

# SKALD: Learning-Based Shot Assembly for Coherent Multi-Shot Video Creation

## Supplementary Material

### 6. More Details on Data Preparation

**MSV3C training set.** We collected videos from the V3C1 dataset [22], comprising 28,450 videos of 2–5 minutes duration. While V3C1 provides scene segmentation, we found its scene cuts did not accurately align with shot transitions. Therefore, we used TransNetV2 [23] to split videos into individual shots. We filtered out shots shorter than 2 seconds, static frames, and shots containing excessive text overlays. To ensure video diversity, we computed the average CLIP embedding for each shot and analyzed variance across videos. We excluded the lowest variance quartile, typically consisting of interview-style videos with limited visual diversity. For each remaining video, we randomly sampled between 3 and 12 consecutive shots, forming representative multi-shot sequences suitable for training and evaluation. The resulting dataset encompasses 5,648 carefully curated videos, totaling over 12,000 shots.

**MSV3C testing set.** We divided the MSV3C dataset into training and testing sets with a 9:1 ratio. To simulate real-world video editing scenarios, each shot in the test set was manually annotated with a concise single-sentence scene description. These descriptions served as textual queries to retrieve five visually similar candidate shots per original shot from our internal stock footage collection. This design closely reflects practical shot assembly workflows, where the objective is identifying the optimal coherent sequence among visually and semantically related alternatives.

### 7. User Study

We conducted a comprehensive user study to rigorously assess and compare the video assembly quality between CLIP4Clip [17] and SKALD. The primary objective was evaluating these methods’ ability to generate coherent and engaging visual narratives without relying on audio elements.

Our study employed a side-by-side comparison format. Participants simultaneously viewed two multi-shot videos sharing identical thematic content. Each video explicitly excluded voice-overs and music, enabling unbiased evaluation of visual storytelling proficiency. Participants independently rated the quality of each video on a standardized five-point scale from “Very Poor” to “Excellent,” ensuring precise quantification of their qualitative judgments. Participants were provided with ample time—up to 30 minutes—although task completion typically required less than half that duration.

To maintain data reliability, “golden set” video pairs with predetermined ratings were included as quality con-

trols. Submissions failing these controls were excluded. The study involved 3 distinct rating sessions covering 10 unique video pairs, with each pair evaluated multiple times by different participants. This robust redundancy mitigated subjective bias, ensuring comprehensive evaluation.

### 8. More Qualitative Results

The qualitative analysis presented in Fig. 7, Fig. 8, and Fig. 9 underscores that SKALD consistently outperforms conventional text-based methods in preserving visual continuity and narrative coherence. As illustrated in Fig. 7, traditional text-based retrieval successfully identifies individual pizza-making steps but frequently introduces visual discontinuities such as inconsistent lighting conditions or contextually irrelevant scenes. Conversely, SKALD reliably assembles sequences that flow logically through each cooking stage, sustaining thematic unity. Similarly, in Fig. 8, standard retrieval methods often correctly pinpoint discrete kitchen-cleaning tasks yet fail to ensure seamless visual and contextual transitions. SKALD excels by selecting visually coherent and contextually consistent shots, effectively conveying a natural narrative progression. Finally, the environmental awareness promotion in Fig. 9 exemplifies SKALD’s capacity to preserve smooth transitions across visually distinct but semantically connected scenes, significantly enhancing viewer engagement through sustained narrative coherence.

### 9. Video Demo

We provide three example videos (0001.mp4, 0002.mp4, 0003.mp4) assembled with SKALD. The detailed themes and descriptions corresponding to each scene within these videos are comprehensively documented in the accompanying file video\_demo.rtf. These examples demonstrate SKALD’s practical capability in generating high-quality, thematically cohesive multi-shot videos across diverse scenarios.



Figure 7. Qualitative analysis shows that standard text-based retrieval methods readily match individual pizza-making steps but struggle to maintain cohesive transitions between them. By contrast, our approach consistently assembles visually coherent sequences that accurately follow each stage of dough preparation.



Figure 8. Qualitative analysis indicates that basic text-based methods identify the individual cleaning tasks but often lack seamless transitions between them. In contrast, our approach constructs visually coherent sequences that reflect each stage of kitchen organization.



Figure 9. Qualitative analysis indicates that basic text-based methods identify the individual cleaning tasks but often lack seamless transitions between them. Nonetheless, our method ensures smooth transition between shots for a promotion video on environment awareness.