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

Peng Jin Jinfa Huang Pengfei Xiong Shangxuan Tian Chang Liu
Xiangyang Ji Li Yuan* Jie Chen*

Contents

A. Datasets and Implementation Details

A.1. Datasets

MSRVTT. MSRVTT [23] contains 10K YouTube videos, each with 20 text descriptions. We follow the training protocol in [7, 12] and evaluate on text-to-video and video-to-text search tasks on the 1K-A testing split with 1K video or text candidates defined by [24].

ActivityNet Captions. ActivityNet Captions [9] consists densely annotated temporal segments of 20K YouTube videos. Following [7, 16, 21], 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. DiDeMo [1] contains 10K videos annotated 40K text descriptions. We concatenate descriptions of segments in a video to construct "video-paragraph" for retrieval. We follow the training and evaluation protocol in [14].

MSRVTT-QA. MSRVTT-QA [22] is based on the MSRVTT dataset and has 243K VideoQA pairs.

A.2. Implementation Details

For fair comparisons, we follow common practice [4, 14, 20] to extract the video representations of input videos and the language representations of input texts. In detail, for video representations, we first extract the frames from the video clip as the input sequence of video. Then we use ViT [5] to encode the frame sequence, by exploiting the transformer [11, 19] architecture to model the interactions between image patches. Followed by the CLIP [17], the output from the [class] token is used as the frame embedding. Finally, we obtain the video representation $\boldsymbol{V}_f = \{v_f^i\}_{i=1}^{N_v}$. For text representation, we directly use the text encoder of CLIP to acquire the text representation $\boldsymbol{T}_w = \{t_w^j\}_{j=1}^{N_t}$.

The dimension of the feature is 512. The temporal transformer is composed of 4-layer blocks, each including 8 heads and 512 hidden channels. The temporal position embedding and parameters are initialized from the CLIP's text encoder. We use the Adam optimizer and set the temperature τ to 0.01. The initial learning rate is 1e-7 for text encoder and video encoder and 1e-3 for other modules.

For text-video retrieval, we utilize the CLIP (ViT-B/32) [17] as the pre-trained model. The frame length and caption length are 12 and 24 for MSRVTT. The network is optimized with the batch size of 128 in 5 epochs. We set the caption length to 64 for ActivityNet Captions and DiDeMo.

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 frame length and question length are 12 and 32 for MSRVTT-QA. The network is optimized with the batch size of 32 in 5 epochs.

B. Proof of Theorem 1

We start by reviewing Banzhaf Values [3] and Banzhaf Interaction [8] for a cooperative game.

The cooperative game theory consists of a set $\mathcal{N} = \{1, 2, ..., n\}$ of players with a characteristic function $\phi : 2^n \rightarrow \mathbb{R}$. 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 [18].

Banzhaf Values. The Banzhaf value [3] is one of the most important solution concepts in cooperative games. Formally, the Banzhaf value measures the average marginal contribution of each player across all permutations of the players. It is the unbiased estimation of the importance or contribution of each player in a cooperative game [6], and has thus found many applications from estimating feature importance to pruning neural networks. Given a set $\mathcal{N} = \{1, 2, ..., n\}$ of players and a characteristic function $\phi : 2^n \rightarrow \mathbb{R}$, the Banzhaf value $\mathcal{B}(i|\mathcal{N})$ for player i is defined as the average marginal contribution of player i to all possible coalitions $\mathcal{C} \subseteq \mathcal{N}$ that are formed without i:

$$\mathcal{B}(i|\mathcal{N}) = \sum_{\mathcal{C} \subseteq \mathcal{N} \setminus \{i\}} p(\mathcal{C})(\phi(\mathcal{C} \cup \{i\}) - \phi(\mathcal{C})), \tag{A}$$

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

Banzhaf Interaction. 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 [8] measures the additional benefits brought by the coalition compared with the costs of the lost interactions of these players with others. 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 [8].*** *Given a coalition $\{i, j\} \subseteq \mathcal{N}$, the Banzhaf Interaction $\mathcal{I}([\{i, j\}])$ for the player $[\{i, j\}]$ is defined as:*

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

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 $\mathcal{N}$.

Similar to Banzhaf value axioms [8], 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 $\phi : 2^n \rightarrow \mathbb{R}$, 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, then $\mathcal{B}([\mathcal{C}]|\mathcal{N} \setminus \mathcal{C} \cup \{[\mathcal{C}]\}) = \mathcal{B}(i|\mathcal{N} \setminus \{j\}) + \mathcal{B}(j|\mathcal{N} \setminus \{i\}) + \mathcal{I}([\mathcal{C}])$.*

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

B.1. Symmetry Axiom

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.

Proof. We consider $\mathcal{C} = \{i, j\}, \mathcal{C}^{'} = \{i^{'}, j^{'}\}$ fixed. Let us choose $\mathcal{T} \subseteq \mathcal{N}$, and consider the unanimity game. Clearly, $\phi(\mathcal{T} \cup \{[\{i, j\}\}\}) - \phi(\mathcal{T} \cup \{[\{i^{'}, j^{'}\}]\}) = 0, \phi(\mathcal{T} \cup i) - \phi(\mathcal{T} \cup i^{'}) = 0, \phi(\mathcal{T} \cup j) - \phi(\mathcal{T} \cup j^{'}) = 0$. That is, for every $\mathcal{T} \subseteq \mathcal{N}$, $\mathcal{C} = \{i, j\}$ and $\mathcal{C}^{'} = \{i^{'}, j^{'}\}$ produce the same benefits. Thus, Banzhaf Interaction satisfies Symmetry axiom, *i.e.*, $\mathcal{I}([\mathcal{C}]) = \mathcal{I}([\mathcal{C}^{'}])$.

B.2. Dummy Axiom

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.

Proof. We consider $\mathcal{C} = \{i, j\}$ fixed. Let us choose $\mathcal{T} \subseteq \mathcal{N}$, and consider the unanimity game. Clearly, $\phi(\mathcal{S} \cup \{[\mathcal{C}]\}) - \phi(\mathcal{S}) = 0, \sum_{i \in \mathcal{C}} \phi(\mathcal{S} \cup i) = 0$. For every $\mathcal{T} \subseteq \mathcal{N}$, $\mathcal{C} = \{i, j\}$ has no interaction with any player. Thus, Banzhaf Interaction satisfies Dummy axiom, *i.e.*, $\mathcal{I}([\mathcal{C}]) = 0$.

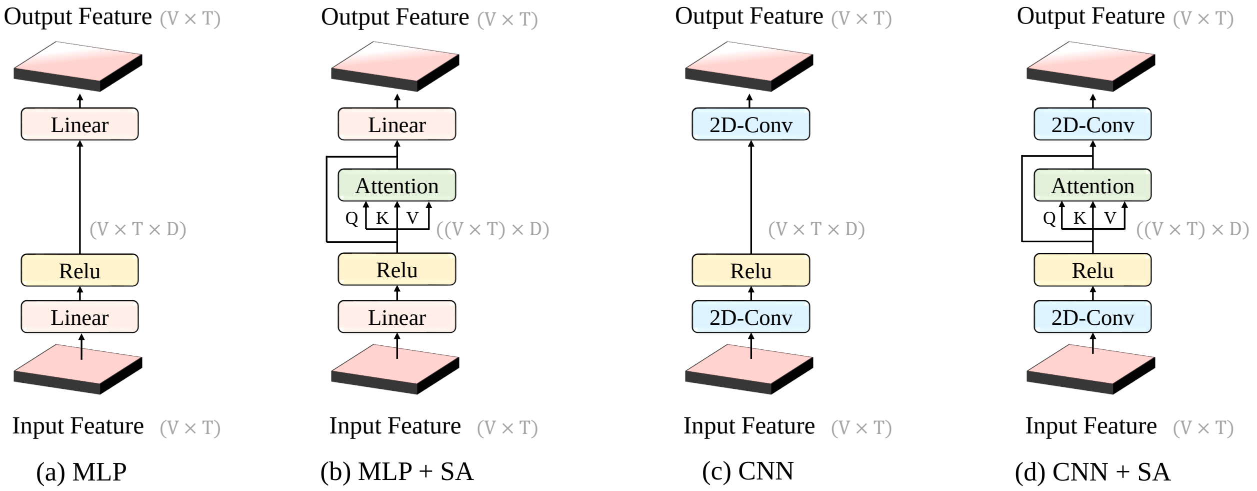
B.3. Additivity Axiom

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.

Proof. Let us choose $\mathcal{T} \subseteq \mathcal{N}$, and consider the unanimity game. Clearly, for the characteristic function $\Phi(*) = \phi(*) + \phi^{'}(*)$, $\Phi(\mathcal{T}) = \phi(\mathcal{T}) + \phi^{'}(\mathcal{T})$. That is, for every $\mathcal{T} \subseteq \mathcal{N}$, the sum of the scores of the two characteristic functions $(\phi(*), \phi^{'}(*))$ is equal to the score of the sum of these characteristic functions $\Phi(*)$. Thus, Banzhaf Interaction satisfies Additivity axiom.

B.4. Recursivity Axiom

We hypothesize that the interaction score should depend on the values of i when j is absent, and j when i is absent.



(a) MLP (b) MLP + SA (c) CNN (d) CNN + SA

Figure A. **The structure of the prediction header.** We choose four several popular structures, *i.e*., "MLP", "CNN", "MLP+SA" and "CNN+SA". V, T, D represent the number of visual tokens, the number of textual tokens, and the number of feature channels, respectively.

And somehow, their interaction should also be taken into account. Specifically, 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.

Proof. We can rewrite Eq. B as $\mathcal{I}([\mathcal{C}]) = \mathcal{B}([\mathcal{C}]|\mathcal{N} \setminus \mathcal{C} \cup \{[\mathcal{C}]\}) - \sum_{i \in \mathcal{C}} \mathcal{B}(i|\mathcal{N} \setminus \mathcal{C} \cup \{i\})$. Clearly, the above formula is equivalent to Recursivity axiom. Thus, Banzhaf Interaction satisfies Recursivity axiom.

C. Discussions

C.1. Banzhaf Interaction Estimator

Since the calculation of the exact Banzhaf Interaction is an NP-hard problem [15], existing methods mainly use sampling-based methods [2, 10] 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 a convolutional layer for encoding features, a self-attention module for capturing global interaction, and a convolutional layer for decoding. The tiny model has 64 hidden channels. The input is the similarity matrix of video frames and text tokens, and the output is the estimation of Banzhaf Interaction.

To explore the impact of the Banzhaf Interaction estimator on our method, we compare the sampling-based method and pre-trained tiny model estimator in Tab. A. Given the costly training time, the ablation study is based on a subset of MSRVTT dataset (3K videos, each with 20 text descriptions). We find that the pre-trained tiny model maintains the estimation accuracy while avoiding intensive computations. The average training time is reduced from 19.79 seconds per iteration to 3.14 seconds per iteration.

Method	Text->Video				Iteration
	R@1↑	R@5↑	R@10↑	MnR↓	Time↓
Baseline	40.0	66.8	77.0	16.5	2.06 s
w/ Sampling-based method	41.5	**68.6**	78.9	**15.1**	19.79 s
w/ Tiny estimator	**41.8**	67.5	**79.0**	15.2	**3.14** s

Table A. **Effect of the Banzhaf Interaction Estimator.** "↑" denotes that higher is better. "↓" denotes that lower is better.

C.2. The Structure of the Prediction Header

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". Fig. A illustrates the structures.

"MLP" consists of a linear layer with a Relu activation function for encoding features and a linear layer for decoding. The dimension of the hidden channels is 64. **"CNN"** consists of a convolutional layer with a Relu activation function for encoding features and a convolutional layer for decoding. The dimension of the hidden channels is 64. **"MLP+SA"** consists of a linear layer with a Relu activation function for encoding features, a self-attention module for capturing global interaction, and a linear layer for decoding. The dimension of the hidden channels is 64. **"CNN+SA"** consists of a convolutional layer with a Relu activation function for encoding features, a self-attention module for capturing global interaction, and a convolutional layer for decoding. The dimension of the hidden channels is 64.

Method	Text->Video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MLP	47.2	73.7	83.5	**2.0**	12.3
CNN	47.3	73.5	**83.7**	**2.0**	12.2
MLP+SA	46.6	74.0	**83.7**	**2.0**	12.3
CNN+SA	**48.6**	**74.6**	83.4	**2.0**	**12.0**

Table B. **Effect of the structure of the prediction header on MSRVTT dataset.** "SA" is the self-attention module. "↑" denotes that higher is better. "↓" denotes that lower is better.

Method	Text->Video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
E->A	48.1	73.6	82.9	**2.0**	11.9
E->O	48.0	74.1	83.2	**2.0**	**11.8**
A->O	48.0	73.0	83.1	**2.0**	12.0
E->A + A->O	48.2	74.1	82.9	**2.0**	**11.8**
E->A + E->O	**48.6**	**74.6**	**83.4**	**2.0**	12.0

Table C. **Ablation study about the self-distillation of our method on MSRVTT dataset.** 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. "↑" denotes that higher is better. "↓" denotes that lower is better.

$\mathcal{L}_I$ Banzhaf Interaction	Deep Supervision	$\mathcal{L}_D$ Self Distillation	Top1 Acc ↑	Top5 Acc ↑
			45.2	73.1
✓			45.8	73.7
	✓		46.0	74.0
	✓	✓	46.0	74.1
✓	✓		46.1	**74.2**
✓	✓	✓	**46.2**	**74.2**

Table D. **Ablation study about the importance of each part on MSRVTT-QA dataset.** "↑" denotes that higher is better.

Method	Text->Video				
	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
Baseline	46.6	73.1	83.0	**2.0**	13.3
One level	47.5	73.7	83.0	**2.0**	12.0
Two levels	48.1	73.6	82.9	**2.0**	**11.9**
Three levels	**48.6**	**74.6**	**83.4**	**2.0**	12.0

Table E. **Effect of the number of semantic levels (the number of stacked token merge modules) on MSRVTT dataset.** "↑" denotes that higher is better. "↓" denotes that lower is better.

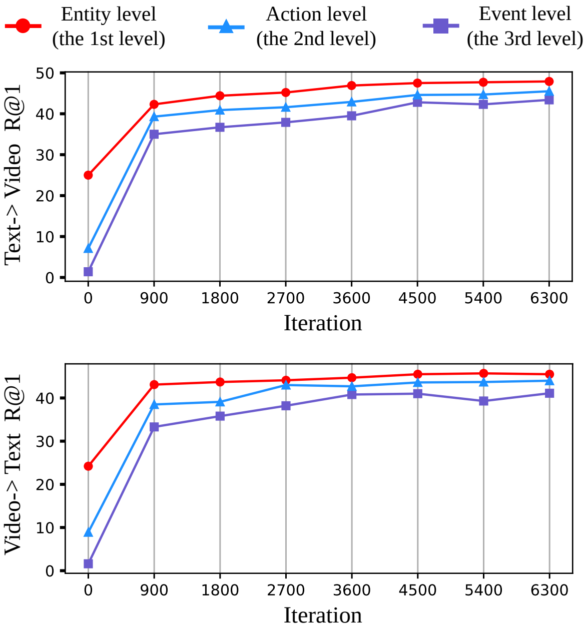


Figure B. **Performance at each semantic level of text-to-video retrieval and video-to-text retrieval task.**

As shown in Tab. B, 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 between video and text. As a result, we adopt "CNN+SA" to achieve the best performance.

C.3. Self-Distillation

Fig. B shows the performance of each semantic level. We find that the entity level converges first in the training process. This is because 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. Thus, we distill the entity-level similarity to the other two semantic levels.

To illustrate the impact of the self-distillation of our method, we conduct ablation experiments on MSRVTT dataset in Tab. C. As we can see, self-distillation improves the generalization ability. Distilling from the entity level to the other two semantic levels achieves the best results. As a result, we distill the entity-level similarity to the other two semantic levels as default in practice.

C.4. Ablation for Video-Question Answering Task

To illustrate the importance of each part of our method for the video-question answering, we conduct ablation experiments on MSRVTT-QA dataset in Tab. D. As we can see, Banzhaf Interaction boosts the baseline with the improvement up to 0.6% at Top1 accuracy. Moreover, deep supervision and self-distillation significantly improve the generalization ability. Self-distillation provides limited improvement for video-question answering compared to text-video retrieval. This is because reasoning relies primarily on high-level semantic features. Therefore, it is difficult

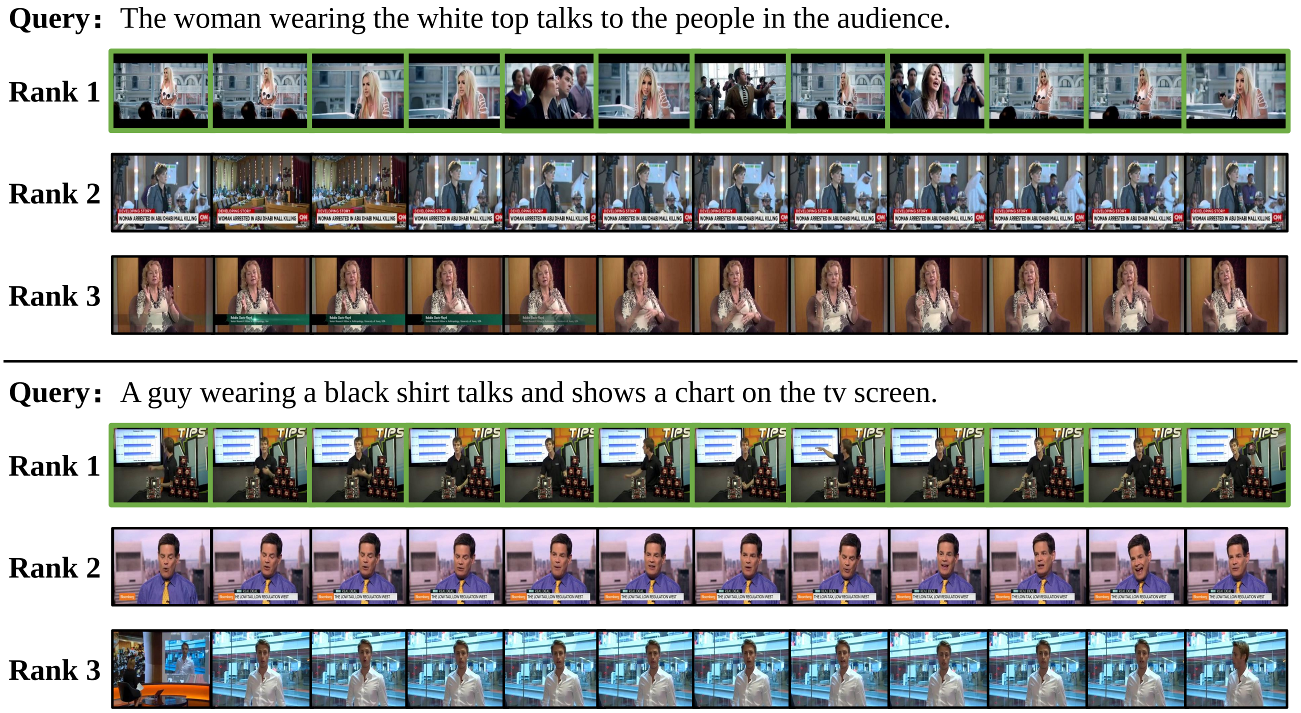

Query: The woman wearing the white top talks to the people in the audience.
Rank 1
Rank 2
Rank 3
Query: A guy wearing a black shirt talks and shows a chart on the tv screen.
Rank 1
Rank 2
Rank 3

Figure C. **Visualization of the text-to-video retrieval.** Only the correct videos are highlighted in green.

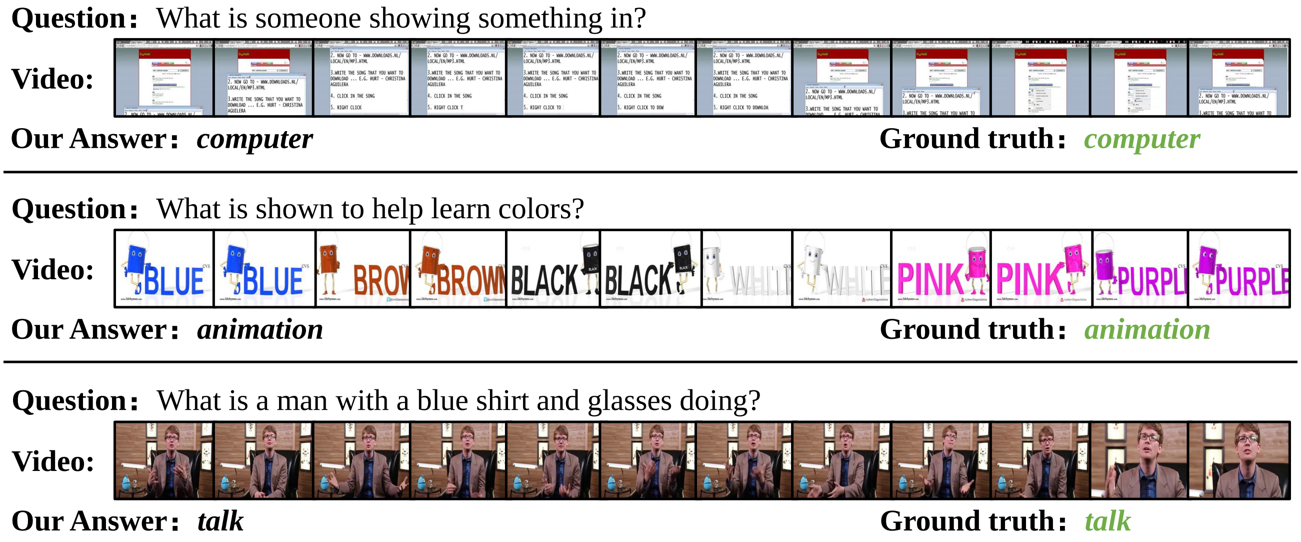

Question: What is someone showing something in?
Video:
Our Answer: computer
Ground truth: computer
Question: What is shown to help learn colors?
Video:
Our Answer: animation
Ground truth: animation
Question: What is a man with a blue shirt and glasses doing?
Video:
Our Answer: talk
Ground truth: talk

Figure D. **Visualization of the video-question answering.**

for low-level semantic features to guide high-level semantic features. Our full model achieves the best performance and outperforms the baseline by 1.0% at Top1 accuracy.

C.5. The Number of Semantic Levels

To efficiently generate coalitions among game players, 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.

To explore the impact of the number of semantic levels on our method, we conduct ablation experiments on MSRVTT dataset in Tab. E. We find that the performance of the model increases with the number of semantic levels. These results

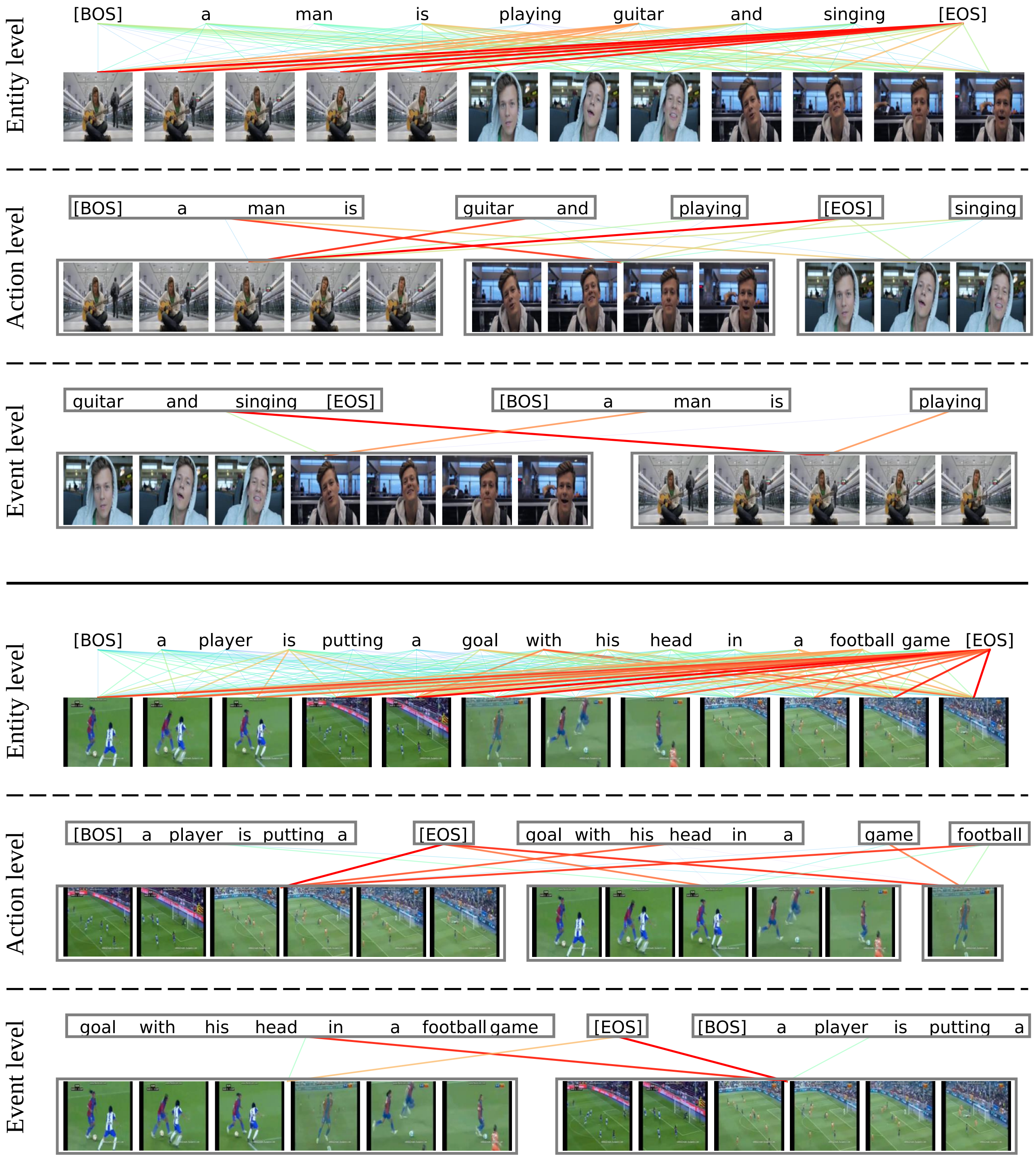

Entity level
[BOS] a man is playing guitar and singing [EOS]
Action level
[BOS] a man is
guitar and
playing
[EOS]
singing
Event level
guitar and singing [EOS]
[BOS] a man is
playing
Entity level
[BOS] a player is putting a goal with his head in a football game [EOS]
Action level
[BOS] a player is putting a
[EOS]
goal with his head in a
game
football
Event level
goal with his head in a football game
[EOS]
[BOS] a player is putting a

Figure E. **Visualization of the hierarchical interaction.** Here, the degree of confidence from high to low is represented by red, orange, green and blue lines, respectively. Entity-level interactions demonstrate the semantic correlation between frames and words. Action-level interactions indicate the semantic correlation between clips and phrases. Event-level interactions show the semantic correlation between segments and paragraphs.

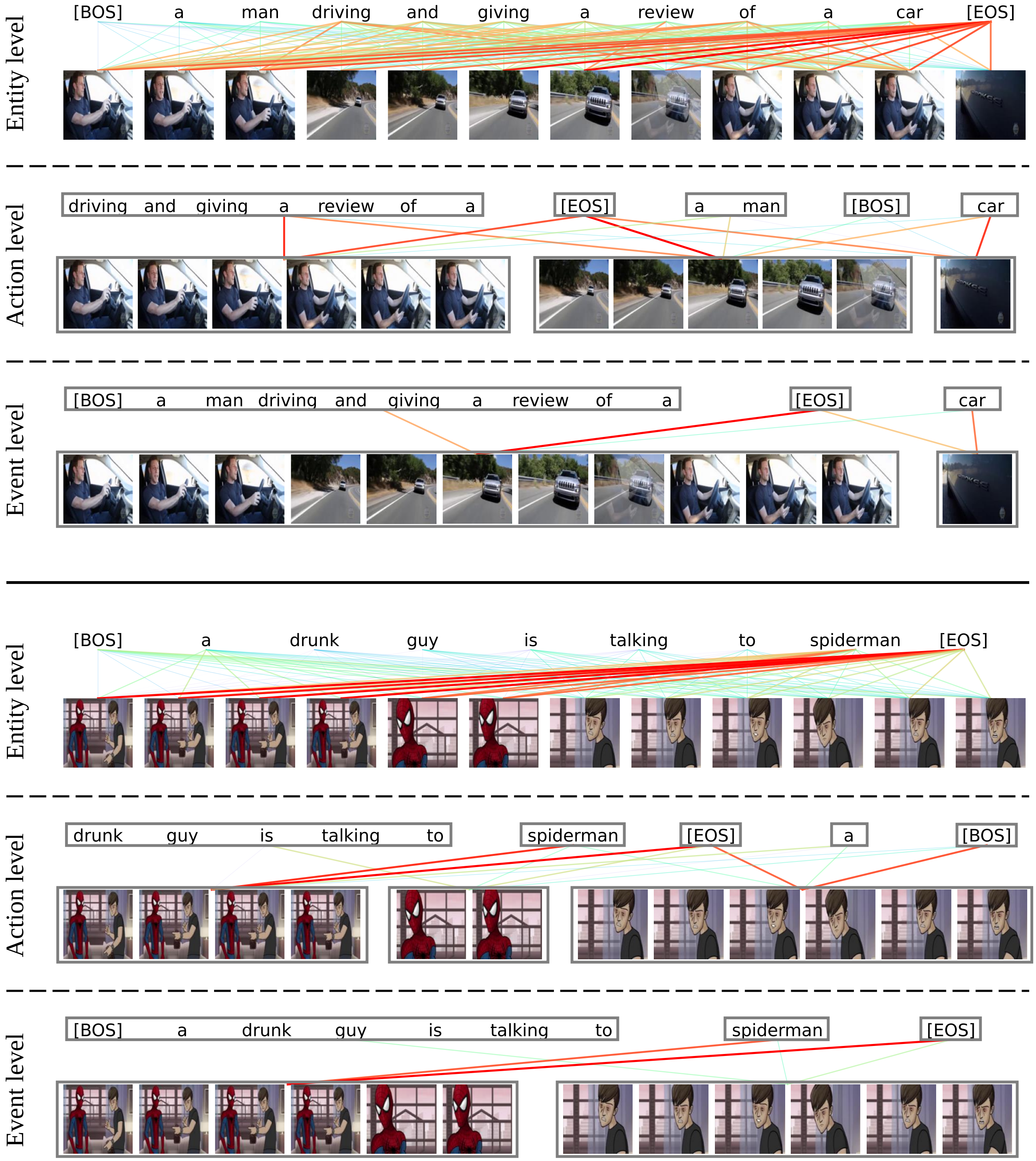

Entity level
[BOS] a man driving and giving a review of a car [EOS]
Action level
driving and giving a review of a
[EOS]
a man
[BOS]
car
Event level
[BOS] a man driving and giving a review of a
[EOS]
car
Entity level
[BOS] a drunk guy is talking to spiderman [EOS]
Action level
drunk guy is talking to
spiderman
[EOS]
a
[BOS]
Event level
[BOS] a drunk guy is talking to
spiderman
[EOS]

Figure F. **Visualization of the hierarchical interaction.** Here, the degree of confidence from high to low is represented by red, orange, green and blue lines, respectively. Entity-level interactions demonstrate the semantic correlation between frames and words. Action-level interactions indicate the semantic correlation between clips and phrases. Event-level interactions show the semantic correlation between segments and paragraphs.

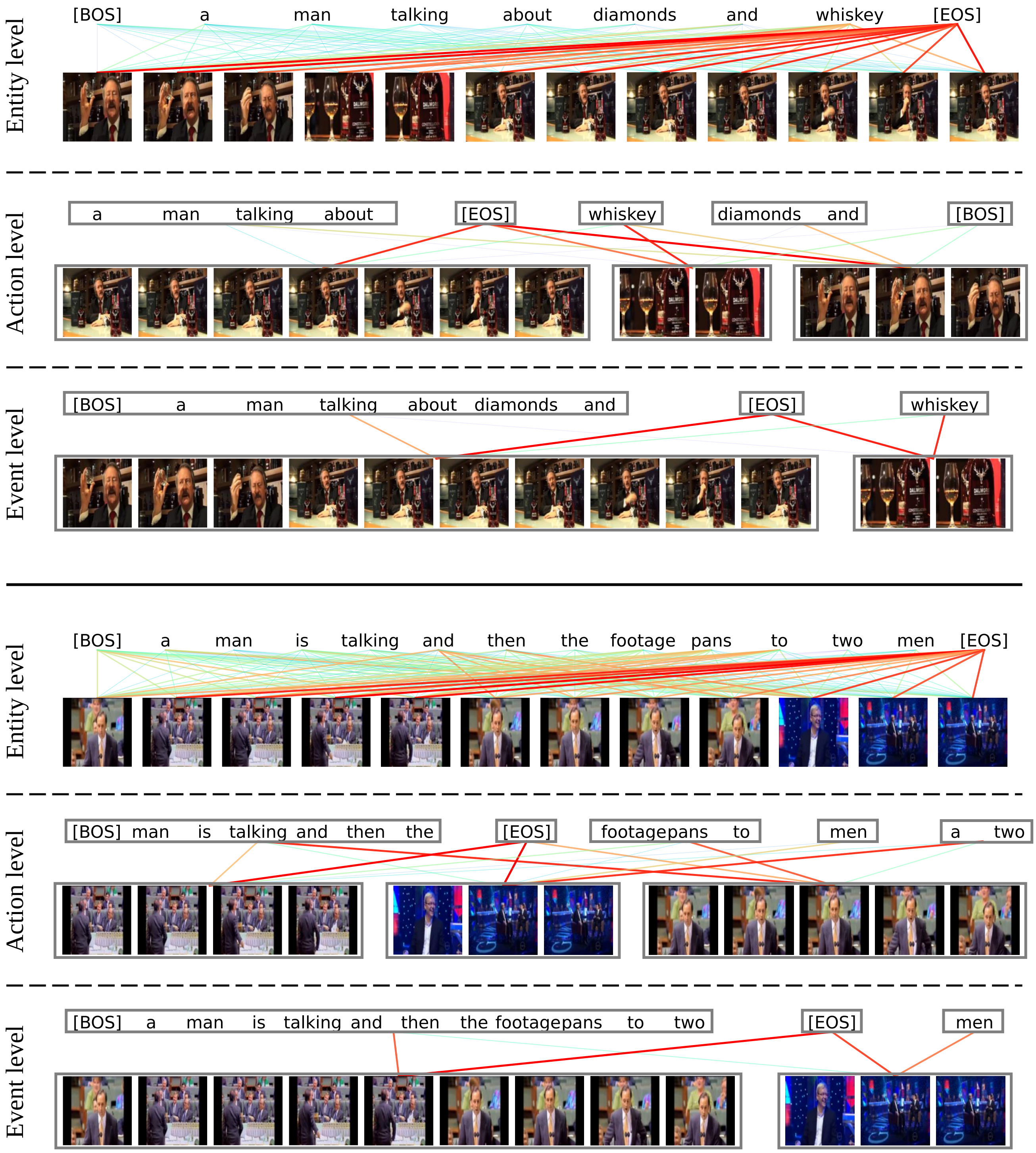

Entity level
[BOS] a man talking about diamonds and whiskey [EOS]
Action level
a man talking about
[EOS]
whiskey
diamonds and
[BOS]
Event level
[BOS] a man talking about diamonds and
[EOS]
whiskey
Entity level
[BOS] a man is talking and then the footage pans to two men [EOS]
Action level
[BOS] man is talking and then the
[EOS]
footagepans to
men
a two
Event level
[BOS] a man is talking and then the footagepans to two
[EOS]
men

Figure G. **Visualization of the hierarchical interaction.** Here, the degree of confidence from high to low is represented by red, orange, green and blue lines, respectively. Entity-level interactions demonstrate the semantic correlation between frames and words. Action-level interactions indicate the semantic correlation between clips and phrases. Event-level interactions show the semantic correlation between segments and paragraphs.

indicate that stacking more token merge modules can provide more coalitions, which enables the model to learn more diverse semantic interaction information. We make a trade-off between the number of semantic levels and computation cost and set the number of semantic levels to 3 in practice.

C.6. Limitations of our Work

The cross-modal contrastive approach typically exploits the coarse-grained labels of video-text pairs to learn a global semantic interaction. 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. Therefore, our method inevitably requires more training time costs. Although our method takes less time than TS2-Net [13] during the inference stage (see Tab. 5 in the main paper), more effort could be paid to obtain an efficient structure in the future.

D. Visualizations

D.1. Text-to-Video Retrieval

We show two retrieval examples from the MSR-VTT testing set for text-to-video retrieval in Fig. C. As shown in Fig. C, our method successfully retrieves the ground-truth video. These results demonstrate that our method can align video and text effectively.

D.2. Video-Question Answering

We show the visualization of the video-question answering in Fig. D. As shown in Fig. D, our method succeeds in getting the ground-truth answer. These results demonstrate that our method can deal with cross-modal inference task effectively.

D.3. Hierarchical Interaction

To better understand the proposed method, we show the visualization of the hierarchical interaction in Fig. E, Fig. F and Fig. G. This experiment shows that our Hierarchical Banzhaf Interaction (HBI) can effectively handle fine-grained semantic interaction with diverse granularity, flexible combination, and vague intensity. More encouragingly, 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.

References

[1] 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/CVF International Conference on Computer Vision*, pages 5803–5812, 2017. 1

[2] 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. 3

[3] John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317, 1964. 2

[4] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[6] Pradeep Dubey. On the uniqueness of the shapley value. *International Journal of Game Theory*, 4(3):131–139, 1975. 2

[7] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of the European Conference on Computer Vision*, volume 12349, pages 214–229, 2020. 1

[8] 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

[9] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 706–715, 2017. 1

[10] Dennis Leech. Computation of power indices. 2002. 3

[11] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *Proceedings of the European Conference on Computer Vision*, pages 110–127. Springer, 2022. 1

[12] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *Proceedings of the British Machine Vision Conference*, page 279, 2019. 1

[13] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 319–335. Springer, 2022. 9

[14] 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 and captioning. *Neurocomputing*, 508:293–304, 2022. 1

[15] Yasuko Matsui and Tomomi Matsui. Np-completeness for calculating power indices of weighted majority games. *Theoretical Computer Science*, 263(1-2):305–310, 2001. 3

[16] 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 *International Conference on Learning Representations*, 2021. 1

[17] 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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 1

[18] 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. 2

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. 1

[20] 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. 1

[21] 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. 1

[22] 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 ACM international conference on Multimedia*, pages 1645–1653, 2017. 1

[23] Jun Xu, Tao Mei, Ting Yao, and Yong 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*, pages 5288–5296, 2016. 1

[24] 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*, pages 471–487, 2018. 1