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. Project page is available at https://jpthu17.github.io/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 [20, 32] and video-question answering [28, 49]. Visual-language learning has recently boomed due to the success of contrastive learning [9–11, 19, 48, 61–64], e.g., CLIP [40], 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 [14, 20, 32] typically models the cross-modal interaction via solely the global similarity of each modality. Specifically, as shown in 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 [20, 32, 45], 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 [18] for its simplicity and efficiency. Banzhaf Interaction is one of the most popular concepts in cooperative games [33]. 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, ActivityNet Captions and DiDeMo) and the video question answering benchmark dataset (MSRVTT-QA) show the advantages of the proposed method. The main contributions are as follows:

- 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 [6, 35]. 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 [13, 54, 60], but there is little exploration in cross-modal learning. Banzhaf Interaction is one of the most popular concepts in cooperative games [33]. Recently, LOUPE [27] 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 [20, 25, 26], such as text-video retrieval [8, 32] and video-question answering [20, 36]. Text-video retrieval [14, 47, 51] requires the model to map text and video to the same latent space, where the similarity between them can be directly calculated [7, 39, 53]. Video-question answering requires the model to predict an answer using visual information [25, 26, 57]. 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 words 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 \((v, t)\), 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.

To learn fine-grained semantic alignment, the input video \(v\) is embedded into frame sequence \(V_f = \{v_f^{i}\}_{i=1}^{N_v}\), where \(N_v\) is the length of video \(v\). The input text \(t\) is embedded into word sequence \(T_w = \{t_w^{j}\}_{j=1}^{N_t}\), where \(N_t\) is the length of text \(t\). Then, the alignment matrix is defined as:

\[
A = [a_{ij}]_{N_v \times N_t}, \quad a_{ij} = \frac{(v_f^i)^T t_w^j}{\|v_f^i\| \|t_w^j\|}
\]

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 \(\max_j a_{ij}\). 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 [45] can be defined as:

\[
\begin{align*}
S_{v, t} &= \frac{1}{2} \left( \sum_{i=1}^{N_v} \omega^v_i \max_j a_{ij} + \sum_{j=1}^{N_t} \omega^t_j \max_i a_{ij} \right), \quad \text{(1)}
\end{align*}
\]

where \([\omega^v_1, \omega^v_2, ..., \omega^v_{N_v}] = \text{Softmax} (\text{MLP}_v(V_f))\) and \([\omega^t_1, \omega^t_2, ..., \omega^t_{N_t}] = \text{Softmax} (\text{MLP}_t(T_w))\) are the weights of the video frames and text words, respectively. Then the cross-modal contrastive loss [43] can be formulated as:

\[
\begin{align*}
\mathcal{L}_C &= -\frac{1}{2B} \sum_{k=1}^B \log \frac{\exp(S_{v_k, t_k}/\tau)}{\sum_{l \neq k} \exp(S_{v_l, t_k}/\tau)} + \\
&= \frac{1}{2B} \sum_{k=1}^B \log \frac{\exp(S_{v_k, t_k}/\tau)}{\sum_{l \neq k} \exp(S_{v_l, t_k}/\tau)}, \quad \text{(2)}
\end{align*}
\]

where \(B\) is the batch size and \(\tau\) is the temperature hyperparameter. 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 \(S_{v, t}\). 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.

3.1.2 Banzhaf Interaction

We start by introducing notation and outlining assumptions about the cooperative game theory. Then, we review Banzhaf Interaction [18] for a cooperative game.

The cooperative game theory consists of a set \(\mathcal{N} = \{1, 2, ..., n\}\) of players with a characteristic function \(\phi\). The characteristic function \(\phi\) 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, and at the same time \(\phi(\{i, j\})\) is large. The Banzhaf Interaction [18] 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 [18].** Given a coalition \(\{i, j\} \subseteq \mathcal{N}\), the Banzhaf Interaction \(I(\{i, j\})\) for the
player \{i, j\} is defined as:

\[ I(\{i, j\}) = \sum_{c \subseteq N \setminus \{i, j\}} p(C) (\phi(C \cup \{i, j\}) - \phi(C \cup \{i\}) - \phi(C \cup \{j\})) \]

where \( p(C) = \frac{1}{2^{|N|}} \) is the likelihood of \( C \) being sampled.

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

### 3.1.3 Video-Text as Game Players

Given features \( V_f = \{v_f^i\}_{i=1}^N \) and \( T_w = \{t_w^j\}_{j=1}^N \), 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 \( N = \{v_f^i\}_{i=1}^N \cup \{t_w^j\}_{j=1}^N \) 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 \( \phi \) should meet all the following criteria: (a) the final score benefits from strongly corresponding semantic pairs \( \{v_f^i, t_w^j\} \), i.e., \( \phi(N) - \phi(N \setminus \{v_f^i, t_w^j\}) < 0 \); (b) the final score is compromised by semantically irrelevant pairs \( \{v_f^i, t_w^j\}_1 \), i.e., \( \phi(N) - \phi(N \setminus \{v_f^i, t_w^j\}) < 0 \); (c) when there are no players to cooperate, the final score is zero, i.e., \( \phi(N) = \phi(\{v_f^i\}_{i=1}^N) = \phi(\{t_w^j\}_{j=1}^N) = \phi(\emptyset) = 0 \), where \( \emptyset \) denotes the empty set.

Note that anything satisfying the above conditions can be used as the characteristic function \( \phi \). For simplicity, we use cross-modality similarity measurement \( S \) as \( \phi \). 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 text word. 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 \( \{v_f^i, t_w^j\}_1 \), referring to Eq. 3, we can cal-
calculate the Banzhaf Interaction $I([v_f^j, t_v^j])$. Due to the disparity in semantic similarity and interaction index, we design a prediction header to predict the fine-grained relationship $R_{i,j}$ 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 [22] between the $I([v_f^j, t_v^j])$ and $R_{i,j}$. Concretely, we define the probability distribution of the video-to-text task and the text-to-video task as:

$$
D^v_{v2t} = [p^v_{1,1}, p^v_{1,2}, \ldots, p^v_{N_v, t}],
$$
$$
D^t_{t2v} = [p^t_{1,j}, p^t_{2,j}, \ldots, p^t_{N_t, j}],
$$

where $p^v_{i,j} = \frac{\exp(I([v_f^j, t_v^j]))}{\sum_{i=1}^{N_v} \exp(I([v_f^j, t_v^j]))}$ and $p^t_{i,j} = \frac{\exp(I([v_f^j, t_v^j]))}{\sum_{i=1}^{N_t} \exp(I([v_f^j, t_v^j]))}$.

Similarly, the probability distribution $D^R_{v2t}$ and $D^R_{t2v}$ are calculated in the same way using $R^{i,j}$, i.e.,

$$
D^R_{v2t} = [p^R_{1,1}, p^R_{1,2}, \ldots, p^R_{N_v, t}],
$$
$$
D^R_{t2v} = [p^R_{1,j}, p^R_{2,j}, \ldots, p^R_{N_t, j}],
$$

where $p^R_{i,j} = \frac{\exp(R^{i,j})}{\sum_{i=1}^{N_i} \exp(R^{i,j})}$. Finally, the Banzhaf Interaction loss $L_I$ is defined as:

$$
L_I = E_{v,t} [KL(D^R_{v2t} \parallel D^v_{v2t}) + KL(D^R_{t2v} \parallel D^t_{t2v})] + KL(D^R_{v2t} \parallel D^v_{v2t}) + KL(D^R_{t2v} \parallel D^t_{t2v})].
$$

The Banzhaf Interaction loss $L_I$ brings the probability distributions of the output $R$ of the prediction header and Banzhaf Interaction $I$ 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 [15], a k-nearest neighbor-based density peaks clustering algorithm, to cluster the visual (textual) tokens. Starting with the frame-level tokens $V_f = \{v_f^j\}_{j=1}^{N_f}$, we first use a one-dimensional convolutional layer to enhance the temporal information between tokens. Then, we compute the local density $\rho_i$ of each token $v_f^j$ according to its $K$-nearest neighbors:

$$
\rho_i = \exp(-\frac{1}{K} \sum_{v_f^j \in \text{KNN}(v_f^j)} \|v_f^j - v_f^i\|^2),
$$

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

$$
\delta_i = \begin{cases}
\min_{j \in \text{KNN}(v_f^j)} \|v_f^j - v_f^i\|^2, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i, \\
\max_{j \in \text{KNN}(v_f^j)} \|v_f^j - v_f^i\|^2, & \text{otherwise}.
\end{cases}
$$

Intuitively, $\rho$ denotes the local density of tokens, and $\delta$ 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 [41,58], we use the weighted average tokens of each cluster to represent the corresponding cluster, where the weight $W = \text{Softmax}(\text{MLP}(V_f))$. 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 $L_C$ and Banzhaf Interaction loss $L_I$, the full objective of semantic alignment can be formulated as $L = L_C + \alpha L_I$, where $\alpha$ is the trade-off hyper-parameter. We train the network at three semantic levels, which are shown as follows,

$$
L^c = L_C + \alpha L_I, \quad L^o = L_C + \alpha L_I, \quad L^e = L_C + \alpha L_I,
$$

where $L^c$, $L^o$, and $L^e$ 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 $S^e_{v,t}$ 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 $S^e_{v,t}$ distilling to action-level similarity $S^a_{v,t}$, we first compute the distribution $D^{v,t}_{v2t}$ and $D^{t,v}_{t2v}$ by replacing $I([v,t])$ with $S^e_{v,t}$ in...
Eq. 4. The distribution $D^{e_{2t}}_D$ and $D^{e_{2t}}_{e_{2t}}$ are calculated using $S_{v,t}$. The $L^{c2a}_D$ loss is defined as:

$$L^{c2a}_D = E_{v,t}[KL(D^{e_{2t}}_D||D^{e_{2t}}_{e_{2t}}) + KL(D^{e_{2t}}_{e_{2t}}||D^{e_{2t}}_D)].$$

(9)

The $L^{c2o}_D$ 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:

$$L_{total} = L^c + L^o + \beta (L^{c2a}_D + L^{c2o}_D),$$

(10)

where $\beta$ 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 $L_I$ significantly improves the performance, while deep supervision and self-distillation can improve the generalization ability.

3.4. Theoretical Analysis

Similar to Banzhaf value axioms [18], the following axioms convey intuitive properties that a cross-modal interaction score should satisfy.

**Axioms 1.** Given a set $N = \{1, 2, ..., n\}$ of players, a characteristic function $\phi : 2^n \rightarrow \mathbb{R}$, and a coalition $C \subseteq N$, following properties are met for the interaction score $I(C)$.

(a) **Symmetry:** If $\forall S \subseteq N, \phi(S \cup \{i\}) = \phi(S \cup \{j\})$, then $I([i]) = I([j])$.
(b) **Dummy:** If $\forall S \subseteq N, \phi(S \cup \{i\}) = \phi(S), \forall i \in C, \phi(S \cup \{i\}) = \phi(S), \forall i \in C \setminus \{j\}, \phi(S \cup \{i\}) = \phi(S), \forall i \in C \setminus \{j\}, \phi(S \cup \{i\}) = \phi(S)$.
(c) **Additivity:** If $\phi(\cdot)$ and $\phi(\cdot)$ have the interaction scores $I([C])$ and $I([C \setminus \{i\}] + I([C \setminus \{i\}]$ respectively, then the interaction score for the game with value function $\phi(\cdot) + \phi(\cdot)$ is $I([C]) + I([C \setminus \{i\}]$.
(d) **Recursivity:** Let $B(\cdot)$ denote the Banzhaf value [4], then $B([C]) = B([C \setminus \{i\}] + B([C \setminus \{j\}]) + B([C \setminus \{i\}] + B([C \setminus \{j\}])$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Receptive field</th>
<th>Robustness</th>
<th>Flexibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine similarity</td>
<td>Element level</td>
<td>Absolute value</td>
<td>Non-adjustable</td>
</tr>
<tr>
<td>Banzhaf Interaction</td>
<td>Set level</td>
<td>Relative value</td>
<td>Adaptable</td>
</tr>
</tbody>
</table>

**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
4. Experiments

4.1. Experimental Settings

Datasets. MSRVTT [50] contains 10K YouTube videos, each with 20 text descriptions. We follow the training protocol in [16, 30] and evaluate on the 1K-A testing split [56]. ActivityNet Captions [21] 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 [1] contains 10K videos annotated 40K text descriptions. We follow the training and evaluation protocol in [32], MSRVTT-QA [49] 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) [16] 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 [34], existing methods mainly use sampling-based methods [2, 23] 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) [40] 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 $R$. To explore the impact of the structure of the prediction header on our method, we compare several popular structures in Tab. 4. We consistently outperform the recently proposed state-of-the-art methods in 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) [40] 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.

Ablation about components. As shown in Tab. 5, Banzhaf Interaction boosts the baseline with the improve-
We evaluate the scale range setting \( \beta \) we set. In Fig. 6b, we show the influence of the hyper-parameter \( \beta \). 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 cannot 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), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), and the Natural Science Foundation of Guangdong Province in China (No. 2019B1515120049).
References


[28] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense


