Video-Text as Game Players: Hierarchical Banzhaf Interaction for Cross-Modal Representation Learning

Peng Jin[1,3] Jinfa Huang[1,3] Pengfei Xiong[4] Shangxuan Tian[4] Chang Liu[5]
Xiangyang Ji[5] Li Yuan[1,2,3*] Jie Chen[1,2,3*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China
[4]Shopee, Shenzhen, China [5]Department of Automation and BNRist, Tsinghua University, Beijing, China
{jp21, jinfahuang}@stu.pku.edu.cn xiongpengfei@gmail.com tianshangxuan@u.nus.edu
{liuchang2022, xyji}@tsinghua.edu.cn chenj@pcl.ac.cn yuanli-ece@pku.edu.cn

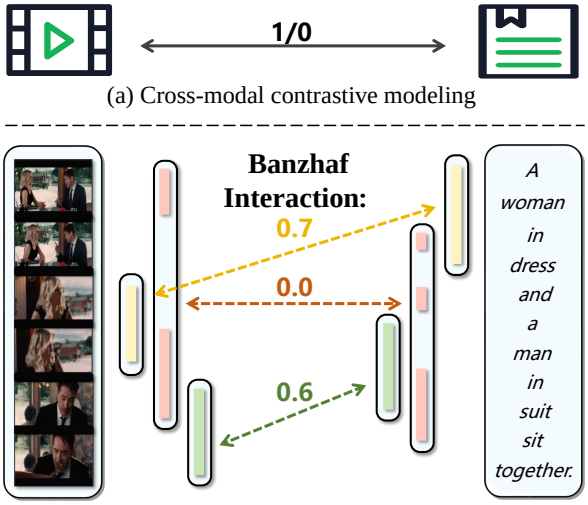
Abstract

Contrastive learning-based video-language representation learning approaches, e.g., CLIP, have achieved outstanding performance, which pursue semantic interaction upon pre-defined video-text pairs. To clarify this coarse-grained global interaction and move a step further, we have to encounter challenging shell-breaking interactions for fine-grained cross-modal learning. In this paper, we creatively model video-text as game players with multivariate cooperative game theory to wisely handle the uncertainty during fine-grained semantic interaction with diverse granularity, flexible combination, and vague intensity. Concretely, we propose Hierarchical Banzhaf Interaction (HBI) to value possible correspondence between video frames and text words for sensitive and explainable cross-modal contrast. To efficiently realize the cooperative game of multiple video frames and multiple text words, the proposed method clusters the original video frames (text words) and computes the Banzhaf Interaction between the merged tokens. By stacking token merge modules, we achieve cooperative games at different semantic levels. Extensive experiments on commonly used text-video retrieval and video-question answering benchmarks with superior performances justify the efficacy of our HBI. More encouragingly, it can also serve as a visualization tool to promote the understanding of cross-modal interaction, which have a far-reaching impact on the community. Code is available at https://github.com/jpthu17/HBI.

1. Introduction

Representation learning based on both vision and language has many potential benefits and direct applicability to cross-modal tasks, such as text-video retrieval [26,32,60] and video-question answering [42,72,80]. Visual-language learning has recently boomed due to the success of contrastive learning [12–14,31,71,86–89], *e.g.*, CLIP [58], to project the video and text features into a common latent space according to the semantic similarities of video-text pairs. In this manner, cross-modal contrastive learning enables networks to learn discriminative video-language representations.

The cross-modal contrastive approach [8, 17, 21, 30, 66] typically models the cross-modal interaction via solely the global similarity of each modality. Specifically, as shown in



(a) Cross-modal contrastive modeling

(b) Multivariate cooperative game modeling (Ours)

Figure 1. (a) Cross-modal contrastive methods only learn a global semantic interaction from the coarse-grained labels of video-text pairs. (b) We model cross-modal alignment as a multivariate cooperative game process. Specifically, we use Banzhaf Interaction to value possible correspondence between video frames and text words and consider it as an additional learning signal.

*Corresponding author: Li Yuan, Jie Chen.

Fig. 1a, it only exploits the coarse-grained labels of video-text pairs to learn a global semantic interaction. However, in most cases, we expect to capture fine-grained interpretable information, such as how much cross-modal alignment is helped or hindered by the interaction of a visual entity and a textual phrase. Representation that relies on cross-modal contrastive learning cannot do this in a supervised manner, as manually labeling these interpretable relationships is unavailable, especially on large-scale datasets. This suggests that there might be other learning signals that could complement and improve pure contrastive formulations.

In contrast to prior works [18, 25, 47, 65], we model cross-modal representation learning as a multivariate cooperative game by formulating video and text as players in a cooperative game, as illustrated in Fig. 1b. Intuitively, if visual representations and textual representations have strong semantic correspondence, they tend to cooperate together and contribute to the cross-modal similarity score. Motivated by this spirit, we consider the set containing multiple representations as a coalition, and propose to quantify the trend of cooperation within a coalition via the game-theoretic interaction index, *i.e.*, Banzhaf Interaction [29] for its simplicity and efficiency. Banzhaf Interaction is one of the most popular concepts in cooperative games [48]. As shown in Fig. 2, it measures the additional benefits brought by the coalition compared with the costs of the lost coalitions of these players with others. When a coalition has high Banzhaf Interaction, it will also have a high contribution to the semantic similarity. Thus, we can use Banzhaf Interaction to value possible correspondence between video frames and text words for sensitive and explainable cross-modal contrast.

To this end, we propose Hierarchical Banzhaf Interaction (HBI). Concretely, we take video frames and text words as players and the cross-modality similarity measurement as the characteristic function in the cooperative game. Then, we use the Banzhaf Interaction to represent the trend of cooperation between any set of features. Besides, to efficiently generate coalitions among game players, we propose an adaptive token merge module to cluster the original video frames (text words). By stacking token merge modules, we achieve hierarchical interaction, *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. In particular, we show that the Banzhaf Interaction index satisfies *Symmetry*, *Dummy*, *Additivity*, and *Recursivity* axiom in Sec. 3.4. This result implies that the representation learned via Banzhaf Interaction has four properties that the features of the contrastive method do not. We find that explicitly establishing the fine-grained interpretable relationships between video and text brings a sensible improvement to already very strong video-language representation learning results. Experiment results on three text-video retrieval benchmark datasets (*MSRVTT* [73], *ActivityNet Captions* [34], and *DiDeMo* [2]) and the video question answering benchmark dataset (*MSRVTT-QA* [72]) show the advantages of the proposed method. The main contributions are as follows:

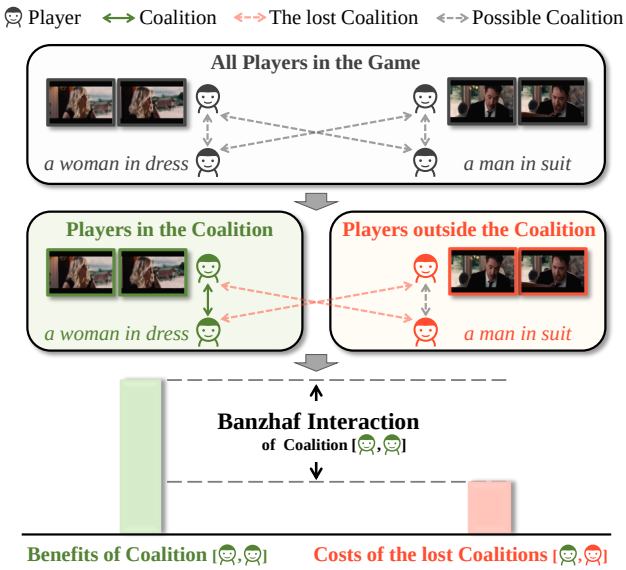


Figure 2. **The intuition of Banzhaf Interaction in video-text representation learning.** We refer the reader to Eq. 3 for the detailed formula. When some players (frames and words) form a coalition, we lose the coalitions of these players with others. In other words, the lost coalition is mutually exclusive from the target coalition. Banzhaf Interaction measures the difference between the benefits of the coalition and the costs of the lost coalitions.

- To the best of our knowledge, we are the first to model video-language learning as a multivariate cooperative game process and propose a novel proxy training objective, which uses Banzhaf interaction to value possible correspondence between video frames and text words for sensitive and explainable cross-modal contrast.
- Our method achieves new state-of-the-art performance on text-video retrieval benchmarks of *MSRVTT*, *ActivityNet Captions* and *DiDeMo*, as well as on the video-question answering task on *MSRVTT-QA*.
- More encouragingly, our method can also serve as a visualization tool to promote the understanding of cross-modal interaction, which may have a far-reaching impact on the community.

2. Related Work

Cooperative Game Theory. The cooperative game theory consists of a set of players with a characteristic function [9, 52]. 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 the cooperative game theory is to allocate different payoffs to game individuals fairly and reasonably. Game theory has found many applications in the field of model interpretability [1, 16, 78, 85], but there is little exploration in cross-modal learning. Banzhaf Interaction is one of the most popular concepts in cooperative games [48]. Recently, LOUPE [41] uses two-player interaction as a vision-language pre-training task. In this paper, we design a new framework of multivariate interaction for video-text representation learning. Besides, our method can be directly co-trained with target task losses for high flexibility.

Visual-Language Learning. Recently, contrastive learning methods show great success in cross-modal tasks [7, 11, 20, 27, 47, 56, 68, 75], such as text-video retrieval [45, 50, 69, 70, 77, 90] and video-question answering [23, 53]. Text-video retrieval [17, 61, 68] requires the model to map text and video to the same latent space, where the similarity between them can be directly calculated [10, 57, 77]. Video-question answering [24, 36, 74] requires the model to predict an answer using visual information [39, 40, 82]. Due to manually labeling the fine-grained relationships being unavailable, cross-modal contrastive learning cannot capture fine-grained information in a supervised manner. To this end, we model video-text as game players with multivariate cooperative game theory and propose to combine Banzhaf Interaction with cross-modal contrastive learning. In contrast to prior works, we explicitly capture the fine-grained semantic relationships between video frames and text tokens via Banzhaf Interaction. Then, we use these relationships as additional learning signals to improve pure contrastive learning.

3. Method

3.1. Multivariate Cooperative Game Modeling

3.1.1 Video-Language Learning

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

To learn fine-grained semantic alignment, the input video $\boldsymbol{v}$ is embedded into frame sequence $\boldsymbol{V}_f = \{v_f^i\}_{i=1}^{N_v}$, where N_v is the length of video $\boldsymbol{v}$. The input text $\boldsymbol{t}$ is embedded into word sequence $\boldsymbol{T}_w = \{t_w^j\}_{j=1}^{N_t}$, where N_t is the length of text $\boldsymbol{t}$. Then, the alignment matrix is defined as: $A = [a_{ij}]^{N_v \times N_t}$, where $a_{ij} = \frac{(v_f^i)^T t_w^j}{\|v_f^i\| \|t_w^j\|}$ represents the alignment score between i_{th} video frame and j_{th} text word. For the i_{th} video frame, we calculate its maximum alignment score as $\max_j a_{ij}$. Then, 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 [65] can be defined as:

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

where $[\omega_v^0, \omega_v^1, ..., \omega_v^{N_v}] = \mathrm{Softmax}(\mathrm{MLP}_v(\boldsymbol{V}_f))$ and $[\omega_t^0, \omega_t^1, ..., \omega_t^{N_t}] = \mathrm{Softmax}(\mathrm{MLP}_t(\boldsymbol{T}_w))$ are the weights of the video frames and text words, respectively. Then the cross-modal contrastive loss [63] can be formulated as:

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

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.

Prior works typically directly apply the cross-modal contrastive loss to optimize the similarity scores $\mathrm{S}_{\boldsymbol{v},\boldsymbol{t}}$. 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.

3.1.2 Banzhaf Interaction

We start by introducing notation and outlining assumptions about the cooperative game theory. Then, we review Banzhaf Interaction [29] for a cooperative game.

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 [62].

In a cooperative game, some players tend to form a coalition: it may happen that $\phi(\{i\})$ and $\phi(\{j\})$ are small, and at the same time $\phi(\{i, j\})$ is large. The Banzhaf Interaction [29] 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 $\{i, j\}$. Then, the reduced 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 [29].*** *Given a coalition $\{i, j\} \subseteq \mathcal{N}$, the Banzhaf Interaction $\mathcal{I}([\{i, j\}])$ for the*

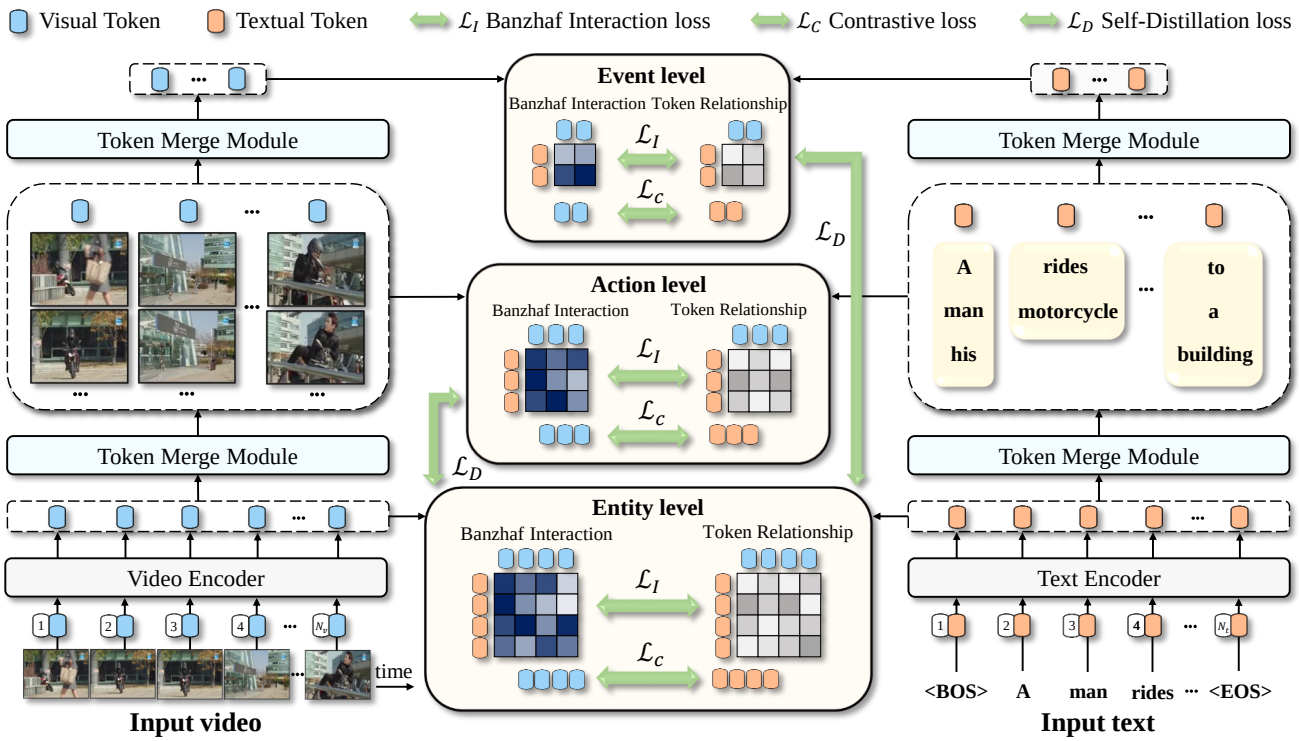


Figure 3. **The overall framework of HBI.** We propose a novel proxy training objective, which uses Banzhaf interaction to value possible correspondence between video frames and text words, and enhance cross-modal representation learning. By stacking token merge modules, we achieve hierarchical interaction, *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. To improve the generalization ability, we use the additional self-distillation loss. The calculation of the exact Banzhaf Interaction is an NP-hard problem. 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 (Sec. 4.1).

player $[\{i,j\}]$ *is defined as:*

$$\mathcal{I}([\{i,j\}]) = \sum_{\mathcal{C}\subseteq\mathcal{N}\setminus\{i,j\}} p(\mathcal{C})[\phi(\mathcal{C}\cup\{[\{i,j\}]\}) + \phi(\mathcal{C}) - \phi(\mathcal{C}\cup\{i\}) - \phi(\mathcal{C}\cup\{j\})], \tag{3}$$

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.

3.1.3 Video-Text as Game Players

Given features $\boldsymbol{V}_f = \{v_f^i\}_{i=1}^{N_v}$ and $\boldsymbol{T}_w = \{t_w^j\}_{j=1}^{N_t}$, 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 the 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 anything satisfying the above conditions can be used as the characteristic function ϕ. For simplicity, we use cross-modality similarity measurement S as ϕ. Then, we can use Banzhaf Interaction to value possible correspondence between video frames and text words, and to enhance cross-modal representation learning.

3.2. Hierarchical Banzhaf Interaction

In the following, we first introduce the simple two-player interaction between a video frame and a word token. Then, we expand the two-player interaction to the multivariate interaction via the token merge module. Fig. 3 illustrates the overall framework of our method.

For a coalition $\{v_f^i, t_w^j\}$, referring to Eq. 3, we can cal-

culate 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. We provide the experiment results of the prediction header with different structures in Tab. 4.

Then, we optimize the Kullback-Leibler (KL) divergence [35] 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:

$$\begin{aligned} \mathcal{D}_{v2t}^{\mathcal{I}} &= [p_{i,1}^{\mathcal{I}}, p_{i,2}^{\mathcal{I}}, ..., p_{i,N_t}^{\mathcal{I}}], \\ \mathcal{D}_{t2v}^{\mathcal{I}} &= [\hat{p}_{1,j}^{\mathcal{I}}, \hat{p}_{2,j}^{\mathcal{I}}, ..., \hat{p}_{N_v,j}^{\mathcal{I}}], \end{aligned} \tag{4}$$

where $p_{i,j}^{\mathcal{I}} = \frac{\exp(\mathcal{I}([\{v_f^i, t_w^j\}]))}{\sum_{k=1}^{N_t} \exp(\mathcal{I}([\{v_f^i, t_w^k\}]))}$, $\hat{p}_{i,j}^{\mathcal{I}} = \frac{\exp(\mathcal{I}([\{v_f^i, t_w^j\}]))}{\sum_{k=1}^{N_v} \exp(\mathcal{I}([\{v_f^k, t_w^j\}]))}$. Similarly, the probability distribution $\mathcal{D}_{v2t}^{\mathcal{R}}$ and $\mathcal{D}_{t2v}^{\mathcal{R}}$ are calculated in the same way using $\mathcal{R}^{i,j}$, *i.e.*, $\mathcal{D}_{v2t}^{\mathcal{R}} = [p_{i,1}^{\mathcal{R}}, p_{i,2}^{\mathcal{R}}, ..., p_{i,N_t}^{\mathcal{R}}], \mathcal{D}_{t2v}^{\mathcal{R}} = [\hat{p}_{1,j}^{\mathcal{R}}, \hat{p}_{2,j}^{\mathcal{R}}, ..., \hat{p}_{N_v,j}^{\mathcal{R}}]$, where $p_{i,j}^{\mathcal{R}} = \frac{\exp(\mathcal{R}_{i,j})}{\sum_{k=1}^{N_t} \exp(\mathcal{R}_{i,k})}, \hat{p}_{i,j}^{\mathcal{R}} = \frac{\exp(\mathcal{R}_{i,j})}{\sum_{k=1}^{N_v} \exp(\mathcal{R}_{k,j})}$. Finally, the Banzhaf Interaction loss $\mathcal{L}_I$ is defined as:

$$\mathcal{L}_I = \mathbb{E}_{\boldsymbol{v},\boldsymbol{t}}[\mathrm{KL}(\mathcal{D}_{v2t}^{\mathcal{R}} \| \mathcal{D}_{v2t}^{\mathcal{I}}) + \mathrm{KL}(\mathcal{D}_{t2v}^{\mathcal{R}} \| \mathcal{D}_{t2v}^{\mathcal{I}})]. \tag{5}$$

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.

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. 4 illustrates the framework of the token merge module.

Specifically, we utilize DPC-KNN [19], a k-nearest neighbor-based density peaks clustering algorithm, to cluster the visual (textual) tokens. Starting with the frame-level tokens $\boldsymbol{V}_f = \{v_f^i\}_{i=1}^{N_v}$, we first use a one-dimensional convolutional layer to enhance the temporal information between tokens. Then, we compute the local density ρ_i of each token v_f^i according to its K-nearest neighbors:

$$\rho_i = \exp(-\frac{1}{K} \sum_{v_f^k \in \mathrm{KNN}(v_f^i)} \|v_f^k - v_f^i\|^2), \tag{6}$$

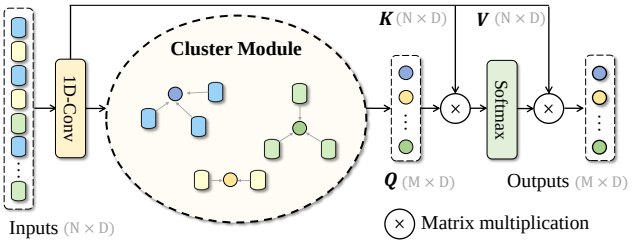


Figure 4. **The token merge module.** "1D-Conv" denotes the one-dimensional convolutional layer. N input tokens with D channels are first clustered into M clusters. Then, we feed the merged tokens as Q and the original tokens as K, V into an attention module.

where $\mathrm{KNN}(v_f^i)$ is the K-nearest neighbors of v_f^i. After that, we compute the distance index δ_i of each token v_f^i:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \|v_f^k - v_f^i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i. \\ \max\limits_{j} \|v_f^k - v_f^i\|^2, & \text{otherwise.} \end{cases} \tag{7}$$

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

We consider those tokens with relatively high $\rho_i \times \delta_i$ as cluster centers, and then assign other tokens to the nearest cluster center according to the Euclidean distances. Inspired by [59, 83], we use the weighted average tokens of each cluster to represent the corresponding cluster, where the weight $W = \mathrm{Softmax}(\mathrm{MLP}_w(\boldsymbol{V}_f))$. Then, we feed the weighted average tokens as queries Q and the original tokens as keys K and values V into an attention module. We treat the output of the attention 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. Training Objective

Combining the cross-modal contrastive loss $\mathcal{L}_C$ and Banzhaf Interaction loss $\mathcal{L}_I$, the full objective of semantic alignment can be 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,

$$\mathcal{L}^e = \mathcal{L}_C^e + \alpha\mathcal{L}_I^e, \quad \mathcal{L}^a = \mathcal{L}_C^a + \alpha\mathcal{L}_I^a, \quad \mathcal{L}^o = \mathcal{L}_C^o + \alpha\mathcal{L}_I^o, \tag{8}$$

where $\mathcal{L}^e$, $\mathcal{L}^a$, and $\mathcal{L}^o$ 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 $\mathrm{S}_{\boldsymbol{v},\boldsymbol{t}}^e$ converges first in the training process, so we distill the entity-level similarity to the other two semantic levels. The analyses and experiments are provided in Appendix.

Starting with entity-level similarity $\mathrm{S}_{\boldsymbol{v},\boldsymbol{t}}^e$ distilling to action-level similarity $\mathrm{S}_{\boldsymbol{v},\boldsymbol{t}}^a$, we first compute the distribution $\mathcal{D}_{v2t}^e$ and $\mathcal{D}_{t2v}^e$ by replacing $\mathcal{I}([\{v,t\}])$ with $\mathrm{S}_{\boldsymbol{v},\boldsymbol{t}}^e$ in

Figure 5. **Visualization of the hierarchical interaction.** We take Video7060 in the MSRVTT as an example. We provide more visualizations in the Appendix. Here, the degree of confidence from high to low is represented by red, orange, green and blue lines, respectively.

Eq. 4. The distribution $\mathcal{D}^a_{v2t}$ and $\mathcal{D}^a_{t2v}$ are calculated using $\mathrm{S}^a_{\boldsymbol{v},\boldsymbol{t}}$. The $\mathcal{L}^{e2a}_D$ loss is defined as:

$$\mathcal{L}^{e2a}_D = \mathbb{E}_{\boldsymbol{v},\boldsymbol{t}}[\mathrm{KL}(\mathcal{D}^a_{v2t} \| \mathcal{D}^e_{v2t}) + \mathrm{KL}(\mathcal{D}^a_{t2v} \| \mathcal{D}^e_{t2v})]. \quad (9)$$

The $\mathcal{L}^{e2o}_D$ loss from entity-level similarity to event-level similarity is calculated in the same way.

The overall loss is the combination of semantically alignment losses and self-distillation losses, which is defined as:

$$\mathcal{L}_{total} = \underbrace{\mathcal{L}^e + \mathcal{L}^a + \mathcal{L}^o}_{\text{deep supervision}} + \beta \underbrace{(\mathcal{L}^{e2a}_D + \mathcal{L}^{e2o}_D)}_{\text{self-distillation}}, \quad (10)$$

where β is the trade-off hyper-parameter. We provide the ablation experiments for each part of the loss function in Tab. 5. We find that Banzhaf Interaction loss $\mathcal{L}_I$ significantly improves the performance, while deep supervision and self-distillation can improve the generalization ability.

3.4. Theoretical Analysis

Similar to Banzhaf value axioms [29], the following axioms convey intuitive properties that a cross-modal interaction score should satisfy.

Axioms 1. *Given a set $\mathcal{N} = \{1, 2, ..., n\}$ of players, a characteristic function $\phi : 2^n \to \mathbb{R}$, and a coalition $\mathcal{C} = \{i, j\} \subseteq \mathcal{N}$, following properties are met for the interaction score $\mathcal{I}([\mathcal{C}])$. (a)* ***Symmetry:*** *If $\forall \mathcal{S} \subseteq \mathcal{N}$, $\phi(\mathcal{S} \cup \{[\mathcal{C}]\}) = \phi(\mathcal{S} \cup \{[\mathcal{C}']\}), \sum_{i \in \mathcal{C}} \phi(\mathcal{S} \cup \{i\}) = \sum_{i' \in \mathcal{C}'} \phi(\mathcal{S} \cup \{i'\})$, then $\mathcal{I}([\mathcal{C}]) = \mathcal{I}([\mathcal{C}'])$; (b)* ***Dummy:*** *If $\forall \mathcal{S} \subseteq \mathcal{N}$, $\phi(\mathcal{S} \cup \{[\mathcal{C}]\}) = \phi(\mathcal{S}), \sum_{i \in \mathcal{C}} \phi(\mathcal{S} \cup i) = 0$, then $\mathcal{I}([\mathcal{C}]) = 0$; (c)* ***Additivity:*** *If $\phi(*)$ and $\phi'(*)$ have the interaction scores $\mathcal{I}([\mathcal{C}])$ and $\mathcal{I}'([\mathcal{C}])$ respectively, then the interaction score for the game with value function $\phi(*) + \phi'(*)$ is $\mathcal{I}([\mathcal{C}]) + \mathcal{I}'([\mathcal{C}])$; (d)* ***Recursivity:*** *let $\mathcal{B}(*)$ denote the Banzhaf value [5], then $\mathcal{B}([\mathcal{C}]|\mathcal{N} \backslash \mathcal{C} \cup \{[\mathcal{C}]\}) = \mathcal{B}(i|\mathcal{N} \backslash \{j\}) + \mathcal{B}(j|\mathcal{N} \backslash \{i\}) + \mathcal{I}([\mathcal{C}])$.*

Method	Receptive field	Robustness	Flexibility
Cosine similarity	Element level	Absolute value	Non-adjustable
Banzhaf Interaction	**Set level**	**Relative value**	**Adaptable ϕ**

Table 1. **Comparison of Banzhaf Interaction and similarity.**

Symmetry states that if changing the value of two coalitions has the same effect on the output under all values of the other variables, then both coalitions should have an identical interaction score. **Dummy** states that if changing the value of a coalition $[\mathcal{C}]$ has no effect on the output under all values of other variables, then the interaction value of $[\mathcal{C}]$ should be zero. **Additivity** states the sum of the interaction scores of the two characteristic functions is equal to the interaction score of the sum of these characteristic functions. **Recursivity** states that if the interaction is positive, then the interaction score of $[\{i, j\}]$ should be greater than simply the sum of individual values. If the interaction is negative, the interaction score of $[\{i, j\}]$ should be less than the sum.

Theorem 1. *The Banzhaf Interaction index satisfies* ***Symmetry***, ***Dummy***, ***Additivity*** *and* ***Recursivity*** *axiom.*

We refer the reader to Appendix for more detail about Theorem 1. This result implies that the representation learned via Banzhaf Interaction has four properties that the features of the contrastive method do not. Besides, we compare Banzhaf Interaction and cosine similarity in Tab. 1, mainly

MSRVTT					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MMT [22]	26.6	57.1	69.6	4.0	24.0
T2VLAD [67]	29.5	59.0	70.1	4.0	-
Support-Set [54]	30.1	58.5	69.3	3.0	-
CLIP4Clip [47]	44.5	71.4	81.6	**2.0**	15.3
EMCL-Net [33]	46.8	73.1	83.1	**2.0**	-
X-Pool [28]	46.9	72.8	82.2	**2.0**	14.3
TS2-Net [46]	47.0	74.5	**83.8**	**2.0**	13.0
HBI (Ours)	**48.6**	**74.6**	83.4	**2.0**	**12.0**

ActivityNet Captions					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
ClipBERT [38]	21.3	49.0	63.5	6.0	-
T2VLAD [67]	23.7	55.5	-	4.0	-
MMT [22]	28.7	61.4	-	3.3	16.0
Support-Set [54]	29.2	61.6	-	3.0	-
CLIP4Clip [47]	40.5	72.4	83.6	**2.0**	7.5
TS2-Net [46]	41.0	72.6	84.5	**2.0**	8.4
EMCL-Net [33]	41.2	72.7	-	**2.0**	-
HBI (Ours)	**42.2**	**73.0**	**84.6**	2.0	**6.6**

DiDeMo					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
FSE [84]	13.9	36.0	11.0	-	-
CE [44]	16.1	41.1	6.0	-	43.7
ClipBERT [38]	20.4	48.0	60.8	6.0	-
TT-CE [15]	21.6	48.6	62.9	6.0	-
Frozen [4]	34.6	65.0	74.7	3.0	-
TS2-Net [46]	41.8	71.6	82.0	**2.0**	14.8
CLIP4Clip [47]	42.8	68.5	79.2	**2.0**	18.9
HBI (Ours)	**46.9**	**74.9**	**82.7**	2.0	**12.1**

(a) Retrieval performance on the **Text->Video** task.

MSRVTT					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
T2VLAD [67]	31.8	60.0	71.1	3.0	-
HiT [43]	32.1	62.7	74.1	3.0	-
CLIP4Clip [47]	42.7	70.9	80.6	**2.0**	11.6
X-Pool [28]	44.4	73.3	84.0	**2.0**	9.0
TS2-Net [46]	45.3	74.1	83.7	**2.0**	9.2
HBI (Ours)	**46.8**	**74.3**	**84.3**	2.0	**8.9**

ActivityNet Captions					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
HSE [84]	18.7	48.1	-	-	-
T2VLAD [67]	24.1	56.6	-	4.0	-
Support-Set [54]	28.7	60.8	-	**2.0**	-
MMT [22]	28.9	61.1	-	4.0	17.1
CLIP4Clip [47]	41.4	**73.7**	85.3	**2.0**	6.7
HBI (Ours)	**42.4**	73.0	**86.0**	2.0	**6.5**

DiDeMo					
Methods	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
FSE [84]	13.1	33.9	12.0	-	
S2VT [64]	13.2	33.6	-	15.0	-
CE [44]	15.6	40.9	-	8.2	42.4
TT-CE [15]	21.1	47.3	61.1	6.3	-
CLIP4Clip [47]	41.4	68.2	79.1	**2.0**	12.4
HBI (Ours)	**46.2**	**73.0**	**82.7**	2.0	**8.7**

(b) Retrieval performance on the **Video->Text** task.

Table 2. **Comparisons to current state-of-the-art methods on the MSRVTT [73], ActivityNet Captions [34] and DiDeMo [2] datasets.** "↑" denotes that higher is better. "↓" denotes that lower is better. All results in this table do not use inverted softmax [6].

in three aspects. **(1) Global receptive field.** In contrast to cosine similarity, which only operates at the element level, Banzhaf Interaction operates at the set level to leverage the global context. **(2) Robustness.** Cosine similarity fluctuates by visual and language style. In contrast, Banzhaf Interaction measures the relative value of benefit and opportunity cost to be robust to the style deviation. **(3) Flexibility.** Our framework can use other characteristic functions ϕ besides similarity, which is left for future work to explore. Therefore, Banzhaf Interaction is a promising interaction score to enhance cross-modal representation learning.

4. Experiments

4.1. Experimental Settings

Datasets. **MSRVTT** [73] contains 10K YouTube videos, each with 20 text descriptions. We follow the training protocol in [22, 44, 51] and evaluate on the 1K-A testing split [81]. **ActivityNet Captions** [34] consists of densely annotated temporal segments of 20K YouTube videos. We use the 10K training split to train the model and report the performance on the 5K "val1" split. **DiDeMo** [2] contains 10K videos annotated 40K text descriptions. We follow the training and evaluation protocol in [47]. **MSRVTT-QA** [72] is based on the MSRVTT and has 243K VideoQA pairs.

Metrics. We choose Recall at rank K (R@K), Median Rank (MdR), and mean rank (MnR) [22] to evaluate the retrieval performance. We choose answer accuracy to evaluate the video question answering performance.

Implementation Details. Since the calculation of the exact Banzhaf Interaction is an NP-hard problem [49], existing methods mainly use sampling-based methods [3, 37] 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. We refer the reader to Appendix for the details. For text-video retrieval, we utilize the CLIP (ViT-B/32) [58] as the pre-trained model. 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. More details are in the Appendix.

4.2. Comparison with State-of-the-art

In Tab. 2, we show the results of our method on MSRVTT, ActivityNet Captions, and DiDeMo datasets. Our model consistently outperforms the recently proposed state-of-the-art methods on both text-to-video retrieval and video-to-text retrieval tasks. Tab. 3 shows the results of our method for video-question answering. Massive experiments on text-video retrieval and video-question answering tasks demonstrate the superiority and flexibility of our method.

4.3. Ablation Study

Effect of the prediction header of $\mathcal{R}$. To explore the impact of the structure of the prediction header on our method, we compare several popular structures in Tab. 4. We find that the combination of CNN and attention ("CNN+SA") can capture both local and global interaction, so it is beneficial for predicting the fine-grained relationship.

Ablation about components. As shown in Tab. 5, Banzhaf Interaction boosts the baseline with the improve-

Methods	Accuracy (%)↑
VQA-T [76]	41.5
SiaSamRea [79]	41.6
MERLOT [82]	43.1
Co-Tokenization [55]	45.7
EMCL-QA [33]	45.8
HBI (Ours)	**46.2**

Table 3. **Video-question answering performance on MSRVTT-QA dataset.**

Method	Text->Video			
	R@1↑	R@5↑	R@10↑	MnR↓
Baseline	46.6	73.1	83.0	13.3
MLP	47.2	73.7	83.5	12.3
CNN	47.3	73.5	**83.7**	12.2
MLP+SA	46.6	74.0	**83.7**	12.3
CNN+SA	**48.6**	**74.6**	83.4	**12.0**

Table 4. **Effect of the prediction header on MSRVTT dataset.** "SA" is the self-attention module.

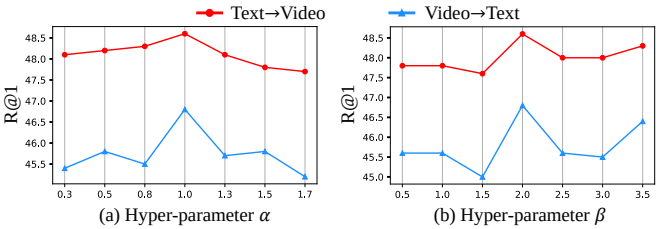


Figure 6. **Effect of the hyper-parameters on MSRVTT dataset.** α and β are the hyper-parameters in Eq. 8 and Eq. 9, respectively.

$\mathcal{L}_I$ Banzhaf Interaction	Deep Supervision	$\mathcal{L}_D$ Self Distillation	Text->Video			
			R@1↑	R@5↑	R@10↑	MnR↓
			46.6	73.1	83.0	13.3
✓			47.4	74.2	82.8	12.1
	✓		47.2	74.1	82.6	12.0
	✓	✓	47.6	73.8	83.2	**11.9**
✓	✓		48.2	73.0	83.1	12.0
✓	✓	✓	**48.6**	**74.6**	**83.4**	12.0

Table 5. **Ablation study about the importance of each part of our method on MSRVTT dataset.**

N_v^a	N_v^o	N_t^a	N_t^o	Text->Video				
				R@1↑	R@5↑	R@10↑	Rsum↑	MnR↓
-	-	-	-	47.5	73.7	83.0	204.2	**12.0**
9	3	18	4	48.2	**75.2**	82.7	206.1	12.4
6	3	12	4	48.3	74.3	83.1	205.7	12.3
6	2	12	3	**48.7**	74.5	82.6	205.8	12.2
3	2	6	3	48.6	74.6	**83.4**	**206.6**	**12.0**

Table 6. **The efficiency of the cluster module.** $N_{\_}^a$ and $N_{\_}^o$ denote the number of clusters at the action level and event level, respectively.

Method	Iteration Time↓	Inference Time↓
CLIP4Clip [47]	**1.63** s	**16.28** s
DRL [65]	1.65 s	16.74 s
EMCL-Net [33]	1.72 s	17.68 s
TS2-Net [46]	2.57 s	19.91 s
Baseline	2.06 s	18.06 s
HBI (Ours)	3.14 s	19.17 s

Table 7. **Time consumption on MSRVTT dataset.**

ment up to 0.8% at R@1. Moreover, deep supervision and self-distillation significantly improve the generalization ability. Our full model achieves the best performance and outperforms the baseline by 2.0% at R@1 for text-to-video retrieval. This demonstrates that the three parts are beneficial for aligning videos and texts.

The efficiency of the cluster module. The ablation results are provided in Tab. 6. N_v^- and N_t^- denote the number of visual and textual clusters, respectively. The first row represents the baseline without the cluster module. We find that large numbers of clusters may make similar tokens classified in different clusters. From Tab. 6, we take the $\{N_v^a, N_v^o, N_t^a, N_t^o\}$ as $\{3, 2, 6, 3\}$ to get the best performance on the sum of recall at rank $\{1, 5, 10\}$ (Rsum).

The efficiency of our method. In Tab. 7, we calculate iteration time and inference time using two Tesla V100 GPUs on MSRVTT dataset. Since the Banzhaf Interaction can be removed during inference, our method only takes additional 1s for processing the test set. This result demonstrates the superiority of our efficient design.

Parameter sensitivity. The parameter α is the hyper-parameter that trades off $\mathcal{L}_C$ and $\mathcal{L}_I$. We evaluate the scale range setting $\alpha \in [0.3, 1.7]$ as shown in Fig. 6a. From Fig. 6a, we adopt $\alpha = 1.0$ to achieve the best performance. In Fig. 6b, we show the influence of the hyper-parameter β. We evaluate the scale range setting $\beta \in [0.5, 3.5]$. We find that the model achieves the best performance at $\beta = 2.0$, so we set $\beta = 2.0$ as default in practice.

4.4. Qualitative Analysis

To better understand the proposed method, we show the visualization of the hierarchical interaction in Fig. 5. We find that the semantic similarities between coalitions are generally higher than the semantic similarities between individual frames and individual words. For example, the coalition "{two, men, talking, after, a}" has a high semantic similarity with the video coalition representing the men talking action. On the contrary, when these words interact with the corresponding frame as individuals, they show low semantic similarity. Interestingly, the model uses the word "fire" instead of the phrase "one puts out a fire" to understand the video-text pair. This is due to insufficient training data, the model can not understand the low-frequency phrase. The visualization illustrates that the proposed method can be used as a tool for visualizing the cross-modal interaction and help us understand the cross-modal model.

5. Conclusion

In this paper, 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 (HBI) 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. More encouragingly, our method can also serve as a visualization tool to promote the understanding of cross-modal interaction.

Acknowledgements. This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201, 2022ZD0118101), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), and the Natural Science Foundation of Guangdong Province in China (No. 2019B1515120049).

References

[1] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev 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*, pages 21406–21415, 2022. 3

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2, 7

[3] Yoram Bachrach, Evangelos Markakis, Ezra Resnick, Ariel D Procaccia, Jeffrey S Rosenschein, and Amin Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, 2010. 7

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew 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*, pages 1728–1738, 2021. 7

[5] John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317, 1964. 6

[6] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 7

[7] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 3

[8] Shuqiang Cao, Bairui Wang, Wei Zhang, and Lin Ma. Visual consensus modeling for video-text retrieval. In *AAAI*, 2022. 1

[9] Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011. 2

[10] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021. 3

[11] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 3

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML 2020: 37th International Conference on Machine Learning*, volume 1, pages 1597–1607, 2020. 1

[13] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1

[15] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 7

[16] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016. 3

[17] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3

[18] Jianfeng Dong, Yabing Wang, Xianke Chen, Xiaoye Qu, Xirong Li, Yuan He, and Xun Wang. Reading-strategy inspired visual representation learning for text-to-video retrieval. *arXiv preprint arXiv:2201.09168*, 2022. 2

[19] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016. 5

[20] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

[21] Valentin Gabeur, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Masking modalities for cross-modal video retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1766–1775, 2022. 1

[22] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV 2020 - European Conference on Computer Vision*, volume 12349, pages 214–229, 2020. 7

[23] Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5046–5055, 2022. 3

[24] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 3

[25] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridgeformer: Bridging video-text retrieval with multiple choice questions. *arXiv preprint arXiv:2201.04850*, 2022. 2

[26] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 1

[27] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020. 3

[28] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei 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*, pages 5006–5015, 2022. 7

[29] Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999. 2, 3, 6

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. *arXiv preprint arXiv:2204.02968*, 2022. 1

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020. 1

[32] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *European Conference on Computer Vision*, pages 444–461. Springer, 2022. 1

[33] Peng Jin, JinFa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. Expectation-maximization contrastive learning for compact video-and-language representations. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 7, 8

[34] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017. 2, 7

[35] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997. 5

[36] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 3

[37] Dennis Leech. Computation of power indices. 2002. 7

[38] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing 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*, pages 7331–7341, 2021. 7

[39] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Toward 3d spatial reasoning for human-like text-based visual question answering. *arXiv preprint arXiv:2209.10326*, 2022. 3

[40] Hao Li, Xu Li, Belhal Karimi, Jie Chen, and Mingming Sun. Joint learning of object graph and relation graph for visual question answering. *arXiv preprint arXiv:2205.04188*, 2022. 3

[41] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *arXiv preprint arXiv:2208.02515*, 2022. 3

[42] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022. 1

[43] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. *arXiv preprint arXiv:2103.15049*, 2021. 7

[44] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, page 279, 2019. 7

[45] Yang Liu, Qingchao Chen, and Samuel Albanie. Adaptive cross-modal prototypes for cross-domain visual-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14954–14964, 2021. 3

[46] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. *arXiv preprint arXiv:2207.07852*, 2022. 7, 8

[47] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2, 3, 7, 8

[48] Jean-Luc Marichal and Pierre Mathonet. Weighted banzhaf power and interaction indexes through weighted approximations of games. *European journal of operational research*, 211(2):352–358, 2011. 2, 3

[49] Yasuko Matsui and Tomomi Matsui. Np-completeness for calculating power indices of weighted majority games. *Theoretical Computer Science*, 263(1-2):305–310, 2001. 7

[50] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021. 3

[51] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 7

[52] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994. 2

[53] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 3

[54] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea

Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021. 7

[55] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. *arXiv preprint arXiv:2208.00934*, 2022. 8

[56] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12, 2021. 3

[57] Mengshi Qi, Jie Qin, Yi Yang, Yunhong Wang, and Jiebo Luo. Semantics-aware spatial-temporal binaries for cross-modal video retrieval. *IEEE Transactions on Image Processing*, 30:2989–3004, 2021. 3

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML 2021: 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 7

[59] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021. 5

[60] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20020–20029, 2022. 1

[61] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019. 3

[62] Jianyuan Sun, Hui Yu, Guoqiang Zhong, Junyu Dong, Shu Zhang, and Hongchuan Yu. Random shapley forests: cooperative game-based random forests with consistency. *IEEE transactions on cybernetics*, 2020. 3

[63] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 3

[64] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 7

[65] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*, 2022. 2, 3, 8

[66] Wei Wang, Junyu Gao, Xiaoshan Yang, and Changsheng Xu. Many hands make light work: Transferring knowledge from auxiliary tasks for video-text retrieval. *IEEE Transactions on Multimedia*, 2022. 1

[67] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 7

[68] Jiwei Wei, Yang Yang, Xing Xu, Xiaofeng Zhu, and Heng Tao Shen. Universal weighting metric learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[69] Keyu Wen, Jin Xia, Yuanyuan Huang, Linyang Li, Jiayan Xu, and Jie Shao. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2208–2217, 2021. 3

[70] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3650–3660, 2021. 3

[71] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018. 1

[72] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1, 2, 7

[73] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 2, 7

[74] Li Xu, He Huang, and Jun 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*, pages 9878–9888, 2021. 3

[75] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 3

[76] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 8

[77] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 3

[78] Yu Yang, Seungbae Kim, and Jungseock Joo. Explaining deep convolutional neural networks via latent visual-semantic filter attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8333–8343, 2022. 3

[79] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In *Proceedings of the 35th International Conference on*

Neural Information Processing Systems, volume 34, pages 26462–26474, 2021. 8

[80] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1

[81] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 7

[82] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, volume 34, pages 23634–23651, 2021. 3, 8

[83] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 5

[84] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–401, 2018. 7

[85] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10877–10886, 2021. 3

[86] Shaofeng Zhang, Meng Liu, Junchi Yan, Hengrui Zhang, Lingxiao Huang, Xiaokang Yang, and Pinyan Lu. M-mix: Generating hard negatives via multi-sample mixing for contrastive learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2461–2470, 2022. 1

[87] Shaofeng Zhang, Lyn Qiu, Feng Zhu, Junchi Yan, Hengrui Zhang, Rui Zhao, Hongyang Li, and Xiaokang Yang. Align representations with base: A new approach to self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16600–16609, 2022. 1

[88] Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2021. 1

[89] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[90] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8746–8755, 2020. 3