, "$E->A$" indicates the distillation from "E" to "A". We find that the entity level converges first in the training process. This indicates that higher-level semantic features are merged from lower-level semantic features. When lower-level semantic features do not converge, it is difficult for higher-level semantic features to learn semantic information. Based on this observation, we propose using lower-level semantic features to guide the learning of higher-level semantic features. As shown in Table 7, self-distillation improves the generalization ability of the model. Distilling from the entity level to the other two semantic levels achieves the best results. Therefore, we distill the entity-level similarity to the other two semantic levels as default in practice.

The Computational Efficiency of Our Method. In Table 8, we calculate the iteration time and inference time on the MSRVTT dataset for the retrieval task. Since the Banzhaf Interaction can be removed during inference, our method only takes an additional 1s for processing the test set (1k
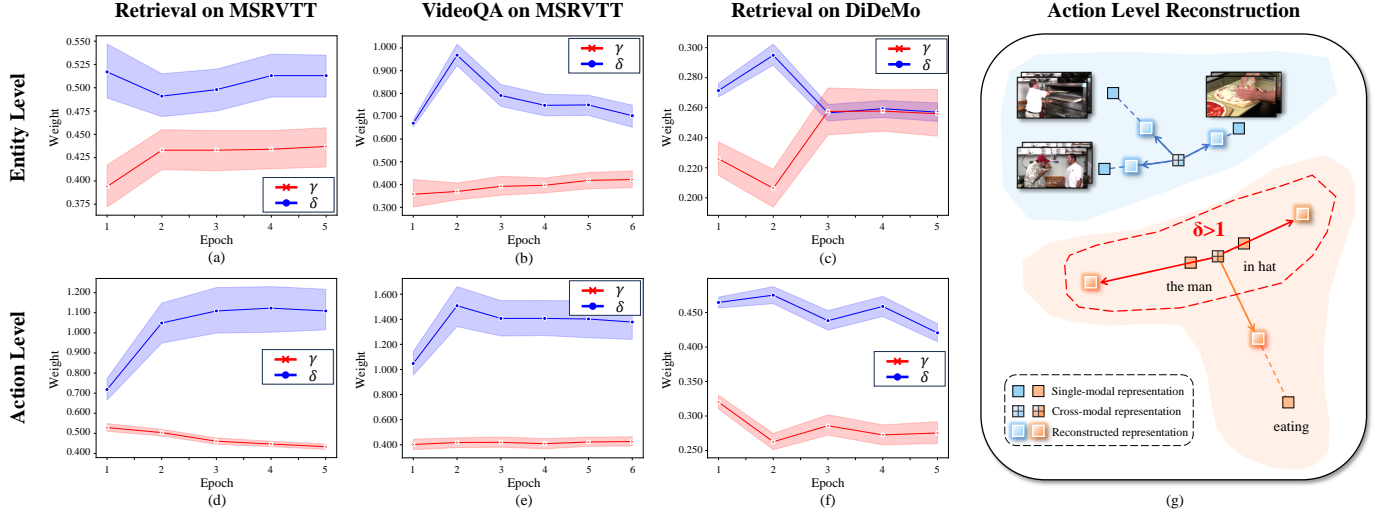
Fig. 9: **The convergence curves of learnable video weight γ and text weight δ for both the text-to-video retrieval and VideoQA tasks.** The video weight γ stably converges across both entity and action levels, while the text weight δ initially fluctuates and converges to different weights at different reconstruction levels.
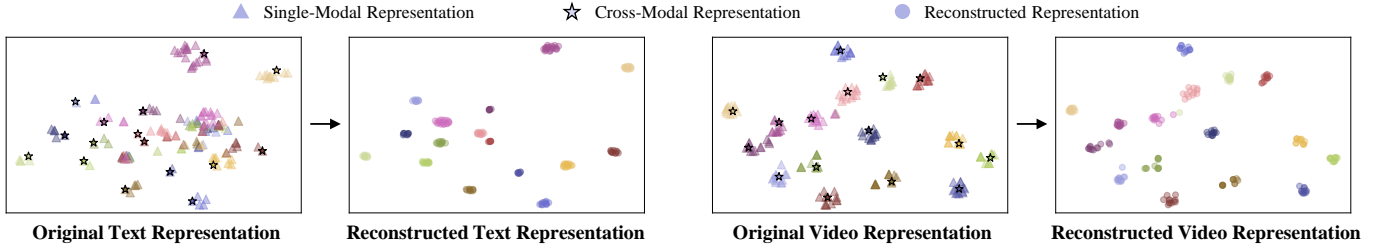

Fig. 10: **Visualization (t-SNE [91]) of the distribution of original and reconstructed representations on the MSRVTT.** We use consistent colors to denote tokens originating from the same text and video. The reconstructed representation maintains the distinctiveness of cross-modal representations while preserving detailed word and frame-level representations.

TABLE 6: **Ablation study for the cluster token number on MSRVTT dataset.** N^a and N^o denote the number of clusters in action level and event level, respectively.

TABLE 7: **Ablation study for the effect of self-distillation loss on MSRVTT.** E, A, O denote entity level, action level, and event level. "E->A + E->O" performs the best.

TABLE 8: **Iteration time and inference time on the MSRVTT dataset.** We report the average time using two Tesla V100 GPUs.

N_v^a	N_v^o	N_t^a	N_t^o	Text-to-Video				
				R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
9	3	18	4	48.7	73.8	**83.8**	**2.0**	12.8
9	3	16	4	48.6	74.1	82.8	**2.0**	12.9
6	2	16	4	**49.4**	**74.6**	**83.8**	**2.0**	**12.2**
6	2	12	3	48.9	74.5	83.0	**2.0**	12.7
3	2	8	3	48.7	74.1	82.9	**2.0**	12.8

Method	Text-to-Video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
E->A	48.7	74.3	83.5	**2.0**	12.9
E->O	48.5	73.5	81.6	**2.0**	12.8
A->O	48.2	74.3	83.2	**2.0**	13.0
E->A + A->O	48.8	74.0	82.9	**2.0**	12.8
E->A + E->O	**49.4**	**74.6**	**83.8**	**2.0**	**12.2**

Method	Iteration Time↓	Inference Time↓
CLIP4Clip [56]	**1.63** s	**16.28** s
EMCL [3]	1.72 s	17.68 s
TS2-Net [58]	2.57 s	19.91 s
Baseline	2.06 s	18.06 s
Ours	3.15 s	19.18 s

videos and 1k text queries). Notably, our method takes less time than TS2-Net during the inference stage. This result demonstrates the superiority of our efficient design.

4.4 Discussions

Convergence curves of learnable γ and δ. As shown in Fig. 9, we observe that the video weight γ is more consistent compared to the text weight δ. At both the entity and action levels, γ stably converges to 0.4~0.5 on the MSRVTT dataset and to 0.2~0.3 on the DiDeMo dataset. This suggests that in video representation, the cross-modal features are beneficial and contribute more significantly than the single-modal features. Although the weight trends are similar in the blue line of Fig. 9(b) and Fig. 9(e), the text weight δ

converges to different values across different reconstruction levels. At the entity level, text weight δ initially fluctuates before stabilizing between 0.7~0.8, indicating that in entity-level reconstruction for text representation, single-modal information plays a larger role compared to cross-modal. However, at the action level, the text weight δ fluctuates initially and converges around 1.4 (greater than 1). We believe the model adjusts the weight ratio to make the single-modal information negative, pushing the action-level features away from the central point. This phenomenon is further illustrated in Fig. 9(g). The degree of clustering among features in the video and text modalities is related to the semantic similarity of the features. At the entity level, there is a significant semantic gap between features in the
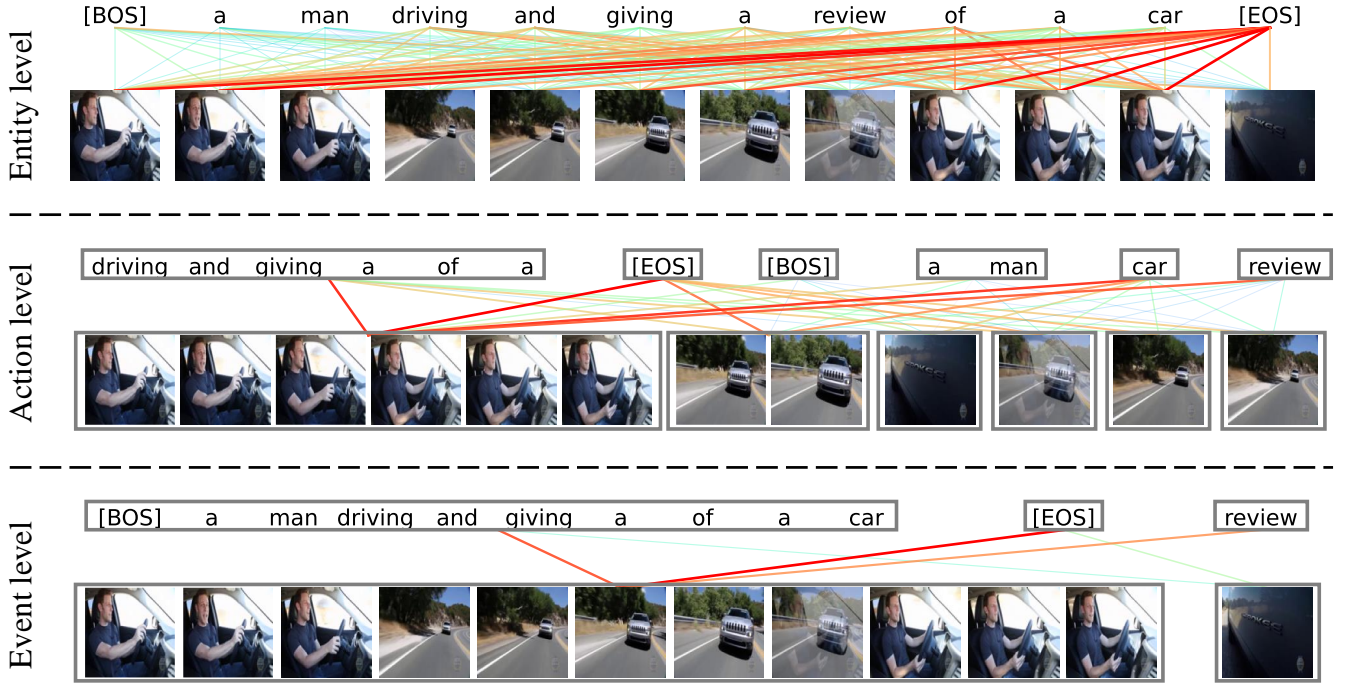
Fig. 11: **Visualization of the hierarchical interaction. We take a representative sample from the MSRVTT dataset as an example.** Here, the degree of confidence from high to low is represented by red, orange, green, and blue lines, respectively.

video and text modalities, with each feature representing different entity information. This noticeable gap helps better match player groups in the game-theoretic interaction. However, the example in Fig. 9(g) shows that at the action level, the semantic overlap between different text features is much greater, leading to a more concentrated distribution in high-dimensional space, which is actually detrimental to the subsequent game-theoretic interaction. As a result, after incorporating text weights, the model adaptively reduces cross-modal information to widen the feature distribution. In contrast, the action-level video features do not face this issue. We consider this because video frames are semantically richer and naturally more dispersed than captions.

Visualization of the Representations. As shown in Fig. 10, we find that single-modal representations tend to be widely dispersed and less distinguishable compared to cross-modal representations. However, cross-modal representations typically capture the entire text or video at a coarse-grained level, lacking granularity at the word and frame levels. Our proposed reconstructed representation maintains the distinctiveness of cross-modal representations while also ensuring word and frame-level representations.

Visualization of the Hierarchical Interaction. As shown in Fig. 11, we find that the semantic similarities between coalitions are generally higher than the semantic similarities between individual frames and individual words. For example, in the action level, the coalition "{driving, and, giving, a, of, a}" has a high semantic similarity with the video coalition representing the man driving action. On the contrary, when these words interact with the corresponding frame as individuals at the entity level, they show low semantic similarity, while most of the similarity scores are taken by the [EOS] token. This indicates that the model at the entity level learns a coarse-grained alignment between

text tokens and video clips. With semantic information, the action level learns a better alignment between text clusters and video clip clusters. The visualization in Fig. 11 also illustrates that our proposed hierarchical interaction method can be used as a tool for visualizing the cross-modal interaction during the cross-modality reasoning tasks. It significantly improves the interpretability of the model and it can help us understand the reasoning details of the models.

5 CONCLUSION

In this work, we creatively model cross-modal representation learning as a multivariate cooperative game by formulating video and text as players in a cooperative game. Specifically, we propose Hierarchical Banzhaf Interaction to value possible correspondence between video frames and text words for sensitive and explainable cross-modal contrast. Although manually labeling the fine-grained relationships between videos and text is unavailable, our method shows a promising alternative to obtaining fine-grained labels based on Banzhaf Interaction. Furthermore, to mitigate the bias in calculations within Banzhaf Interaction, we propose a reconstructed representation that leverages the advantages of both single-modal and cross-modal representations. We extend our original structure into a flexible encoder-decoder framework, allowing the model to adapt to various downstream tasks. To validate the effectiveness and versatility of our method, we conduct extensive experiments encompassing text-video retrieval, video-question answering, and video captioning. More encouragingly, our method can also serve as a visualization tool to promote the understanding of cross-modal interaction.

ACKNOWLEDGEMENT

REFERENCES

[1] Z. Cheng, K. Li, P. Jin, S. Li, X. Ji, L. Yuan, C. Liu, and J. Chen, "Parallel Vertex Diffusion for Unified Visual Grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 1326–1334.

[2] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and Prompt: Video-and-Language Pre-training with Entity Prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4953–4963.

[3] P. Jin, J. Huang, F. Liu, X. Wu, S. Ge, G. Song, D. A. Clifton, and J. Chen, "Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022, pp. 30 291–30 306.

[4] H. Li, Y. Jia, P. Jin, Z. Cheng, K. Li, J. Sui, C. Liu, and L. Yuan, "FreestyleRet: Retrieving Images from Style-Diversified Queries," in *Proceedings of the European Conference on Computer Vision*, 2025, pp. 258–274.

[5] W. Wang, J. Gao, X. Yang, and C. Xu, "Many Hands Make Light Work: Transferring Knowledge from Auxiliary Tasks for Video-Text Retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 2661–2674, 2022.

[6] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang, "Dual Encoding for Video Retrieval by Text," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4065–4080, 2021.

[7] T. Han, W. Xie, and A. Zisserman, "Temporal Alignment Networks for Long-term Video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2906–2916.

[8] P. Jin, H. Li, Z. Cheng, K. Li, X. Ji, C. Liu, L. Yuan, and J. Chen, "DiffusionRet: Generative Text-Video Retrieval with Diffusion Model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2470–2481.

[9] M. Grabisch and M. Roubens, "An axiomatic approach to the concept of interaction among players in cooperative games," *International Journal of Game Theory*, vol. 28, pp. 547–565, 1999.

[10] J.-L. Marichal and P. Mathonet, "Weighted Banzhaf power and interaction indexes through weighted approximations of games," *European Journal of Operational Research*, vol. 211, no. 2, pp. 352–358, 2011.

[11] P. Jin, J. Huang, P. Xiong, S. Tian, C. Liu, X. Ji, L. Yuan, and J. Chen, "Video-Text as Game Players: Hierarchical Banzhaf Interaction for Cross-Modal Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2472–2482.

[12] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5288–5296.

[13] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-Captioning Events in Videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 706–715.

[14] L. Anne Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing Moments in Video with Natural Language," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 5803–5812.

[15] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang, "Video Question Answering via Gradually Refined Attention over Appearance and Motion," in *Proceedings of the ACM International Conference on Multimedia*, 2017, pp. 1645–1653.

[16] Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao, "ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9127–9134.

[17] J. Wei, Y. Yang, X. Xu, X. Zhu, and H. T. Shen, "Universal Weighting Metric Learning for Cross-Modal Retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6534–6545, 2021.

[18] H. Li, J. Huang, P. Jin, G. Song, Q. Wu, and J. Chen, "Weakly-Supervised 3D Spatial Reasoning for Text-Based Visual Question Answering," *IEEE Transactions on Image Processing*, vol. 32, pp. 3367–3382, 2023.

[19] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the 'Video' in Video-Language Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2917–2927.

[20] J. Li, X. He, L. Wei, L. Qian, L. Zhu, L. Xie, Y. Zhuang, Q. Tian, and S. Tang, "Fine-Grained Semantically Aligned Vision-Language Pre-Training," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022, pp. 7290–7303.

[21] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan, "Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 700–13 710.

[22] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, W. HongFa, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. W. Zhang, Z. Li, W. Liu, and L. Yuan, "LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment," in *Proceedings of the International Conference on Learning Representations*, 2024.

[23] H. Li, D. Long, L. Yuan, Y. Wang, Y. Tian, X. Wang, and F. Mo, "Decoupled peak property learning for efficient and interpretable ecd spectra prediction," in *Nature Computational Science*, 2024.

[24] G. Xu, P. Jin, L. Hao, Y. Song, L. Sun, and L. Yuan, "LLaVA-o1: Let Vision Language Models Reason Step-by-Step," *arXiv preprint arXiv:2411.10440*, 2024.

[25] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the Best Pooling Strategy for Visual Semantic Embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 789–15 798.

[26] X. Yang, L. Zhu, X. Wang, and Y. Yang, "DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 6540–6548.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[28] M. Cheng, Y. Sun, L. Wang, X. Zhu, K. Yao, J. Chen, G. Song, J. Han, J. Liu, E. Ding, and J. Wang, "ViSTA: Vision and Scene Text Aggregation for Cross-Modal Retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5184–5193.

[29] L. Xu, H. Huang, and J. Liu, "SUTD-TrafficQA: A Question Answering Benchmark and an Efficient Network for Video Reasoning over Traffic Events," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9878–9888.

[30] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical Conditional Relation Networks for Video Question Answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9972–9981.

[31] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, and T.-S. Chua, "Video Question Answering: Datasets, Algorithms and Challenges," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 6439–6455.

[32] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, "Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7331–7341.

[33] A. Piergiovanni, K. Morton, W. Kuo, M. S. Ryoo, and A. Angelova, "Video Question Answering with Iterative Video-Text Co-Tokenization," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 76–94.

[34] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1728–1738.

[35] P.-Y. Huang, M. Patrick, J. Hu, G. Neubig, F. Metze, and A. Hauptmann, "Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models," in *Proceedings of the Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, 2021, pp. 2443–2459.

[36] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly Localizing and Describing Events for Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7492–7500.

[37] Y. Pan, T. Yao, H. Li, and T. Mei, "Video Captioning with Transferred Semantic Attributes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6504–6512.

[38] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly Modeling Embedding and Translation to Bridge Video and Language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.

[39] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao, and T. Mei, "Retrieval Augmented Convolutional Encoder-Decoder Networks for Video Captioning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1s, pp. 1–24, 2023.

[40] T. S. Ferguson, *A Course in Game Theory*. World Scientific, 2020.

[41] J.-P. Jeukenne, A. Lejeune, and C. Mahaux, "Optical-model potential in finite nuclei from Reid's hard core interaction," *Physical Review C*, vol. 16, no. 1, p. 80, 1977.

[42] J. Sun, H. Yu, G. Zhong, J. Dong, S. Zhang, and H. Yu, "Random Shapley Forests: Cooperative Game Based Random Forests with Consistency," *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 205–214, 2020.

[43] E. Aflalo, M. Du, S.-Y. Tseng, Y. Liu, C. Wu, N. Duan, and V. Lal, "VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 406–21 415.

[44] A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," in *IEEE Symposium on Security and Privacy*, 2016, pp. 598–617.

[45] P. Jin, H. Li, Z. Cheng, J. Huang, Z. Wang, L. Yuan, C. Liu, and J. Chen, "Text-Video Retrieval with Disentangled Conceptualization and Set-to-Set Alignment," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, pp. 938–946.

[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[47] S. Kullback, *Information Theory and Statistics*. Courier Corporation, 1997.

[48] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowledge-Based Systems*, vol. 99, pp. 135–145, 2016.

[49] K. Li, Z. Wang, Z. Cheng, R. Yu, Y. Zhao, G. Song, C. Liu, L. Yuan, and J. Chen, "ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7162–7172.

[50] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2021, pp. 13 937–13 949.

[51] S.-V. Bogolin, I. Croitoru, H. Jin, Y. Liu, and S. Albanie, "Cross Modal Retrieval with Querybank Normalisation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5194–5205.

[52] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 214–229.

[53] X. Wang, L. Zhu, and Y. Yang, "T2VLAD: Global-Local Sequence Alignment for Text-Video Retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5079–5088.

[54] I. Croitoru, S.-V. Bogolin, M. Leordeanu, H. Jin, A. Zisserman, S. Albanie, and Y. Liu, "TEACHTEXT: CrossModal Generalized Distillation for Text-Video Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 583–11 593.

[55] M. Patrick, P.-Y. Huang, Y. Asano, F. Metze, A. G. Hauptmann, J. F. Henriques, and A. Vedaldi, "Support-set bottlenecks for video-text representation learning," in *Proceedings of the International Conference on Learning Representations*, 2021.

[56] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[57] S. K. Gorti, N. Vouitsis, J. Ma, K. Golestan, M. Volkovs, A. Garg, and G. Yu, "X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5006–5015.

[58] Y. Liu, P. Xiong, L. Xu, S. Cao, and Q. Jin, "TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 319–335.

[59] B. Fang, W. Wu, C. Liu, Y. Zhou, Y. Song, W. Wang, X. Shu, X. Ji, and J. Wang, "UATVR: Uncertainty-Adaptive Text-Video Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 723–13 733.

[60] C. Deng, Q. Chen, P. Qin, D. Chen, and Q. Wu, "Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 648–15 658.

[61] S. Zhao, L. Zhu, X. Wang, and Y. Yang, "CenterCLIP: Token Clustering for Efficient Text-Video Retrieval," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 970–981.

[62] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.

[63] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use What You Have: Video Retrieval Using Representations From Collaborative Experts," in *Proceedings of the British Machine Vision Conference*, 2019.

[64] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[65] Y. Yu, J. Kim, and G. Kim, "A Joint Sequence Fusion Model for Video Question Answering and Retrieval," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 471–487.

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

[67] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, "UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation," *arXiv preprint arXiv:2002.06353*, 2020.

[68] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[69] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[70] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[71] Y. Matsui and T. Matsui, "Np-completeness for calculating power indices of weighted majority games," *Theoretical Computer Science*, vol. 263, no. 1, pp. 305–310, 2001.

[72] Y. Bachrach, E. Markakis, E. Resnick, A. D. Procaccia, J. S. Rosenschein, and A. Saberi, "Approximating power indices: theoretical and empirical analysis," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 105–122, 2010.

[73] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, Y. Shan, X. Qie, and M. Z. Shou, "All in One: Exploring Unified Video-Language Pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6598–6608.

[74] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Zero-Shot Video Question Answering via Frozen Bidirectional Language Models," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022, pp. 124–141.

[75] Y. Wang, M. Liu, J. Wu, and L. Nie, "Multi-Granularity Interaction and Integration Network for Video Question Answering," *IEEE*

Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7684–7695, 2023.

[76] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant Grounding for Video Question Answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2928–2937.

[77] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Learning to Answer Visual Questions from Web Videos," *arXiv preprint arXiv:2205.05019*, 2022.

[78] J. Liu, G. Wang, J. Xie, F. Zhou, and H. Xu, "Video Question Answering With Semantic Disentanglement and Reasoning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3663–3673, 2024.

[79] Y. Li, K. Min, S. Tripathi, and N. Vasconcelos, "SViTT: Temporal Learning of Sparse Video-Text Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 919–18 929.

[80] H. Li, P. Jin, Z. Cheng, S. Zhang, K. Chen, Z. Wang, C. Liu, and J. Chen, "TG-VQA: Ternary Game of Video Question Answering," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023, pp. 1044–1052.

[81] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "SWINBERT: End-to-End Transformers with Sparse Attention for Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[82] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end Generative Pretraining for Multimodal Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 959–17 968.

[83] S. Chen and Y.-G. Jiang, "Motion Guided Region Message Passing for Video Captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1543–1552.

[84] Z. Zhang, Z. Qi, C. Yuan, Y. Shan, B. Li, Y. Deng, and W. Hu, "Open-book Video Captioning with Retrieve-Copy-Generate Network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9837–9846.

[85] J. Perez-Martin, B. Bustos, and J. Pérez, "Attentive Visual Semantic Specialized Network for Video Captioning," in *Proceedings of the International Conference on Pattern Recognition*, 2021, pp. 5767–5774.

[86] X. Luo, X. Luo, D. Wang, J. Liu, B. Wan, and L. Zhao, "Global semantic enhancement network for video captioning," *Pattern Recognition*, vol. 145, p. 109906, 2024.

[87] Y. Shen, X. Gu, K. Xu, H. Fan, L. Wen, and L. Zhang, "Accurate and Fast Compressed Video Captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 558–15 567.

[88] P. Song, D. Guo, X. Yang, S. Tang, and M. Wang, "Emotional Video Captioning with Vision-based Emotion Interpretation Network," *IEEE Transactions on Image Processing*, vol. 33, pp. 1122–1135, 2024.

[89] J. Perez-Martin, B. Bustos, and J. Perez, "Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3039–3049.

[90] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, and X. Li, "CLIP4Caption: CLIP for Video Caption," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 4858–4862.

[91] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

Hao Li received the BE degree in the School of Computer Science from Peking University, China, in 2021. He is currently working toward a Ph.D. degree at Peking University, Beijing, China. His research interests include cross-modality learning and its downstream tasks: video-text retrieval, image caption, and visual question answering.

Li Yuan received the B.Eng. degree from University of Science and Technology of China in 2017, and Ph.D. degree from National University of Singapore in 2021. He is currently a tenure-track Assistant Professor with School of Electrical and Computer Engineering at Peking University. He has published more than 20 papers on top conferences/journals. His research interests include deep learning, image processing, and computer vision.

Shuicheng Yan (Fellow, IEEE) is currently the Honorary Advisor of Kunlun Tech and was formerly the Chief Scientist at Sea. He is a member of the Singapore Academy of Engineering and a Fellow of AAAI, ACM, IEEE, and IAPR, and his research interests include computer vision, machine learning, and multimedia analytics. To date, Prof. Yan has published more than 800 papers in top international journals and conferences with an H-index of 140+. Prof. Yan Shuicheng's team has won more than ten awards in two core competitions, Pascal VOC and ImageNet (ILSVRC). In addition, his team has won more than ten best paper and best student paper awards, especially a grand slam at ACM Multimedia, the top conference in multimedia, including three best paper awards, two best student paper awards, and a best demo award.

Peng Jin received the BE degree in the School of Electronic Information from Tsinghua University, China, in 2021. He is currently working toward a Ph.D. degree at Peking University, Beijing, China. He has published more than 10 papers on top international journals and conferences. His research interests include cross-modality learning and its downstream tasks: multimodal LLMs, video-text retrieval, image caption, and visual question answering.

Jie Chen (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Harbin Institute of Technology, China, in 2002 and 2007, respectively. He joined as the Faculty Member of the Shenzhen Graduate School, Peking University, in 2019. He is currently an Associate Professor with the School of Electronic and Computer Engineering, at Peking University. Since 2018, he has been working with the Pengcheng Laboratory, in China. From 2007 to 2018, he worked as a Senior Researcher with the Center for Machine Vision and Signal Analysis, the University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory, University of Maryland, and School of Electrical and Computer Engineering, Duke University, respectively. He was a Co-Chair of International Workshops at ACCV, ACM MM, CVPR, ICCV, and ECCV. He was a Guest Editor of special issues for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, and Neurocomputing. He has been selected as a finalist for the 2022 Gordon Bell Special Prize for HPC-Based COVID-19 Research. His research interests include computer vision, large language model and its application for protein engineering. He has published over 200 papers in top international journals and conferences. He is an Associate Editor of the Visual Computer.