BEAT: RHYTHM-ELASTIC ALIGNMENT FOR AGENTIC MUSIC-GUIDED MOVIE TRAILER GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic movie trailer generation selects shots from a full-length film and synchronizes them with chosen music. Existing methods either relegate music alignment to post-processing or enforce rigid one-to-one shot-music mappings, overlooking that professional editing rhythm is *elastic*: energetic passages favor rapid cuts, while quieter passages allow shots to span multiple bars. We introduce **BEAT**, a framework that addresses this gap with two core components: MuVA, a compact music-visual alignment encoder trained with Sinkhorn-regularized two-stage learning, and Bar-DP, an energy-adaptive dynamic programming algorithm that produces elastic many-to-one alignments. These components are integrated into a five-phase agentic pipeline that grounds the core alignment in learned cross-modal features while coordinating higher-level creative decisions through structured text signals. To support comprehensive evaluation, we also introduce TrailerArena, a benchmark with 20+ metrics across four complementary dimensions. On TrailerArena, BEAT achieves state-of-the-art performance across shot selection, ordering, and perceptual quality, while producing fully composed trailers end-to-end.

1 INTRODUCTION

Movie trailers condense a film into a short engaging audiovisual sequence while avoiding key plot revelations. A chosen soundtrack provides emotional direction and rhythmic structure, motivating *music-guided trailer generation* (Xu et al., 2015; Wang et al., 2024b): selecting and arranging shots from a given movie to fit the music. This requires coherent shot selection and *elastic* timing: energetic passages favor rapid cuts, while quieter passages allow a shot to span multiple bars.

Learning-based approaches provide useful mechanisms for shot selection, but shot compatibility and editing duration remain distinct decisions. IPOT (Wang et al., 2024b) and MMSC (Zhu et al., 2025) model shot-music matching through optimal transport or cross-attention, while SSMP (Zhu et al., 2026) predicts shot sequences and handles music alignment in post-processing. A rigid one-to-one matching does not express how a shot can extend across several bars as musical energy changes. Trailer generation therefore calls for an alignment mechanism that couples learned shot-music compatibility with variable shot durations.

Agentic systems offer complementary capabilities for planning and iterative editing. TRAILDREAMS (Balestri et al., 2026) coordinates trailer creation through LLM-driven sequence selection and audio generation; DIRECT (Li et al., 2026a) organizes video mashup creation through hierarchical planning; and CutClaw (Zhao et al., 2026) coordinates agents for music-synchronized editing of long-form footage. However, repeated model calls across planning and editing can incur substantial inference costs. When editing decisions depend on model reasoning and prompt design, rhythmic constraints become difficult to specify and verify explicitly. This motivates a framework that combines agent-based planning with a trainable alignment model and explicit control over shot durations.

Our key idea is to learn bar-shot compatibility, select shot durations according to musical energy, and coordinate creative decisions through agents. We introduce **BEAT** (**B**ar-level **E**lastic **A**lignment for agentic **T**railer generation), which implements this idea through a compact alignment model and elastic selection within a five-phase agentic pipeline. MuVA scores candidate shots for each musical bar; Bar-DP uses these scores and musical energy to select shots and their bar spans. LLM/VLM agents coordinate planning, composition, and iterative critique through structured text signals. Given a movie and a music, BEAT produces a fully composed trailer in a single run. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Learned alignment with elastic timing.** We introduce MuVA, a compact cross-modal encoder for bar-shot compatibility, and Bar-DP, an energy-adaptive selection algorithm that jointly determines shot choice and duration. Their many-to-one alignments support rapid cuts and sustained shots spanning multiple bars within a unified formulation (§3.2, §3.3).
- **Agentic trailer generation.** We integrate learned alignment and elastic selection into a five-phase workflow from music analysis to final composition. Structured text signals connect user instructions, narrative planning, and tagline generation with iterative critique, allowing agent feedback to guide re-selection while retaining explicit control over shot spans (§3.4).
- **Comprehensive trailer evaluation.** We introduce TrailerArena to assess selection, ordering, composition, and perceptual quality, including Fréchet Shot Distance for comparing selected-shot feature distributions. BEAT leads the compared methods in F1, SoftF1@5, and FSD on both Test-8 and Test-73, with ablations examining component contributions (§4, §5).

2 Related Work

Movie Trailer Generation. Early approaches used hand-crafted heuristics on multimodal features (Smith et al., 2017; Smeaton et al., 2006; Hesham et al., 2018), graph-based narrative analysis to detect turning points (Papalampidi et al., 2021), or cinematic grammar to extract logos, dialogue, and action scenes (Chen et al., 2004; Liu & Jiang, 2015; Liu et al., 2023). More recent methods learn trailer composition from paired movie-trailer data. TGT (Argaw et al., 2024) models trailer generation as sequence-to-sequence translation with an autoregressive decoder, while IPOT (Wang et al., 2024b) jointly optimizes shot selection and music alignment through inverse partial optimal transport. MMSC (Zhu et al., 2025) learns weakly supervised trailerness scores through multimodal cross-attention with semantic consistency constraints. SSMP (Zhu et al., 2026) trains a Transformer encoder for masked prediction with self-corrective re-masking. Despite these advances, rigid one-to-one shot-music correspondences do not capture the elastic editing rhythm of professional trailers.

Multi-Agent Systems for Video Production. Agentic workflows coordinate video editing (Wang et al., 2024a; Li et al., 2026b; Ding et al., 2025; Zhou et al., 2026) and film production (Wu et al., 2025b; Xu et al., 2025; Balestri et al., 2026) through specialized roles and tools. EditDuet (Sandoval-Castañeda et al., 2025) couples an editor with a critic, while DIRECT (Li et al., 2026a) organizes video mashup creation into a Screenwriter–Director–Editor hierarchy. CineAgents (Zhang et al., 2026) combines hierarchical narrative memory with iterative planning for instruction-driven cinematic compilation. Closer to music-guided editing, CutClaw (Zhao et al., 2026) coordinates narrative planning, editing, and review to synchronize selected footage with music. BEAT combines agent coordination with a trained bar-shot alignment model and energy-adaptive elastic selection, using structured text signals to convey instructions, taglines, and critique.

3 Proposed Method

We present BEAT, a framework for music-guided movie trailer generation. Given a movie and a piece of trailer music, BEAT selects and arranges movie shots to produce a trailer that is rhythmically synchronized with the music, narratively coherent, and spoiler-free. We first formalize the problem (§3.1), then describe the three core components: the Music-Visual Aligner (MuVA, §3.2), the elastic Bar-DP selection algorithm (§3.3), and the agentic generation pipeline (§3.4).

3.1 Problem Formulation

Let $\mathcal{M} = \{s_1, s_2, \ldots, s_I\}$ denote the set of I shots obtained by segmenting a full-length movie via a shot boundary detector (Soucek & Lokoc, 2024). Each shot s_i is characterized by a visual feature vector $\mathbf{v}_i \in \mathbb{R}^{d_v}$ extracted by a pretrained visual encoder, and a temporal duration $\delta_i > 0$.

The trailer music is segmented into J musical bars $\mathcal{B} = \{b_1, b_2, \ldots, b_J\}$ via change-point detection on the audio signal (Truong et al., 2020). Each bar b_j is represented by an audio feature vector $\mathbf{a}_j \in \mathbb{R}^{d_a}$ extracted by a pretrained audio encoder, a duration $\ell_j > 0$, and an energy level $e_j \in [0, 1]$ computed as the normalized RMS amplitude.

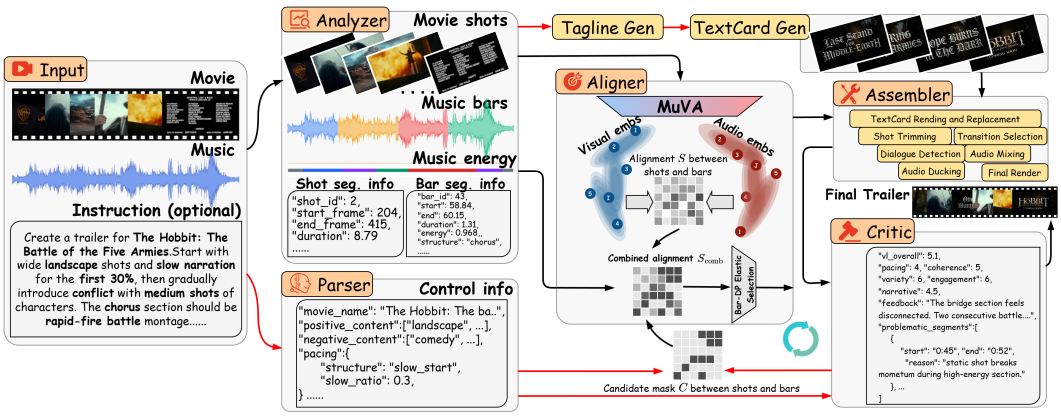


Figure 1: **Overview of BEAT.** Given a movie, trailer music, and an optional text instruction, the pipeline operates in five phases. The **Analyzer** segments the movie into shots and the music into bars, extracting visual/audio features and per-bar energy estimates. An optional **Parser** converts free-form instructions into a structured control information that modifies the candidate mask. The **Aligner** computes cross-modal alignment scores, which are fused with energy-dynamics matching and fed to Bar-DP for elastic shot selection. The **Assembler** generates taglines, renders styled text cards, applies transitions and dialogue ducking, and produces the final trailer. The **Critic** reviews the rendered trailer and returns timestamped feedback to trigger re-selection if needed. Structured text signals (red arrows) coordinate modules across phases.

Elastic Alignment. We define a trailer as an *elastic alignment* $\boldsymbol{\pi}$ that maps the music bar sequence to a shot assignment sequence. Formally, $\boldsymbol{\pi} = \big((c_1, r_1), \ldots, (c_K, r_K)\big)$, where segment k assigns shot $c_k \in \{1, \ldots, I\}$ to bars $[r_k, r_{k+1})$, with $r_1 = 1, r_{K+1} = J+1$. The number of segments $K \leq J$ is determined by the algorithm, and a single shot may span multiple consecutive bars (many-to-one mapping), while each bar belongs to exactly one shot. This elastic formulation generalizes the rigid one-to-one mapping used in prior works (Wang et al., 2024b; Zhu et al., 2026; 2025; Xu et al., 2015).

Objective. The goal is to find an elastic alignment $\boldsymbol{\pi}^*$ that maximizes a composite quality score:

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \Pi} \sum_{k=1}^{K} \Big[\underbrace{\bar{f}(c_k, r_k, r_{k+1})}_{\text{alignment}} + \lambda_{\text{sm}} \cdot \underbrace{g(c_{k-1}, c_k)}_{\text{smoothness}} + \lambda_{\text{cut}} \cdot \underbrace{h(e_{r_k : r_{k+1}})}_{\text{cut bonus}} \Big], \tag{1}$$

subject to no-repeat constraints $c_k \neq c_{k'}$ for $k \neq k'$ and duration constraints $\sum_{j=r_k}^{r_{k+1}-1} \ell_j \leq \delta_{c_k}/\eta$ for spans of multiple bars. Here, $\bar{f}$ measures the average alignment quality between a shot and its assigned bars, g penalizes extreme visual discontinuities between consecutive shots (without enforcing similarity, since trailers often use deliberate contrast), and h provides an energy-adaptive cut bonus. Π denotes the space of valid elastic alignments.

3.2 MuVA: Music-Visual Aligner

MuVA is the core alignment model that scores the compatibility between each musical bar b_j and each movie shot s_i. It produces an alignment score matrix $\mathbf{S} \in \mathbb{R}^{J \times I}$ where $\mathbf{S}_{ji}$ indicates how well shot s_i matches bar b_j.

3.2.1 Architecture

MuVA takes as input the raw audio features $\mathbf{a}_j \in \mathbb{R}^{d_a}$ extracted by CLAP (Wu et al., 2023) (d_a=512) and visual features $\mathbf{v}_i \in \mathbb{R}^{d_v}$ extracted by ImageBind (Girdhar et al., 2023) (d_v=1024). Linear projections with RMSNorm map both modalities into a shared d-dimensional space, yielding $\tilde{\mathbf{a}}_j, \tilde{\mathbf{v}}_i \in \mathbb{R}^d$. Each modality is then processed independently by L Transformer blocks with Rotary Position Embeddings (RoPE) to capture temporal ordering, producing self-attended representations $\bar{\mathbf{a}}_j$ and $\bar{\mathbf{v}}_i$. The two streams interact through L_{ca} layers of bidirectional cross-attention (audio attends

to visual, then visual attends to the updated audio), yielding the final cross-modal representations $\hat{\mathbf{a}}_j$ and $\hat{\mathbf{v}}_i$. The alignment score between bar j and shot i is then computed via temperature-scaled cosine similarity:

$$\mathbf{S}_{ji} = \frac{\langle \hat{\mathbf{a}}_j, \hat{\mathbf{v}}_i \rangle}{\|\hat{\mathbf{a}}_j\| \cdot \|\hat{\mathbf{v}}_i\|} \cdot \frac{1}{\tau}, \tag{2}$$

where τ is a learnable temperature parameter.

3.2.2 Two-Stage Training

MuVA is trained in two stages to leverage both abundant video-music data and scarce trailer-specific annotations.

Stage 1: Pretraining on Video-Music Data. We pretrain on large video-music pairs from V2M (Tian et al., 2025). Since these pairs lack bar-shot alignment labels, we construct soft targets based on temporal IoU between music bars and visual shots. Bar-shot pairs with greater temporal overlap receive higher alignment scores, while non-overlapping pairs are assigned zero. The pretraining loss combines KL divergence on alignment distributions with an InfoNCE (Oord et al., 2018) contrastive term on pooled representations:

$$\mathcal{L}_{\text{pretrain}} = \mathrm{KL}\big(\mathrm{softmax}(\mathbf{S}^{\text{soft}}/\tau_t) \parallel \mathrm{softmax}(\mathbf{S}/\tau_t)\big) + \lambda_{\text{con}} \cdot \mathcal{L}_{\text{InfoNCE}}. \tag{3}$$

Stage 2: Fine-tuning on Trailer Alignments. We fine-tune on trailer-movie pairs with binary bar-to-shot annotations $\mathbf{S}^{\text{gt}}$, allowing multiple positive shots per bar. For each annotated bar, we normalize the target as $q_{ji} = \mathbf{S}^{\text{gt}}_{ji} / \sum_{i'} \mathbf{S}^{\text{gt}}_{ji'}$ and define $\mathbf{p}_j = \mathrm{softmax}(\mathbf{S}_{j,:}/\tau_t)$ over candidate shots. KL alignment supervises per-bar correctness, but without structural constraints, a few visually distinctive shots can dominate scores across many bars. We add Sinkhorn regularization (Cuturi, 2013), which iteratively projects $\mathbf{S}$ toward a doubly-stochastic matrix, enforcing that each bar focuses on few shots and each shot is not claimed by too many bars:

$$\mathbf{S}^{(0)} = \exp(\mathbf{S}/\tau_s), \quad \mathbf{S}^{(t+1)} = \mathcal{N}_c\big(\mathcal{N}_r(\mathbf{S}^{(t)})\big), \quad t = 0, \ldots, T{-}1, \tag{4}$$

where $\mathcal{N}_r$ and $\mathcal{N}_c$ normalize rows and columns respectively, and τ_s controls the sharpness of the target. The regularization loss $\mathcal{L}_{\text{sink}}$ encourages the model's raw scores to approximate this structured target. We use $T{=}3$ iterations, applying a gentle structural prior; heavier settings ($T \geq 10$) over-constrain the model and degrade performance. The full fine-tuning loss is:

$$\mathcal{L}_{\text{ft}} = \underbrace{\frac{1}{J}\sum_{j=1}^{J} \mathrm{KL}\big(\mathbf{q}_j \parallel \mathbf{p}_j\big)}_{\mathcal{L}_{\text{KL}} \text{ for selecting correctly}} + \lambda_{\text{sink}} \cdot \underbrace{\mathrm{KL}\big(\mathbf{S}^{(T)} \parallel \mathrm{softmax}(\mathbf{S})\big)}_{\mathcal{L}_{\text{sink}} \text{ for distributing evenly}} + \lambda_{\text{con}} \cdot \underbrace{\mathcal{L}_{\text{InfoNCE}}}_{\substack{\mathcal{L}_{\text{InfoNCE}} \text{ for globally} \\ \text{aligning modalities}}}, \tag{5}$$

Bars without positive annotations contribute zero to $\mathcal{L}_{\text{KL}}$; losses are averaged over examples.

3.3 Bar-DP: Energy-Adaptive Elastic Selection

Given the alignment score matrix $\mathbf{S} \in \mathbb{R}^{J \times I}$ from MuVA, Bar-DP approximates the optimum in Eq. (1) via beam search over the bar sequence. The key insight is that editing pace should adapt to musical dynamics: high-energy passages warrant rapid one-bar-one-shot cuts, while low-energy passages benefit from sustained shots spanning multiple bars.

State Representation. A state at bar position j is a tuple $\sigma = (\text{score}, \text{assignments}, \text{used}, \text{last})$. Here score $\in \mathbb{R}$ is the accumulated quality, assignments records shot-bar span allocations, used $\subseteq \{1, \ldots, I\}$ tracks assigned shots and their visually similar neighbors, and last is the most recently assigned shot. We maintain a separate beam $\mathcal{H}_j$ for states whose next unassigned bar is j; assigning a span of k bars advances a state to $\mathcal{H}_{j+k}$.

Transition Scoring. At each bar j, we consider assigning one of the top-M candidate shots from $\mathbf{S}_{j,:}$ with a bar span of $k \in \{1, \ldots, k_{\max}\}$:

$$\text{score}(i, j, k) = \underbrace{\frac{1}{k} \sum_{j'=j}^{j+k-1} \mathbf{S}_{j'i}}_{\text{avg. alignment}} + \lambda_{\text{sm}} \cdot \underbrace{\cos(\mathbf{v}_{\text{last}}, \mathbf{v}_i)}_{\text{smoothness}} + \underbrace{\phi(e_{j:j+k})}_{\text{adaptive cut bonus}}, \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and the energy-adaptive cut bonus ϕ is defined as:

$$\phi(e_{j:j+k}) = \lambda_{\text{cut}} \cdot (\bar{e}_{j:j+k} - 0.15), \tag{7}$$

with $\bar{e}_{j:j+k} = \frac{1}{k} \sum_{j'=j}^{j+k-1} e_{j'}$ being the average energy of the bar span. When $\bar{e} = 1$ (high energy), the bonus is $+\lambda_{\text{cut}} \cdot 0.85$, encouraging cuts. When $\bar{e} = 0$ (low energy), the penalty is $-\lambda_{\text{cut}} \cdot 0.15$, discouraging unnecessary transitions.

Duration Constraint. A shot s_i may span $k > 1$ bars only if its duration δ_i can accommodate the total bar span:

$$\sum_{j'=j}^{j+k-1} \ell_{j'} \leq \frac{\delta_i}{\eta}, \tag{8}$$

where η controls the permitted playback slowdown.

Neighbor Exclusion. To prevent redundant selections, we exclude previously used shots and their visually similar neighbors:

$$\text{used}' = \text{used} \cup \{i' : \cos(\mathbf{v}_i, \mathbf{v}_{i'}) > \theta_{\text{sim}}\}, \tag{9}$$

where θ_{sim} is the similarity threshold. We denote this exclusion set, including i itself, by $\mathcal{N}(i)$.

Algorithm 1 Bar-DP Elastic Selection

Require: $\mathbf{S}^{\text{comb}}$, $\{\ell_j\}$, $\{\delta_i\}$, $\{e_j\}$, $\mathcal{N}$
Ensure: Approximate elastic alignment $\hat{\boldsymbol{\pi}}$
  $\mathcal{H}_1 \leftarrow \{(0, [\,], \emptyset, \texttt{null})\}$
  $\mathcal{H}_2, \ldots, \mathcal{H}_{J+1} \leftarrow \emptyset$
  for $j = 1$ **to** J **do**
    $\mathcal{H}_j \leftarrow$ Top-$W(\mathcal{H}_j)$
    for $(sc, asgn, \mathcal{U}, prev) \in \mathcal{H}_j$ **do**
      for $i \in$ Top-$M(\mathbf{S}^{\text{comb}}_{j,:}) \setminus \mathcal{U}$ **do**
        for $k = 1$ **to** $\min(k_{\max}, J - j + 1)$ **do**
          if $\mathbf{S}^{\text{comb}}_{j+k-1,i} = -\infty$ **or** $(k > 1$ **and** $\sum_{j'=j}^{j+k-1} \ell_{j'} > \delta_i / \eta)$ **then**
            break
          end if
          $\Delta \leftarrow \text{score}(i, j, k)$
          $\mathcal{H}_{j+k} \leftarrow \mathcal{H}_{j+k} \cup \{(sc + \Delta,$
            $asgn \,\|\, (i, j, k), \mathcal{U} \cup \mathcal{N}(i), i)\}$
        end for
      end for
    end for
  end for
  if $\mathcal{H}_{J+1} = \emptyset$ **then return** failure
  $\sigma^* \leftarrow \arg\max_{\sigma \in \mathcal{H}_{J+1}} \sigma.sc$
  return alignment encoded by $\sigma^*.asgn$

Algorithm 1 expands the top-W states per bar using finite-score candidates and valid spans. Transitions follow Eq. (6) with $\mathbf{S}^{\text{comb}}$, last = *prev*, and zero initial smoothness; $\|$ appends an assignment. Each selected span $k_t^* = r_{t+1} - r_t$ is recovered from the best complete path; an empty $\mathcal{H}_{J+1}$ means no feasible completion was found. Hyperparameters are listed in Appendix A.3.

Complexity Analysis. Expanding W states over J bars, M candidate shots, and up to $k_{\max}$ spans costs $O(J \cdot W \cdot M \cdot k_{\max})$. Precomputing $\mathbf{S} \in \mathbb{R}^{J \times I}$ and shot similarities for neighbor exclusion costs $O(J \cdot I \cdot d)$ and $O(I^2 \cdot d_v)$, respectively. With the settings in Appendix A.3, selection takes under 1.0 s on a single CPU core, negligible compared to feature extraction and video rendering.

3.4 Agentic Trailer Generation Pipeline

Professional trailer editors do not produce a final cut in a single pass: they analyze the music, survey the footage, draft a rough cut, review it, and revise repeatedly. BEAT mirrors this workflow through five phases connected by structured text signals, as shown in Figure 1. Unlike black-box agentic systems (Zhao et al., 2026), where all reasoning is delegated to LLMs, BEAT grounds the core alignment in a trained model. LLM-generated text coordinates higher-level decisions: user instructions flow as keywords into the candidate mask, VLM taglines flow into composition, and critic feedback flows back to re-selection.

Phase 1–2: Analysis and Candidate Preparation. **MusicAnalyzer** segments the music into bars via change-point detection (Truong et al., 2020), extracts CLAP features (Wu et al., 2023), and estimates per-bar energy and musical structure. **NarrativeGuard** filters the shot pool by excluding

Table 1: TrailerArena evaluation dimensions, quality questions, and representative metrics. The complete metrics can be found in Appendix A.5, Table 5.

Dim.	Question	# Metrics	Representative Metrics
Selection	Did the method select the right shots?	8	F1, SoftF1@K, FSD
Ordering	Are shots ranked and ordered correctly?	8	LD, AA, τ, R@K, mAP@t
Composition	Is the assembled trailer well-composed?	3	SDTW, AQ, BeatAlign
Perceptual	Does the trailer look/feel good to a viewer?	6	VL-Overall

spoiler-prone regions (final 15%, opening logos) and scoring shot importance based on visual distinctiveness and position, producing a safe mask $\mathbf{m} \in [0,1]^I$. The safe mask is broadcast to form a candidate mask $\boldsymbol{C} \in \{0,1\}^{J\times I}$. When a user provides a free-form text instruction (e.g., "focus on action scenes"), an LLM parses it into structured control signals, and ImageBind embeds the keywords into visual feature space to refine $\boldsymbol{C}$.

Phase 3: Alignment and Selection. MuVA computes the alignment matrix $\mathbf{S}$ (§3.2), which is fused with energy-dynamics matching and shot importance:

$$\mathbf{S}^{\text{comb}}_{ji} = \begin{cases} \mathbf{S}_{ji} + \lambda_e e_j d_i + \lambda_{\text{imp}} p_i, & C_{ji} = 1, \\ -\infty, & C_{ji} = 0. \end{cases} \tag{10}$$

The candidate mask is applied after score fusion so excluded shots remain infeasible. $\mathbf{e} \in \mathbb{R}^J$ is per-bar energy, $\mathbf{d} \in \mathbb{R}^I$ is per-shot visual dynamics (ℓ_2 norm of consecutive feature differences), and $\mathbf{p} \in \mathbb{R}^I$ is shot importance. The energy term increases the preference for dynamic shots as musical energy rises. The shot selector produces the final sequence from $\mathbf{S}^{\text{comb}}$ based on Algorithm 1.

Phase 4: Composition. A VLM-based Tagline Generator (Qwen3-VL-30B) watches a montage of key shots and produces short tagline texts with a recommended style; a Text Card Generator based on Qwen-Image (Wu et al., 2025a) renders them as stylized overlays. Each shot is trimmed to its bar duration with energy-driven transitions (dissolve / flash / hard cut). Silero VAD (Silero Team, 2024) detects dialogue for audio ducking, and all components are rendered into the final trailer via FFmpeg.

Phase 5: VLM Critic. Qwen3-VL-30B (Bai et al., 2025) watches the rendered trailer at 1fps and rates it across five perceptual dimensions (1–10 each). If quality is insufficient, the critic returns timestamped text feedback identifying problematic segments and reasons, which is mapped back to bar indices to ban corresponding shots and trigger re-selection and re-rendering for up to 3 rounds.

4 TrailerArena: A Benchmark for Trailer Generation

Prior trailer generation methods primarily use set-based (e.g., F1) and ordering metrics (e.g., LD, AA) (Wang et al., 2024b; Zhu et al., 2026). Index-based overlap can reward shots concentrated in a narrow movie region while overlooking visual similarity. These metrics leave visual distributions, overall trajectory similarity, and perceptual appeal unassessed. **TrailerArena** addresses these gaps with 20+ metrics across four complementary dimensions (Table 1).

A. Shot Selection. Set-based metrics (F1, F1@K, IoU) measure binary overlap with GT; soft metrics (SoftF1@K, Chamfer Distance) award partial credit for near-misses. Neither directly measures *visual* similarity. Analogous to FID, **Fréchet Shot Distance** (FSD) computes the Fréchet distance between Gaussians fitted to ImageBind features of predicted and GT shot sets:

$$\text{FSD} = \|\boldsymbol{\mu}_{\hat{\mathcal{S}}} - \boldsymbol{\mu}_{\mathcal{S}^*}\|^2 + \text{Tr}\big(\boldsymbol{\Sigma}_{\hat{\mathcal{S}}} + \boldsymbol{\Sigma}_{\mathcal{S}^*} - 2(\boldsymbol{\Sigma}_{\hat{\mathcal{S}}}\boldsymbol{\Sigma}_{\mathcal{S}^*})^{1/2}\big). \tag{11}$$

The mean and covariance terms capture mismatches in content *type* and diversity, respectively, providing a method-agnostic measure of visual-semantic distributions. Appendix A.5 (Table 6) compares FSD, F1, and Chamfer Distance through diagnostic scenarios.

B. Ordering. Levenshtein Distance (LD) and Alignment Accuracy (AA) evaluate shot ordering on the overlap subsequence, following the SSMP protocol Zhu et al. (2026). LD penalizes both missing shots and wrong ordering; AA measures pairwise ordering agreement. Kendall τ provides a complementary global ordering signal on the full predicted sequence.

Table 2: TrailerArena main results on Test-8 (8 movies) and Test-73 (73 movies). Best in **bold**, second best underlined. Methods for video summarization are marked with a dagger (†).

		Selection			Ordering			Composition		Perceptual
Test	Method	F1↑	SF1@5↑	FSD↓	LD↓	AA↑	τ↑	SDTW↓	AQ↑	VL-Overall↑
Test-8	VASNet† (Fajtl et al., 2018)	0.096	0.276	0.197	100.6	0.441	-0.004	0.351	4.026	–
	CLIP-It† (Narasimhan et al., 2021)	0.074	0.231	0.247	101.2	0.414	-0.025	0.382	4.026	–
	OTVS† (Wang et al., 2023)	0.076	0.243	0.201	101.4	0.475	-0.040	0.368	3.724	–
	Muvee† (Ganhör, 2014)	0.065	0.129	0.362	103.5	0.370	**0.271**	0.396	4.255	5.53
	CutClaw (Zhao et al., 2026)	0.036	0.292	0.433	100.3	0.027	0.034	0.352	4.147	6.13
	V2T (Irie et al., 2010)	0.054	0.154	0.240	103.8	0.529	0.025	0.404	4.288	5.85
	PPBVAM (Xu et al., 2015)	0.083	0.213	0.191	101.5	0.531	0.040	0.373	4.288	5.25
	IPOT (Wang et al., 2024b)	0.131	0.278	0.259	**99.8**	0.538	0.055	0.349	3.721	6.48
	MMSC (Zhu et al., 2025)	0.139	0.271	0.238	101.1	0.394	−0.017	0.392	3.724	6.30
	SSMP (Zhu et al., 2026)	0.162	0.324	0.198	100.0	**0.568**	−0.036	0.390	4.856	6.15
	BEAT (Ours)	**0.193**	**0.373**	**0.143**	100.2	0.529	0.060	**0.330**	**4.919**	**6.58**
Test-73	VASNet† (Fajtl et al., 2018)	0.077	0.235	0.260	84.8	0.441	0.003	0.367	4.182	–
	CLIP-It† (Narasimhan et al., 2021)	0.046	0.173	0.316	85.8	0.389	0.013	0.371	4.096	–
	OTVS† (Wang et al., 2023)	0.077	0.235	0.276	84.8	0.400	-0.012	0.375	4.237	–
	IPOT (Wang et al., 2024b)	0.104	0.262	0.241	83.6	0.505	0.028	0.365	**4.291**	5.86
	MMSC (Zhu et al., 2025)	0.197	0.280	0.222	82.5	0.500	0.020	0.367	4.197	5.81
	SSMP (Zhu et al., 2026)	0.226	0.325	0.212	**77.3**	**0.526**	**0.095**	0.333	4.262	5.81
	BEAT (Ours)	**0.233**	**0.403**	**0.175**	81.1	0.523	0.049	**0.323**	4.263	**5.94**

C. Composition. Sequence DTW (SDTW) compares predicted and GT trajectories in visual feature space, jointly capturing selection, ordering, and visual similarity. Aesthetic Quality (AQ) measures whether the method picks visually striking frames via a pretrained CLIP aesthetic predictor.

D. Perceptual. A vision-language model (Qwen3-VL-30B) Bai et al. (2025) watches each generated trailer and rates it on Pacing, Coherence, Variety, Engagement, and Narrative (1–10 scale).

5 Experiments

5.1 Implementation Details

Dataset. For the training datasets, Stage-1 pretraining uses ∼4,500 video-music pairs from V2M (Tian et al., 2025). Stage-2 fine-tuning uses 870 trailer-movie pairs with ground-truth bar-to-shot alignments: 472 from MMSC (Zhu et al., 2025) and 398 from CMTD (Wang et al., 2024b). For the test datasets, following MMSC (Zhu et al., 2025), we use Test-8 and Test-73 for fair evaluation.

Baselines. We compare our proposed method with the state-of-the-art trailer generation methods V2T (Irie et al., 2010), M2T, PPBVAM (Xu et al., 2015), IPOT (Wang et al., 2024b), MMSC (Zhu et al., 2025), and SSMP (Zhu et al., 2026). When evaluating the capability of shot selection, we also include video summarization baselines VASNet (Fajtl et al., 2018), CLIP-It (Narasimhan et al., 2021), OTVS (Wang et al., 2023), agentic editing model CutClaw (Zhao et al., 2026), and commercial software Muvee (Ganhör, 2014).

Model architecture and Hyperparameter settings. We use ImageBind (Girdhar et al., 2023) (d_v=1024) as the visual encoder and CLAP HTSAT-tiny (Wu et al., 2023) (d_a=512) at 48kHz as the acoustic encoder. We use TransNetV2 (Soucek & Lokoc, 2024) for shot detection and Ruptures (Truong et al., 2020) for change-point detection. All experiments are conducted on 1 NVIDIA H200 GPU. We use Qwen-Image (Wu et al., 2025a) for text-over-video shot generation. Appendix A.2 provides the MuVA architecture and training hyperparameters.

5.2 Quantitative Evaluation on TrailerArena

Table 2 compares all four TrailerArena dimensions on Test-8 and Test-73. BEAT leads shot selection (F1, SoftF1@5, FSD) on both sets. SoftF1@5 additionally credits nearby shot indices, while the lowest FSD indicates the closest match to the GT visual feature distribution. VASNet and OTVS achieve reasonable F1 but much higher FSD, favoring general visual saliency over trailer-appropriate content. SSMP leads AA on both sets and LD on Test-73, while IPOT leads LD on Test-8. BEAT

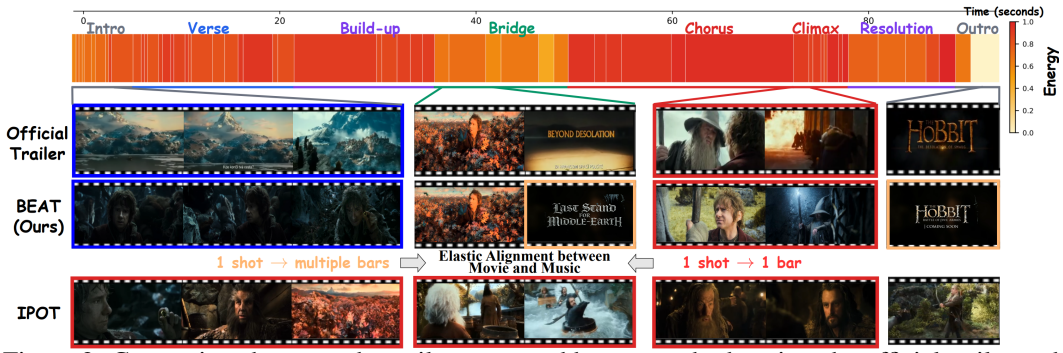


Figure 2: Comparison between the trailer generated by our method against the official trailer and IPOT (Wang et al., 2024b). The top row shows the breakdown of the trailer's music into bars, along with the energy level for each bar. The shots enclosed in blue boxes are long takes that span multiple music bars, while the red boxes indicate quick cuts where each bar corresponds to a separate shot. The shots enclosed in gold boxes are text-over-video shots generated by our pipeline using Qwen-Image (Wu et al., 2025a).

ranks second on both LD and AA on Test-73. Muvee leads τ on Test-8 by selecting few, long shots that preserve chronological order; BEAT ranks second, with RoPE-based self-attention capturing temporal structure. BEAT also achieves the lowest SDTW on both sets, indicating the closest overall trajectory to GT in selection, ordering, and visual content. It leads VLM-rated perceptual quality on both sets, with the largest gain on Test-8. Table 4 further examines the effects of individual pipeline components and selection strategies.

5.3 Qualitative Analysis

Figure 2 compares BEAT and IPOT with the official trailer on one test movie. BEAT adapts its editing rhythm to musical dynamics: it sustains shots across multiple bars during the low-energy bridge (blue boxes), mirroring the official trailer's pacing, and uses rapid one-bar-one-shot cuts during the high-energy chorus (red boxes). IPOT instead assigns one shot per bar throughout, producing a uniform rhythm. BEAT also integrates stylized text cards at low-energy moments (gold boxes) directly into its generation pipeline. In our user study, 26 participants rate trailers from all methods on two Test-8 movies using a five-point Likert scale for rhythm alignment, shot relevance, engagement, and overall quality. BEAT significantly outperforms all five baselines in rhythm and engagement, and MMSC, IPOT, PPBVAM, and CutClaw in overall quality; SSMP scores significantly higher on relevance. Appendix C provides the study design, statistical analysis, and detailed results.

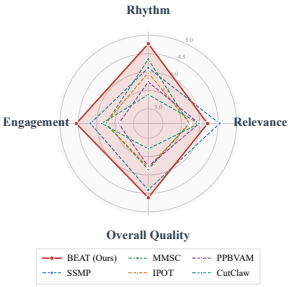


Figure 3: User study.

Table 3: **Ablation study** for MuVA on Test-73. Default settings are marked as gray.

Arch	Objective	FSD↓	AA↑	SDTW↓
SwinT	CLAP	**0.175**	**0.523**	**0.323**
ViT-H/14	ImageBind	0.186	0.490	0.352
ViT-B/32	Mel-CLIP	0.181	0.511	0.334

(a) **Audio encoders.**

$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{InfoNCE}}$	$\mathcal{L}_{\text{Sink}}$	FSD↓	AA↑	SDTW↓
✓	✗	✗	0.191	0.506	0.331
✓	✓	✗	0.181	0.519	0.329
✓	✓	✓	**0.175**	**0.523**	**0.323**

(b) **Loss terms** for Stage 2.

MV	Movie-Trailer	FSD↓	AA↑	SDTW↓
✓	✗	0.227	0.415	**0.312**
✓	✓	**0.175**	0.523	0.323
✗	✓	0.184	**0.539**	0.333

(c) **Training stages.**

CMTD	MMSC	#Pairs	FSD↓	AA↑	SDTW↓
✓	✗	398	0.217	0.452	0.347
✗	✓	472	0.178	0.517	0.324
✓	✓	870	**0.175**	**0.523**	**0.323**

(d) **Training datasets** for Stage 2.

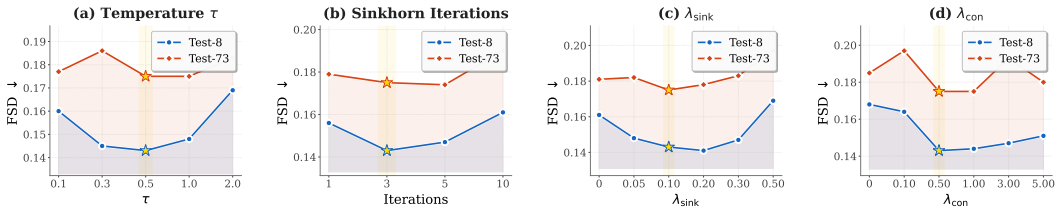


Figure 4: Ablation study of training hyperparameters. Stars mark defaults.

Table 4: **Ablation study** for the module of BEAT on Test-73.

Configuration	F1↑	AA↑	τ↑	VL-Overall↑
MuVA	0.233	0.523	0.049	5.94
+ MusicAnalyzer	0.231	0.532	0.057	6.02
+ VLM Critic	0.215	0.511	0.065	6.14
+ NarrativeGuard	0.195	0.528	0.040	5.79

(a) **Role of each module.**

Configuration	AQ↑	VL-Overall↑
MuVA	4.263	5.94
+ Text Cards	4.310	5.90
+ Transition	4.211	6.03

(b) **Trailer Composition**.

Strategy	F1↑	SF1@5↑	FSD↓	LD↓	AA↑	τ↑	SDTW↓	AQ↑	VL-Overall↑
Greedy	0.233	0.403	0.175	81.1	0.523	0.049	0.323	4.263	5.94
Bar-DP	0.217	0.378	0.180	82.7	0.530	0.054	0.327	4.346	6.23

(c) **Shot selection strategy.**

5.4 Ablation Study

MuVA training (Table 3). We ablate four design choices for training MuVA. **(a)** CLAP outperforms ImageBind Audio and Mel-CLIP across all metrics, likely because its LAION-Audio-630K pretraining provides stronger music-specific representations than general-purpose audio encoders. **(b)** Each loss term contributes incrementally: KL alignment provides the base supervision, InfoNCE anchors the cross-modal embedding space, and Sinkhorn regularization further improves both FSD and AA by preventing degenerate score concentration. **(c)** Two-stage training achieves the best FSD. Skipping V2M pretraining hurts distributional quality; skipping trailer fine-tuning yields poor AA because the model cannot specialize to trailer-specific ordering. **(d)** Combining CMTD and MMSC (870 pairs) outperforms either alone: CMTD-only suffers from limited scale, while MMSC-only achieves competitive FSD but weaker ordering.

Pipeline modules (Table 4). **(a)** Each agent shifts the quality profile in a predictable way. MusicAnalyzer improves τ and VL by providing energy-aware bar segmentation that enables rhythm-matched editing. VLM Critic trades F1 for higher VL by re-selecting shots for coherence. NarrativeGuard reduces F1 because excluding late-movie shots narrows the candidate pool, but is essential for spoiler-free output. **(b)** Text cards improve AQ but slightly reduce VL, while transitions show the opposite trade-off. **(c)** Bar-DP trades F1 for improved VL, AA, and τ: elastic multi-bar spans produce more perceptually appealing trailers, consistent with professional editing practice where long takes are preferred during low-energy passages.

6 Conclusion

We presented BEAT, a music-guided movie trailer generation framework that combines learned bar-shot compatibility, energy-adaptive shot durations, and agent-based creative planning. MuVA and Bar-DP provide explicit control over shot selection and timing, while structured text signals coordinate planning, composition, and iterative critique. The resulting pipeline takes a full-length movie and a chosen soundtrack as input and produces a fully composed trailer. We also introduced TrailerArena for evaluation across four complementary dimensions, with experiments demonstrating leading shot-selection performance on both test sets. BEAT illustrates how trainable audiovisual alignment and explicit editing constraints can support flexible agentic workflows for trailer production.

Reproducibility Statement

Section 3 describes MuVA, its training objectives, the Bar-DP selection algorithm, and the trailer composition pipeline. Section 5 specifies the training data, evaluation splits, baselines, feature encoders, and hardware. Appendix A.2 provides the MuVA architecture and training hyperparameters. Section 4 defines the TrailerArena evaluation framework. Appendix A clarifies the evaluated system configurations and metrics, Appendix B provides additional quantitative results and runtime measurement boundaries, and Appendix C documents the human-study protocol and statistical analysis. To support reproduction of the reported results, we will publicly release the features extracted from movies, trailers, and music. We will also provide links to official trailers to facilitate comparisons with real-world trailers.

Ethics Statement

BEAT supports creative work through music-guided editing of existing movie footage. Responsible use centers on respect for the rights and applicable licenses of movie footage, music, datasets, and pretrained models. Clear identification of automatically assembled trailers and human review before distribution can help preserve the context of source material and support transparent communication with audiences. Human oversight can further support responsible use by helping identify and address potential biases in pretrained models and model-based evaluation, as discussed in Appendix D.

AI Use Statement

AI tools assisted with research ideation and execution, including discussing methodological alternatives, supporting code implementation and debugging, and inspecting experimental records. They also helped organize, revise, and polish the manuscript for clarity and readability. All AI-assisted text and analysis were reviewed by the authors, and every reported number, claim, and citation was verified against our own experimental logs and the cited sources. We take responsibility for the final content of this work.

References

Dawit Mureja Argaw, Mattia Soldan, Alejandro Pardo, Chen Zhao, Fabian Caba Heilbron, Joon Son Chung, and Bernard Ghanem. Towards automated movie trailer generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7445–7454. IEEE, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Roberto Balestri, Pasquale Cascarano, Mirko Degli Esposti, and Guglielmo Pescatore. Trailer reimagined: An innovative, llm-driven, expressive automated movie summary framework (traildreams). *arXiv preprint arXiv:2602.02630*, 2026.

Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, and Ja-Ling Wu. Action movies segmentation and summarization based on tempo analysis. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 251–258, 2004.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Zihan Ding, Xinyi Wang, Junlong Chen, Per Ola Kristensson, and Junxiao Shen. Prompt-driven agentic video editing system: Autonomous comprehension of long-form, story-driven media. *arXiv preprint arXiv:2509.16811*, 2025.

Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian conference on computer vision*, pp. 39–54. Springer, 2018.

Roman Ganhör. Muvee: An alternative approach to mobile video trimming. In *2014 IEEE International Symposium on Multimedia*, pp. 229–236. IEEE, 2014.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Mohammad Hesham, Bishoy Hani, Nour Fouad, and Eslam Amer. Smart trailer: Automatic generation of movie trailer using only subtitles. In *2018 First International Workshop on Deep and Representation Learning (IWDRL)*, pp. 26–30. IEEE, 2018.

Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 839–842, 2010.

Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2023.

Ke Li, Maoliang Li, Jialiang Chen, Jiayu Chen, Zihao Zheng, Shaoqi Wang, and Xiang Chen. Direct: Video mashup creation via hierarchical multi-agent planning and intent-guided editing. *arXiv preprint arXiv:2604.04875*, 2026a.

Yuzhi Li, Haojun Xu, and Feng Tian. From shots to stories: Llm-assisted video editing with unified language representations. *Displays*, 95:103575, 2026b.

Wu-Qin Liu, Min-Xuan Lin, Hai-Bin Huang, Chong-Yang Ma, Yu Song, Wei-Ming Dong, and Chang-Sheng Xu. Emotion-aware music driven movie montage. *Journal of Computer Science and Technology*, 38(3):540–553, 2023.

Xingchen Liu and Jianming Jiang. Semi-supervised learning towards computerized generation of movie trailers. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2990–2995. IEEE, 2015.

Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie summarization via sparse graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13631–13639, 2021.

Marcelo Sandoval-Castañeda, Bryan Russell, Josef Sivic, Gregory Shakhnarovich, and Fabian Caba Heilbron. Editduet: A multi-agent system for video non-linear editing. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Silero Team. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2024.

Alan F Smeaton, Bart Lehane, Noel E O'Connor, Conor Brady, and Gary Craig. Automatically selecting shots for action movie trailers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 231–238, 2006.

John R Smith, Dhiraj Joshi, Benoit Huet, Winston Hsu, and Jozef Cota. Harnessing ai for augmenting creativity: Application to movie trailer creation. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1799–1808, 2017.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11218–11221, 2024.

Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18782–18793. IEEE, 2025.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 699–714, 2024a.

Yutong Wang, Hongteng Xu, and Dixin Luo. Self-supervised video summarization guided by semantic inverse optimal transport. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6611–6622, 2023.

Yutong Wang, Sidan Zhu, Hongteng Xu, and Dixin Luo. An inverse partial optimal transport framework for music-guided trailer generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9739–9748, 2024b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025b.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Hongteng Xu, Yi Zhen, and Hongyuan Zha. Trailer generation via a point process-based visual attractiveness model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2198–2204, 2015.

Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909*, 2025.

Peixuan Zhang, Chang Zhou, Ziyuan Zhang, Hualuo Liu, Chunjie Zhang, Jingqi Liu, Xiaohui Zhou, Xi Chen, Shuchen Weng, Si Li, et al. A benchmark and multi-agent system for instruction-driven cinematic video compilation. In *European Conference on Computer Vision*, pp. 213–230. Springer, 2026.

Shifang Zhao, Yihan Hu, Ying Shan, Yunchao Wei, and Xiaodong Cun. CutClaw: Agentic hours-long video editing via music synchronization. *arXiv preprint arXiv:2603.29664*, 2026.

Hengji Zhou, Lingxuan Huang, Jian Wang, Bing Zhou, Si Wu, Lianghao Xia, and Chao Huang. VideoAgent: All-in-one framework for video understanding and editing. *arXiv preprint arXiv:2606.23327*, 2026.

Sidan Zhu, Yutong Wang, Hongteng Xu, and Dixin Luo. Weakly-supervised movie trailer generation driven by multi-modal semantic consistency. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pp. 10234–10242, 2025.

Sidan Zhu, Hongteng Xu, and Dixin Luo. Self-paced and self-corrective masked prediction for movie trailer generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7684–7694, 2026.

Appendix

CONTENTS

A EXPERIMENTAL PROTOCOL AND METRICS 13
A.1 TASK AND EVALUATION DATA . . . 13
A.2 MUVA ARCHITECTURE AND TRAINING HYPERPARAMETERS . . . 13
A.3 BAR-DP HYPERPARAMETERS . . . 13
A.4 SYSTEM CONFIGURATIONS . . . 14
A.5 EVALUATION METRICS . . . 14
B ADDITIONAL QUANTITATIVE RESULTS 15
B.1 STANDALONE MUSIC-BAR-TO-SHOT RETRIEVAL . . . 16
B.2 SHOT-SELECTION PRECISION AND RECALL . . . 16
B.3 RHYTHMIC SYNCHRONIZATION . . . 16
B.4 COMPUTATIONAL COST . . . 17
C HUMAN EVALUATION 18
C.1 STUDY DESIGN AND PARTICIPANTS . . . 18
C.2 STATISTICAL ANALYSIS . . . 18
C.3 RESULTS . . . 18
D DISCUSSION AND LIMITATIONS 18

A EXPERIMENTAL PROTOCOL AND METRICS

A.1 TASK AND EVALUATION DATA

BEAT addresses music-guided editing: given a source movie and a selected soundtrack, it selects and times shots according to the music's semantics, structure, and dynamics. This setting assumes that music is available before visual editing, while professional workflows may also select or revise music during editing. It complements methods that generate or retrieve music for fixed visuals.

We evaluate on the established Test-8 and Test-73 splits using TrailerArena's metrics and protocols, which also apply to other generated and ground-truth (GT) trailer collections. Reproducing the results requires access to the source movies and trailers; copyright restrictions and unavailable media links may limit this access.

A.2 MUVA ARCHITECTURE AND TRAINING HYPERPARAMETERS

MuVA uses a latent dimension of d=512, L=6 Transformer blocks, L_{ca}=3 cross-attention layers, and 8 attention heads. Stage 1 pretraining runs for 50 epochs with a batch size of 64, an initial learning rate of 3×10^{-4}, and a cosine learning-rate schedule with warm-up. The optimizer parameters are β_1=0.9 and β_2=0.999. Stage 2 fine-tuning uses movie-trailer pairs for 20 epochs, with a batch size of 16 and a learning rate of 5×10^{-5}. We set the InfoNCE and Sinkhorn weights to λ_{con}=0.5 and λ_{sink}=0.1, respectively, and use Sinkhorn temperature τ_s=0.5 with T=3 normalization iterations.

A.3 BAR-DP HYPERPARAMETERS

For Bar-DP (Section 3.3), we use beam width W=50, retain the top M=20 candidate shots per bar, and allow spans of up to $k_{\max}$=5 bars. The smoothness and cut-bonus weights are λ_{sm}=0.3 and λ_{cut}=0.5, respectively. The duration constraint uses η=0.9, allowing a shot to span up to $1/\eta \approx 1.11$ times its original duration through mild slow-motion. The visual similarity threshold for neighbor exclusion is θ_{sim}=0.80. The single-CPU-core runtime reported in Section 3.3 uses these settings with approximately J=60 music bars and I=1500 movie shots.

Table 5: Full metric inventory for Table 1. Arrows indicate the preferred direction; GT indicates whether a reference trailer or its annotations are required.

Dimension	Metric	GT	Measurement
Selection (8)	Precision ↑	Yes	Fraction of selected shots present in GT.
	Recall ↑	Yes	Fraction of GT shots recovered.
	F1 ↑	Yes	Harmonic mean of shot-set precision and recall.
	F1@K ↑	Yes	F1 with hard shot-index tolerance K.
	IoU ↑	Yes	Intersection over union of selected and GT shot sets.
	SoftF1@K ↑	Yes	F1 with linearly decreasing credit for nearby indices.
	Chamfer distance ↓	Yes	Symmetric mean nearest-neighbor shot-index distance.
	FSD ↓	Yes	Distance between selected and GT visual feature distributions.
Ordering (8)	LD ↓	Yes	Edit distance under the overlap-subsequence protocol.
	AA ↑	Yes	Pairwise order agreement among overlapping shots.
	Kendall τ ↑	Yes	Rank correlation between predicted and GT sequences.
	R@1 ↑	Yes	Per-bar GT-shot retrieval success at rank 1.
	R@5 ↑	Yes	Per-bar GT-shot retrieval success within the top 5.
	R@10 ↑	Yes	Per-bar GT-shot retrieval success within the top 10.
	MRR ↑	Yes	Mean reciprocal rank of the GT-aligned shot.
	mAP@t ↑	Yes	Threshold-tolerant per-bar ranking score (see text).
Composition (3)	SDTW ↓	Yes	DTW distance between visual feature trajectories.
	AQ ↑	No	Mean pretrained aesthetic score of selected-shot frames.
	BeatAlign ↑	No	Fraction of music-bar boundaries near a shot transition.
Perceptual (6)	VL-Pacing ↑	No	VLM rating of editing pace.
	VL-Coherence ↑	No	VLM rating of coherence across shots.
	VL-Variety ↑	No	VLM rating of visual variety.
	VL-Engagement ↑	No	VLM rating of viewer engagement.
	VL-Narrative ↑	No	VLM rating of narrative quality.
	VL-Overall ↑	No	Mean of the five perceptual ratings.

A.4 System Configurations

The main comparison in Table 2 evaluates MuVA with greedy, rigid one-to-one bar-to-shot selection. The VLM Critic and generated text cards are disabled. Table 4 separately evaluates elastic selection with Bar-DP and the optional agentic components. These comparisons separate the effects of the alignment model, selection strategy, and optional composition modules.

When enabled, the critic inspects a 1-fps representation for semantic and narrative issues, such as irrelevant or incoherent segments. MuVA and Bar-DP perform fine-grained alignment using audio- and shot-level representations. This sampling rate supports high-level inspection but may miss brief cuts and transition artifacts. Appendix B.3 therefore evaluates rhythmic synchronization separately.

A.5 Evaluation Metrics

Table 5 expands the four dimensions in Table 1 into 25 entries (8 selection, 8 ordering, 3 composition, and 6 perceptual). For this inventory, F1@K, SoftF1@K, and mAP@t each count as one parameterized metric family; R@1, R@5, and R@10 are listed separately. The perceptual count includes the five individual ratings and their mean. Additional threshold variants and diagnostic scores do not add entries to this count.

The ordering dimension includes both sequence-order and per-bar retrieval metrics. The latter require a bar-by-shot score matrix; their protocol is described in Appendix B.1. The evaluation script's mAP@t uses the reciprocal rank of the first shot within t indices of the GT-aligned shot, averaged over bars, with $t \in \{0, 1, 2, 3\}$. It is thus a tolerance-based reciprocal-rank score, rather than conventional

average precision over all relevant shots. SoftF1 is evaluated at $K \in \{1, 3, 5\}$, with $K{=}5$ in the main comparison. BeatAlign uses a ± 0.2 s window around music-bar boundaries; separate detected-event rhythm diagnostics are reported in Appendix B.3.

Seven of the nine metrics in Table 2 use the GT trailer: F1, SoftF1@5, FSD, LD, AA, Kendall τ, and SDTW. AQ and VL-Overall complement these reference-based measurements with visual-attractiveness and model-based perceptual scores.

Shot-set precision and recall. For predicted and GT shot sets $\hat{\mathcal{S}}$ and $\mathcal{S}^*$, respectively,

$$P = \frac{|\hat{\mathcal{S}} \cap \mathcal{S}^*|}{|\hat{\mathcal{S}}|}, \qquad R = \frac{|\hat{\mathcal{S}} \cap \mathcal{S}^*|}{|\mathcal{S}^*|}, \qquad F1 = \frac{2PR}{P + R}. \tag{12}$$

Precision measures the fraction of selected shots present in GT; recall measures the fraction of GT shots recovered. Appendix B.2 reports both quantities alongside the F1 summary used in the main table.

Comparison of shot-selection metrics. Table 6 contrasts the information captured by F1, Chamfer Distance, and FSD. FSD operates on visual features rather than shot indices. Its mean and covariance terms characterize shifts in content and feature variation, complementing exact-overlap and index-proximity measurements. For the synthetic examples, GT shots have indices $\{10, 20\}$ and scalar features $\{0, 2\}$. Shots 9 and 19 have features 0 and 2, respectively, while shots 11 and 21 both have feature 0. The illustrative Chamfer distance is the average of the two directional mean nearest-neighbor distances, using absolute index differences. Thus, $\{9, 19\}$ and $\{11, 21\}$ have identical F1 and index-distance scores, but only the former matches the GT feature distribution. Identical shot sets yield F1$=$1 and FSD$=$0; FSD does not distinguish distributions with identical means and covariances.

Table 6: Shot-selection metrics: properties and controlled synthetic examples. The examples illustrate index overlap, index proximity, and visual-distribution agreement; they are not experimental results.

	F1↑	**Chamfer Dist.↓**	**FSD↓**
Space	Shot indices	Shot indices	Visual features
Granularity	Binary (hit/miss)	NN distance	Distribution (μ+Σ)
Measures	Index overlap	Index proximity	Content + diversity
Synthetic examples: GT indices $\{10, 20\}$*, features* $\{0, 2\}$			
Identical GT shots: $\{10, 20\}$	1	0	0
Matching features: $\{9, 19\}$	0	1	0
Collapsed features: $\{11, 21\}$	0	1	> 0
Half GT, matching features: $\{10, 19\}$	0.5	0.5	0

Aesthetic Quality (AQ). In the selection-only setting, we sample the midpoint frame of each unique selected source-movie shot, extract a CLIP ViT-L/14 representation, and apply a pretrained aesthetic MLP. Trailer-level AQ is the mean frame score. It measures visual attractiveness; ordering, pacing, and music synchronization require separate temporal metrics.

VL-Overall. In the main comparison, VLM ratings are computed after generation and provide no feedback to the system. The critic-enabled ablation uses related VLM judgments during refinement, making VL-Overall partly dependent on the refinement criterion. Appendix D discusses this dependence, and Appendix C provides independent human ratings.

B Additional Quantitative Results

We evaluate local bar-shot retrieval, trailer-level shot selection, and the rhythmic structure of rendered trailers, then report selection and composition runtimes separately.

B.1 Standalone Music-Bar-to-Shot Retrieval

To isolate the alignment model, we evaluate MuVA's bar-by-shot compatibility matrix on Test-73 without Bar-DP, agents, the critic, or the compositor. Each of the 6,415 valid music bars is a query over all shots in its source movie, with an average of approximately 1,381 candidates per movie.

Exact R@K records whether the annotated GT-aligned shot appears among the top K candidates. MRR averages the reciprocal rank $1/r$ of that shot. We average query scores within each trailer, then average over all 73 trailers to give each trailer equal weight. A temporally tolerant R@10 also accepts either immediately adjacent shot (± 1 shot), accounting for ambiguity near editing boundaries. These query-level retrieval metrics evaluate a different property from the shot-set recall defined in Appendix A.

Table 7: Standalone bar-to-shot retrieval on Test-73. All values are proportions; higher is better. Dashes denote unreported values.

	Exact match				± 1 shot
Method	R@1	R@5	R@10	MRR	R@10
Random ranking	0.0007	0.0036	0.0072	–	–
IPOT	0.0019	0.0087	0.0139	0.0085	0.0296
Frozen ImageBind cosine	0.0141	0.0361	0.0510	0.0284	0.0732
MuVA	**0.0318**	**0.0879**	**0.1177**	**0.0635**	**0.1505**

Table 7 shows that MuVA improves every reported retrieval metric over frozen ImageBind cosine similarity and IPOT. Exact R@10 increases from 0.0510 for frozen ImageBind to 0.1177 for MuVA, while MRR increases from 0.0284 to 0.0635. These gains support learning bar-shot compatibility, although absolute recall remains modest in the large candidate pools. Retrieval evaluates local matching; selection and composition must additionally account for trailer-level constraints.

B.2 Shot-Selection Precision and Recall

We next examine the shot sets produced by the main-comparison configuration. Table 8 expands the F1 results in Table 2 into precision and recall for IPOT, MMSC, SSMP, and BEAT on both test splits.

Table 8: Exact shot-set precision, recall, and F1 on both test splits. Higher is better. F1 values match Table 2.

Split	Method	Precision	Recall	F1
Test-8	IPOT	0.122	0.143	0.131
	MMSC	0.130	0.150	0.139
	SSMP	0.156	0.168	0.162
	BEAT	**0.180**	**0.208**	**0.193**
Test-73	IPOT	0.097	0.113	0.104
	MMSC	0.184	0.214	0.197
	SSMP	0.211	0.243	0.226
	BEAT	**0.217**	**0.252**	**0.233**

BEAT achieves the highest precision and recall among these four methods on both splits. The F1 gains reflect both more accurate selections and greater GT-shot coverage. These trailer-level set metrics complement per-bar retrieval, but neither measures shot timing relative to music. The next analysis evaluates this temporal aspect.

B.3 Rhythmic Synchronization

The rigid one-to-one configuration places cuts at predefined music-bar boundaries, giving IPOT, MMSC, SSMP, and BEAT the same BeatAlign score of 1.0. To assess rhythmic structure beyond the

imposed bar grid, we evaluate Test-73 using the official All-In-One beat/downbeat tracker (Kim & Nam, 2023) and independently detected visual cuts.

Reference-free evaluation measures whether cuts occur near tracked beats or downbeats. Reference-based evaluation measures whether generated and GT trailers cut near the same tracked music events. We report beat and downbeat F1 separately, using ± 70 ms as the primary tolerance and ± 200 ms for sensitivity analysis.

Table 9: Rhythmic synchronization on Test-73 using tracked music events and detected visual cuts. All columns report F1; higher is better. Reference-free scores measure cut-to-music alignment, whereas reference-based scores measure agreement with the GT rhythmic editing pattern.

	Reference-free		Reference-based	
Method	Beat	Downbeat	Beat	Downbeat
Primary tolerance: ± 70 *ms*				
IPOT	0.101	0.054	0.095	0.118
MMSC	0.104	0.052	0.080	0.066
SSMP	0.104	0.050	0.085	0.103
MuVA-Greedy	**0.128**	**0.059**	**0.099**	**0.124**
MuVA-BarDP	0.084	0.045	0.065	0.082
BEAT-Full	0.065	0.055	0.073	0.093
Sensitivity analysis: ± 200 *ms*				
IPOT	0.277	0.151	0.245	0.284
MMSC	0.299	0.151	0.232	0.265
SSMP	0.281	0.143	0.223	0.249
MuVA-Greedy	**0.345**	**0.169**	**0.265**	**0.303**
MuVA-BarDP	0.213	0.127	0.180	0.198
BEAT-Full	0.183	0.145	0.166	0.210

The event-based scores distinguish methods that bar-grid BeatAlign treats equally (Table 9). MuVA-Greedy achieves the highest beat and downbeat F1 in both evaluation settings at both tolerances, ahead of MuVA-BarDP and BEAT-Full. Sustained shots reduce cut density and may trade musical-event coverage for visual continuity; elastic pacing therefore does not uniformly improve beat matching. The human rhythm ratings in Appendix C complement these event-level scores by assessing the perceived quality of the editing rhythm.

B.4 Computational Cost

We measure selection time with cached features on one NVIDIA H200 and report median runtimes in Table 10. MuVA-Greedy takes 0.0094 s and MuVA with Bar-DP takes 0.297 s, compared with 0.0005 s for IPOT and 0.0045 s for MMSC. These timings exclude feature extraction and cover selection rather than complete video generation.

Table 10: Median runtime in seconds. Selection uses cached features on one NVIDIA H200. Agentic composition is timed separately and covers a broader set of operations.

Method / timed scope	Median time (s)
Selection only, cached features	
IPOT	0.0005
MMSC	0.0045
MuVA-Greedy	0.0094
MuVA + Bar-DP	0.297
Agentic composition pipeline	115.3

The reported median runtime of the agentic composition pipeline is 115.3 s, dominated by VLM inference, text-card generation, and video rendering. These operations cost substantially more than

shot-index selection. The separate timing scopes quantify this overhead but do not establish an end-to-end speed advantage over the baselines.

C Human Evaluation

The user study complements automatic measurements of shot overlap, alignment, and rhythmic events with viewer assessments of the composed trailers. We report the design, participant-level analysis, and results for the original and expanded cohorts.

C.1 Study Design and Participants

Following the user-study setting of MMSC, we use two Test-8 movies: *300: Rise of an Empire* and *The Hobbit: The Desolation of Smaug*. Participants view trailers generated by BEAT, MMSC, PPBVAM, SSMP, IPOT, and CutClaw, with each movie's official trailer provided as a quality reference. Generated trailers are anonymized, and the method order differs between the two movies to reduce order bias. Every participant rates every generated trailer on a five-point Likert scale using four questions:

- **Rhythm:** How well do the visual cuts align with the rhythm of the music?
- **Engagement:** How engaging is the trailer?
- **Relevance:** How well does the trailer convey the theme and style of the source movie?
- **Overall quality:** How would you rate the trailer's overall quality, including how closely it resembles a real movie trailer?

The study initially included 12 participants, primarily aged 20–30, and was expanded with 14 additional participants: ten aged 20–30, two aged 30–40, and two aged 40–50. Section 5.3 reports the full 26-participant cohort; the two cohorts are nested, not independent.

C.2 Statistical Analysis

For each participant and method, we average the two movie scores. We apply a repeated-measures Friedman test across the six methods, followed by two-sided paired Wilcoxon signed-rank tests with Holm correction across the five BEAT–baseline comparisons within each metric and cohort ($\alpha = 0.05$). Zero differences are excluded; we use exhaustive sign permutations for $N = 12$ and a tie-adjusted normal approximation without continuity correction for $N = 26$. The omnibus test indicates differences among methods; superiority to individual baselines is assessed by the corrected pairwise tests.

C.3 Results

For the original 12 participants, pairwise comparisons favor BEAT primarily in rhythm, engagement, and overall quality, with no significant advantage in relevance. In the expanded cohort, BEAT receives significantly higher rhythm and engagement ratings than all five baselines, and higher overall-quality ratings than MMSC, IPOT, PPBVAM, and CutClaw.

For relevance, BEAT scores significantly above MMSC and IPOT but below SSMP; differences from PPBVAM and CutClaw are not significant. These results support perceptual gains on specific dimensions for the two evaluated movies. Limited movie coverage and participant diversity constrain their generalizability.

D Discussion and Limitations

Rhythmic alignment and perceived editing quality capture complementary aspects of trailer generation. MuVA-Greedy achieves the highest tracked-event F1, while human ratings favor BEAT's rhythm and engagement on the evaluated movies. Frequent cuts can increase musical-event coverage, whereas sustained shots can preserve visual continuity. These results motivate evaluating elastic editing with both temporal metrics and viewer judgments.

The optional modules also affect evaluation dimensions differently (Table 4). Text cards raise AQ from 4.263 to 4.310 but lower VL-Overall from 5.94 to 5.90; the critic raises VL-Overall from 5.94 to 6.14 but reduces GT F1 from 0.233 to 0.215. Neither yields uniform gains. Because the critic and evaluator use related VLM judgments, the perceptual-score increase may partly reflect agreement with the refinement criterion. A dedicated human comparison of critic-enabled and critic-disabled outputs is needed to isolate its benefit.

Several limitations remain. Alignment inherits potential biases from ImageBind and CLAP and may favor salient visual or acoustic content over narratively important but subtle scenes. Generated text cards and transitions may contain artifacts, and selected shots may form incoherent sequences. The critic's limited temporal resolution can leave brief failures undetected.

The two-movie human study and qualitative examples provide limited coverage of possible outputs, while the scale of public movie-trailer datasets constrains broader evaluation. Generalization to documentaries, animation, and non-Western cinema remains underexplored. Evaluating more genres with broader participant populations, while holding movie and music inputs fixed across methods, would strengthen the evidence. Applications to music videos, promotional montages, and highlight reels also remain untested.

Practical deployment depends on access to source media and the VLM and image-generation components. Tagline generation and critique add inference cost, and model inference and rendering dominate composition time. Changes in component availability or behavior can also affect reproducibility. Fast shot selection therefore does not by itself imply low-cost trailer production.