V-Trans4Style: Visual Transition Recommendation for Video Production Style Adaptation

The exponential growth of digital video content, driven by both professional studios and independent creators, and social media platforms, has intensified the demand for rapid, flexible, and high-quality video editing solutions. Central to compelling video storytelling are visual transitions, which connect individual clips into a coherent narrative by signalling changes in time, location or mood. Transitions shape viewer experience through pacing and emotional cues, making them critical for stylistic consistency across diverse production styles such as cinematic film, documentary, vlog or animation. However, crafting transitions that effectively marry narrative flow with style is a complex, manual task requiring substantial skill and time. Existing editing tools often rely on static templates or lack content-awareness, limiting creativity and accessibility for many users.

To overcome these challenges, we introduce V-Trans4Style, a novel learning-based algorithm that automates visual transition recommendation by integrating content and production style awareness. Central to the method is a transformer-based encoder-decoder network that learns to recommend visually and temporally consistent transitions from sequences of video clips. This encoder-decoder is the only trainable component, trained on a large annotated video dataset to model the underlying dynamics of transitions. To support flexible adaptation across a variety of production styles without retraining, V-Trans4Style incorporates a style conditioning module that operates exclusively during inference. This module applies activation maximization on the encoder's latent embeddings to iteratively refine the transition sequence from the decoder, aligning transitions to user-specified styles while preserving narrative coherence. This two-stage, bottom-up architecture ensures that the recommended transitions reinforce both the natural flow of the video content and the intended visual identity, moving beyond brittle, rule-based, or static template solutions for robust style adaptation across diverse production domains.

To support adaptive training and benchmarking, we introduce AutoTransition++, a new dataset of 6,000 videos spanning five production styles, with 1,379 of them receiving detailed, human-verified style labels. This large-scale, style-annotated resource fills a critical gap in evaluating and training data-driven video editing models. Empirical validation on this dataset shows that V-Trans4Style achieves up to 80% improvements in transition recall and rank metrics, along with approximately 12% higher style similarity compared to state-of-the-art baselines. Furthermore, a comprehensive user study with 102 participants confirmed the system's practical effectiveness: users consistently preferred videos whose transitions were refined by the style conditioning module, with over 70% favoring them as closer to the intended style.

We hope that our work serves as a foundation for future exploration and a deeper understanding of video production styles and their interaction with various editing elements. Ultimately, this will enable richer and more personalized storytelling experiences.