A Supplementary Material

In this supplementary material, we provide following ablation studies and analysis. (1) More detailed analysis of the collected dataset. (2) Ablation on freezing part of SlowFast and varying the model size. (3) Analysis of recommendation results. We show that they are reasonable and plausible by following common video editing guidelines. (4) Experimental results of adding "direct cut" in transitions recommendation. (5) Visualization of demo videos rendered with 30 transitions types.

A.1 Analysis of Dataset

We conduct experiments to analyze more relations between transitions and the properties of visual/audio inputs. Specifically, we use two classifiers of video style/audio mood trained on in-house datasets to get labels of video shots and audio respectively. Then we get the statistical results of the frequencies of transitions w.r.t. each category, followed by a column-wise normalization. As visualized in Fig. A.1, one can get some hints of how transitions match with visual/audio semantics. To give a few examples, "floodlight" and "black fade" appear more with the visual style of "fresh" and "nostalgic" respectively. Both "star" and "heart" tend to come with the audio mood of "sweet" and "romantic". Note these results comply with general preferences in video editing, which evidences the feasibility of learning meaningful correspondence from inputs to transitions using the dataset.

A.2 Model Architecture Ablation

Impact of freezing SlowFast backbone. When training the recommendation model, we freeze stage one to stage three of the SlowFast network by default. To verify its impact, we do some experiments with different freezing proportions of SlowFast network in the setting with only visual as input. From the results in Table A.1, we can see that freezing all parameters severely damages the performance of the model than freezing the parameters from stage 1-3. While not

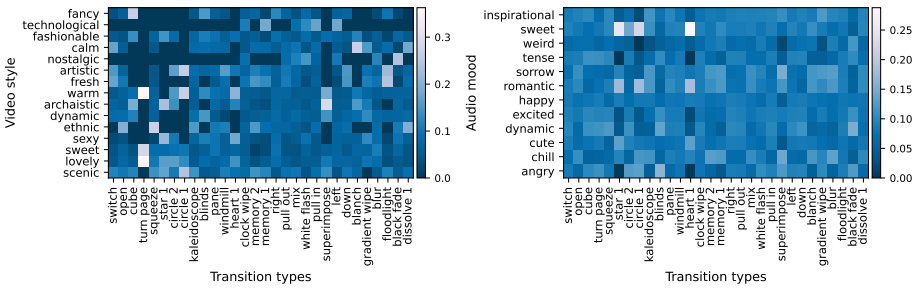

Fig. A.1. The relationships between video style, audio mood and transition types.

Table A.1. The impact of freezing the SlowFast backbone in visual-only setting. By default, we freeze stage 1-3 to facilitate testing.

SlowFast Freezing	Recall@1	Recall@5	Mean Rank
Freeze all parameters	22.39%	26.53%	6.097
Freeze stage 1-3	25.40%	66.33%	5.665
No freezing	**25.97%**	**66.95%**	**5.579**

Table A.2. Ablation study on the model size. Where N is the number of transformer encoder layers, d_{model} is the dimension of transformer layers, and d_{matching} is the dimension of the common matching space.

	N	d_{model}	d_{matching}	Recall@1	Recall@5	Mean Rank
base	2	2048	2048	**28.06%**	**66.85%**	**5.480**
(a)	2	2048	1024	26.59%	67.09%	5.493
	2	2048	512	26.40%	67.07%	5.499
(b)	2	1024	2048	25.52%	66.71%	5.598
	2	4096	2048	26.93%	66.55%	5.541
(c)	1	2048	2048	25.77%	66.45%	5.623
	4	2048	2048	27.26%	66.47%	5.528

freezing any parameters only improves performance by a small amount than freezing the parameters from stage 1-3. So in order to balance the performance and efficiency of the algorithm in training, we choose to freeze stage one to stage three of the SlowFast network by default.

Variations on the model architecture. In this experiment, we study the influence of the different model architecture and the embedding dimension. In Table A.2 (a), the influence of different dimension in the embedding matching space is investigated. Reducing the dimension of matching embedding does harm to the performance. From Table A.2 (b) and (c), we observe that a smaller feature dimension or number of the transformer layers in the transformer network also hurts performance. At the same time, a larger feature dimension or number of the transformer layers has adverse effects on performance, indicating that large model size may cause severe over-fitting.

A.3 Results Analysis

In this section, we show through concrete examples that our method indeed learns the general guidelines of using transitions in video editing. Therefore, the videos generated by our method comply with common cinematography knowledge and aesthetic principles.

As shown in Fig. A.2, our method can capture the visual changes in neighboring video shots and thus recommend suitable video transitions. In Fig. A.2 (a), since the neighboring video shots have similar scenes, our method ranks gentle transitions higher, such as “mix”, “black fade”, “dissolve” and “blur”, in order

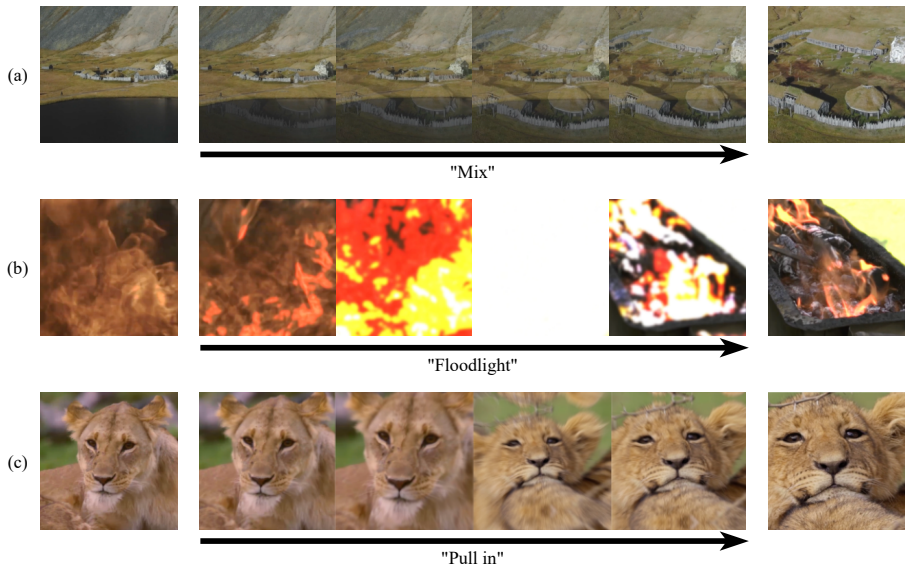

Fig. A.2. The video transition recommendations generated by our method. In each row, the leftmost and the rightmost images represents the video shots before and after transitions respectively. The frames in middle represent the video clip in transition. For better visualization, please refer to videos (a)-(c) under folder "video/fig-A-2".

to keep the coherence of video narration. In Fig. A.2 (b) where neighboring video shots have large scene changes and brightness changes, our method recommends more dynamic transitions like "floodlight". In Fig. A.2 (c) where a long video shot switches to a shorter one, our method recommends the "pull in" to ensure natural transition .

In Fig. A.3 (a), we show an example that our method is able to adapt to the characteristics of music. Since the background music used in this example is with soft tune, our method prefers to recommend gentle transitions. Otherwise using abrupt transitions may break the visual-auditory harmonious. In contrast, since the weighted random pick method does not consider visual/audio contents, its results are less reasonable (e.g. using too much "flashing") as shown in Fig. A.3 (b).

We show another comparing example among three methods in Fig. A.4. Compared with weighted random pick (Fig. A.4 (a)), our method (Fig. A.4 (b)) recommends better results which are reflected by the consistency in sequential transitions predictions, as well as the nice matching with visual and audio contents. While the professional editor (Fig. A.4 (c)) may be advantageous in capturing the details in videos like the specific movement pattern of basketball players to select dynamic transitions, we note our method is much faster in terms of efficiency. This comparison result also motivates us to use more fine-grained video embedding to further improve our method, which we leave for future work.

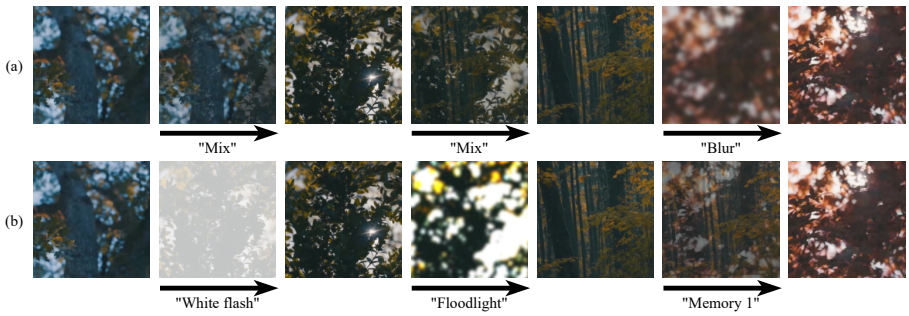

Fig. A.3. Transitions recommended for video shots with soft background music (see videos (a)-(b) under folder "video/fig-A-3"). (a) and (b) corresponds to the results of our method and weighted random pick respectively.

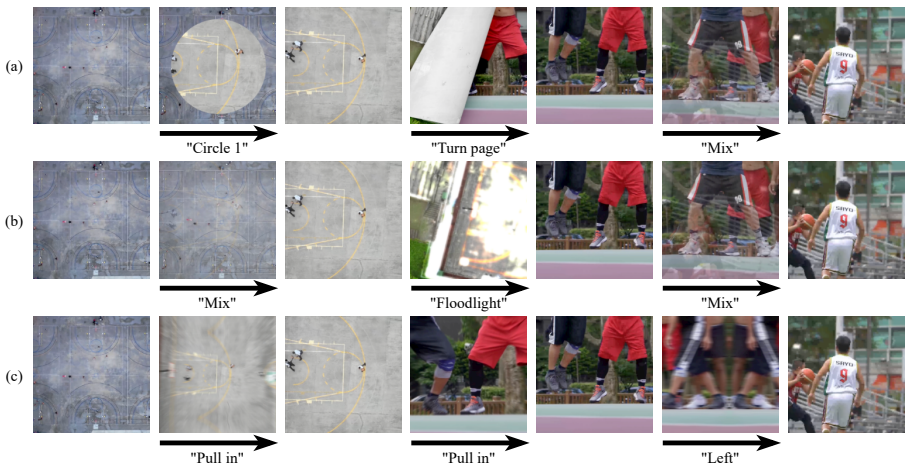

Fig. A.4. A comparison of the transitions selected by three different methods (see videos (a)-(c) under folder "video/fig-A-4"). Transitions in (a), (b) and (c) are selected by weighted random pick, our method and professional editor respectively.

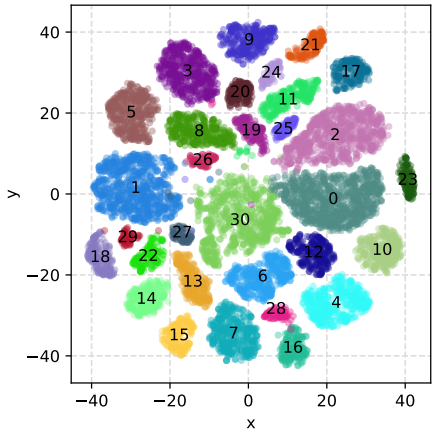

Fig. A.5. t-SNE visualization of the pre-trained transition embedding (with "direct cut", the index of "direct cut" is 30).

Table A.3. Ablative experiments on direct cut.

Modal	with "direct cut"	R@1 ↑	R@5 ↑	Mean Rank ↓
Visual+Audio		28.06%	66.85%	5.480
Visual+Audio	✓	30.57%	67.98%	5.347

A.4 "Direct Cut" and More User Study

We note that "direct cut" is also widely used in video editing by directly connecting shots without using any effects. We add "direct cut" and re-train/re-evaluate our model to verify the extendibility of our proposed framework. We draw similar conclusions from this new experiment as in the main paper and verify both quantitatively and qualitatively that our method can handle this case. Fig. A.5 shows the learned embeddings of transitions including "direct cut". The semantic relationships are still preserved. As shown in Tab. A.3, the performance of model with "direct cut" is on par with that without "direct cut", showing the model can successfully learn the matching from input to transitions.

To further verify the effectiveness of our method, we conduct a more comprehensive user study following the practice introduced in Section 6.4 in the main paper. In this experiment, we hire 10 professional and 10 amateur editors, and direct cut is allowed to use in all comparing methods. As show in Fig. A.6, in terms of both voting result and average score, our method surpasses amateur editors and achieves comparable results with professional editors, therefore demonstrating the effectiveness of our method.

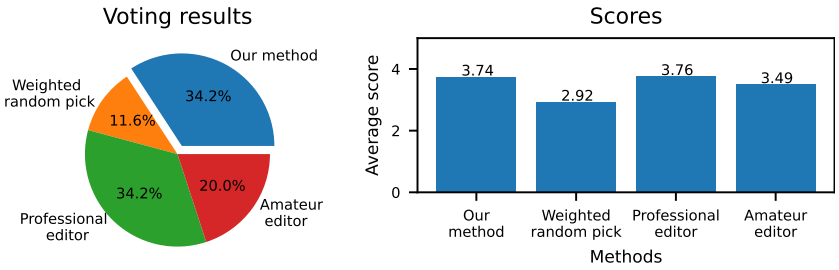

Fig. A.6. User study results with more editors, "direct cut" is allowed in all comparing methods.

A.5 Example videos of all transition effects

To help readers understand video transitions more straightforwardly, we provide example videos created using video transitions and a fixed pair of video shots. Overall 30 demo videos are included in the zipped file, under the folder "video/transitions".