LET RHYTHM TELL YOUR STORY

Anonymous authors
Paper under double-blind review

ABSTRACT

In audio-driven video editing, jointly determining ordered video clips' relative durations and positions on a given music timeline is a cross-modal structured prediction problem with multiple valid solutions and global constraints. We formulate this as constrained boundary-trajectory prediction, jointly specifying relative clip durations and timeline positions. Our primary evaluation uses a given start time. We introduce **Cutalio**, a unified framework for trajectory data construction, modeling, inference, and relational evaluation. Using real edits and independent-start augmentations, we train a multimodal masked model that fuses video, music, and rhythmic evidence. The model jointly predicts unknown boundaries; differentiable dynamic programming computes distributions over feasible paths under configurable outer constraints, including per-clip available source durations, an optional target edit duration, and known cuts, and feeds these distributions back into subsequent predictions. At inference, the framework can generate a full timeline from the given start or complete it around fixed cuts. Progressive reveal, multi-trajectory sampling, and model-based ranking provide test-time scaling. Because absolute cut error against a single reference cannot adequately assess multiple valid arrangements, we introduce pairwise relational metrics for duration ordering, duration ratios, local acoustic changes, and sequence-level acoustic trends. Experiments show that music content and rhythmic cues are complementary, known anchors aid completion, and independent-start augmentation helps primarily under multi-trajectory inference.

1 INTRODUCTION

What to show and *when to show it* are distinct decisions in music-driven video editing (Gonzálbez-Biosca et al., 2025). We study timing: given selected, ordered clips and a full music track, predict a start S_0, internal cuts, and an endpoint END, completing any unknown boundaries around known ones. This trajectory determines each clip's playback duration, relative durations across clips, and where the visuals fall on the music timeline (Pearlman, 2025). We neither reselect nor reorder clips, and the edit need not fill the soundtrack. Figure 1 illustrates the task with ordered chase shots.[1]

Cuts are coupled: moving one changes neighboring durations and the music covered. Feasible paths respect clip order and source availability, yet many edits can be valid. We learn relative durations and their musical response from real arrangements (Swenberg & Carlgren, 2021); source availability and optional target edit duration remain outer constraints.

Editing is constrained completion, not chronological continuation: users may fix an internal cut or endpoint before allocating surrounding shots. Causal attention in standard autoregressive LLMs blocks access to later output tokens. Our masked boundary slots exchange representations bidirectionally, using known anchors and planning feedback to coordinate durations without accessing hidden reference times (Ghazvininejad et al., 2019; Janner et al., 2022).

Prior work selects or sequences shots (Wang et al., 2024; Zhu et al., 2025a;b) or assigns shots to predefined musical units (Wang et al., 2026). Agentic editors plan and revise broader workflows (Sandoval-Castaneda et al., 2025; Li et al., 2026; Zhao et al., 2026; Zhou et al., 2026; Lin et al., 2026). We learn boundary-level planning for fixed content and order, not merely history use.

[1] Interactive demo: `https://cutalio-demo.pages.dev/`.

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.

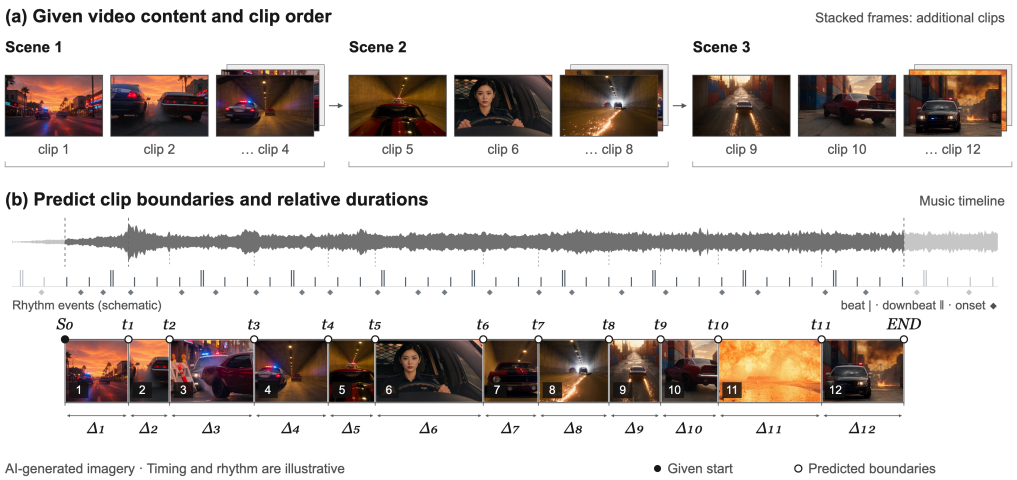

Figure 1: **Learning when to cut, given what to show.** (a) Ordered clips; stacks denote omitted clips. (b) Given S_0, predict cuts and END. Beat/downbeat/onset markers are rhythmic cues, not mandatory cuts. AI-generated imagery; schematic timing/events, not model or detector outputs.

We introduce **Cutalio**, trained on finished edits and independent-start trajectories. Its multimodal masked model scores boundaries from video, music, and rhythm; differentiable DP converts scores into feasible-path distributions under the outer constraints and feeds them into subsequent predictions. Progressive reveal, multi-trajectory sampling, and model ranking provide test-time scaling. The same model generates from a given start or completes around internal anchors; main evaluation gives S_0, with extra anchors studied separately.

Cut error against one reference cannot fully assess valid alternatives. We add pairwise duration order/ratio and local/sequence-level acoustic correspondence metrics. Experiments examine modalities, feedback, anchors, and inference budget; validation finds independent-start augmentation helps mainly under multi-trajectory inference, not the two-pass baseline.

Our contributions are:

1. **Task and constraints:** boundary-trajectory prediction for fixed content, order, and music, separating learned relative temporal relations from outer production constraints.
2. **Joint prediction and planning:** multimodal masked prediction trained with differentiable DP feedback, supporting full trajectories, conditional completion from anchors, and adjustable test-time scaling.
3. **Data and evaluation:** reference and independent-start augmented trajectories, evaluated by pairwise duration and acoustic relations rather than absolute timestamp error alone.

2 Related Work

Music-driven arrangement. IPOT selects and orders shots through music-guided inverse partial transport (Wang et al., 2024); MMSC uses semantic and emotion scores (Zhu et al., 2025a). SSMP completes masked shot indices before post-hoc music alignment (Zhu et al., 2025b), while BEAT assigns shots to consecutive bars, allowing variable spans but cuts only at bar boundaries (Wang et al., 2026). Other work localizes music intervals (Xin et al., 2025), edits concerts (González-Biosca et al., 2025), or aligns motion to beats (Zhang et al., 2025). These methods focus on selection, ordering, or predefined music units; we jointly allocate durations and boundaries for ordered clips.

Agentic editors plan and revise broader workflows: EditDuet uses an Editor–Critic loop (Sandoval-Castaneda et al., 2025), DIRECT coordinates creative roles (Li et al., 2026), CutClaw plans from musical structure (Zhao et al., 2026), and VideoAgent delegates editing subtasks (Zhou et al., 2026). GLANCE passes completed subtimelines to later tasks and resolves cross-segment conflicts (Lin

et al., 2026). Unlike such agent-level feedback, ours returns feasible-path distributions to a trained boundary scorer. We fine-tune Qwen3-Omni (Xu et al., 2025) for direct timestamp prediction without an agent wrapper. Autoregressive models support infilling (Bavarian et al., 2022); our distinction is native constrained completion, not access to supplied future anchors or full music.

Masked prediction and test-time scaling. Mask-Predict and MaskGIT complete partial observations (Ghazvininejad et al., 2019; Chang et al., 2022); MDLM and LLaDA use masked diffusion (Sahoo et al., 2024; Nie et al., 2025), and Diffuser inpaints trajectories with known states or goals (Janner et al., 2022). SSMP already progressively reveals shot indices (Zhu et al., 2025b); masking itself is not new. Our discrete boundaries add monotonicity and duration constraints. Sampling, search, and verification allocate compute differently (Ma et al., 2025): ReMDM remasks tokens (Wang et al., 2025), MEDAL searches reveal paths (Huang et al., 2026), PG-DLM refines denoising histories (Dang et al., 2025), and Fast-dLLM v2 combines block diffusion with caching (Wu et al., 2026). We report feedback passes, reveal depth, and trajectory count by forward calls and runtime, ranking only by model scores.

Differentiable planning. Soft-DTW, smoothed DP, and structured attention connect local scores to global inference (Cuturi & Blondel, 2017; Mensch & Blondel, 2018; Kim et al., 2017); Neural A* learns search guidance (Yonetani et al., 2021). IPOT learns structured allocation through bilevel optimization; Bar-DP searches a precomputed table using selection history and durations. We jointly train path-distribution feedback into subsequent boundary scores, not merely use history.

3 Data Construction

We extract complete boundary trajectories from music-driven edits, preserving clip order and full music context (Figure 2).

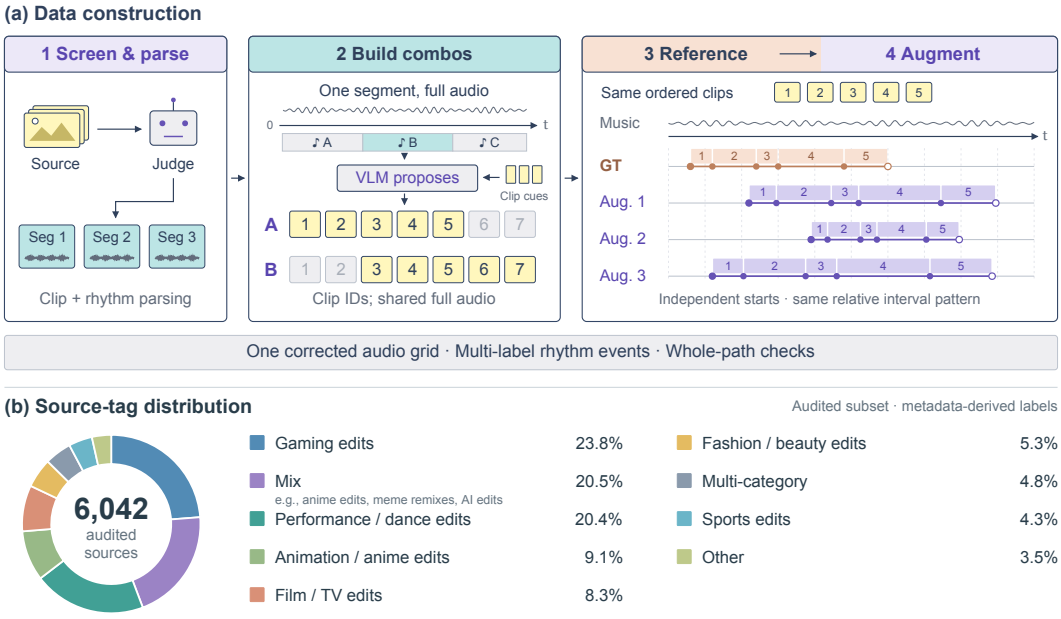

Figure 2: Data construction and audited source tags. (a) VLM-proposed combos share segment audio; GT and independent-start paths retain clip order. Timings are schematic, not exact-ratio guarantees. (b) Unique sources are sorted by share; tags below 2% form Other. Mix means unmatched metadata; Multi-category matches multiple tag groups.

From source videos to combos. After deduplicating public source videos, we use Qwen3.5-Omni (Qwen Team, 2026) to locate one or more beat-synchronous editing intervals per video. Each retained interval defines a *segment*, screened for music clarity, cut–music correspondence, and audio-

visual quality. TransNet V2 (Souček & Lokoč, 2020) splits segments into *clips*; madmom (Böck et al., 2016) extracts rhythm events. An offline language-model planner proposes ordered clip combinations (*combos*) from visual descriptions, musical structure, and editing intent. Each segment's full audio is shared by potentially overlapping combos. Planning text is not a model input. References are automatically parsed and corrected, not exclusively human-annotated.

References and independent-start augmentation. Each N-clip combo has $N + 1$ boundaries (S_0, cuts, END), whose differences give playback durations. Preserving clips and order, augmentation searches the full music timeline for alternative starts, ends, and intervals. Acoustic/rhythmic evidence and All-In-One structure (Kim & Nam, 2023) guide duration-order, adjacent-trend, and rhythmic-density constraints. These rule-constrained paths are not human-verified optimal edits; discrete snapping need not preserve exact ratios. Beat-coordinate duration ordering is checked for pairs above a discrimination threshold. We compare training with and without independent-start augmentation in Section 6.4; alternative controls are retained in Appendix D.1.

Grid alignment. Complete paths map to a physically time-corrected audio-token grid; shared cells retain beat, downbeat, and onset attributes. We reject out-of-range, non-increasing, colliding, or duplicate paths without reference-only candidates or mixed trajectories.

Coverage and availability. Figure 2 covers 6,042 unique sources in an audited subset, not the full collection. Platform tags are weak, unverified labels, not evidence of general-video coverage (Appendix A). We plan to release model and evaluation code, but not media or sample-level data; splits, duration budgets, and mixtures are specified in the experimental protocol.

4 METHOD

Cutalio combines masked multimodal boundary scoring with differentiable DP feedback between predictions (Figure 3).

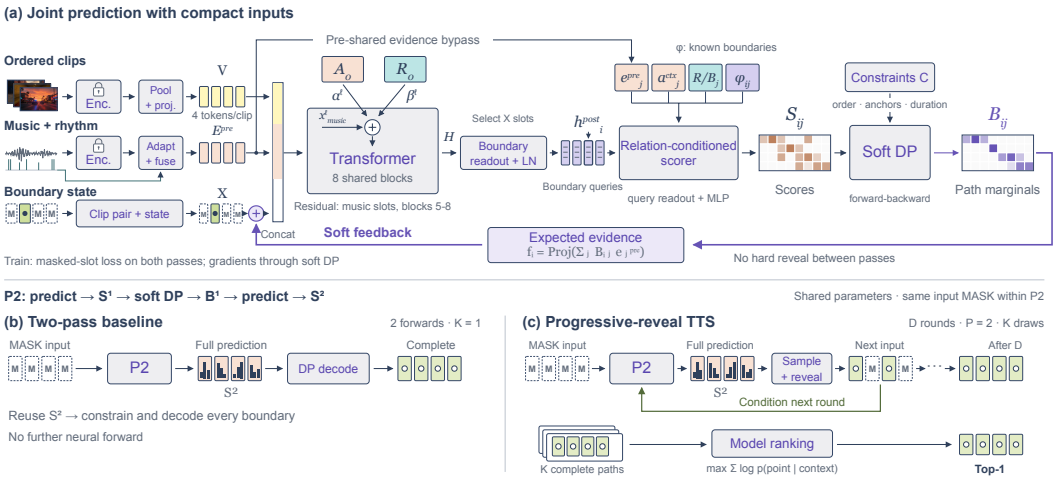

Figure 3: **Cutalio: inputs, planning feedback, and decoding.** (a) X-slot readouts score boundaries; soft DP beliefs feed back. (b) The two-pass baseline (P2) keeps the input MASK; (c) TTS reveals boundaries and model-ranks K paths. M: unknown; bars: predictions; circles: selected; locks: frozen encoders (adapters train). Frames and events are illustrative.

4.1 TRAJECTORY REPRESENTATION AND SHARED BACKBONE

For ordered clips $V = (v_1, \ldots, v_n)$ and full music A, predict $\mathbf{t} = (S_0, t_1, \ldots, t_n = \text{END})$ on the ordered audio grid $\{\tau_j\}_{j=1}^{J}$: $t_i = \tau_{z_i}$ and playback duration $\Delta_i = t_i - t_{i-1}$. Known boundaries $\mathcal{O}$ and constraints $\mathcal{C}$ define feasible paths $\Omega(\mathcal{O}, \mathcal{C})$ enforcing order, music availability, anchors, and duration limits. Duration metadata constrains DP, not embeddings. Main evaluation gives only S_0.

Inputs. Each clip yields four pooled relative-position tokens with order embeddings, but no source duration or absolute timestamps. Time-encoded audio attention gives a_j^{ctx}; rhythm fusion gives e_j^{pre}, which covers all grid candidates and enters backbone and scorer. Boundary tokens combine adjacent clips, slot, observed location or MASK, and DP feedback, never hidden reference locations.

Shared interaction and residual anchors. Eight shared blocks encode the concatenated tokens. Boundary slots attend non-causally to valid context, producing h_i^{post} in parallel; video/music cannot attend back. Before blocks $\ell \in \{5, 6, 7, 8\}$, music slots receive the initial audio projection A_0 and rhythm embedding R_0:

$$\widetilde{x}^{\ell}_{\mathrm{music}} = x^{\ell}_{\mathrm{music}} + \alpha^{\ell} A_0 + \beta^{\ell} R_0. \tag{1}$$

Scalars start at zero; anchors are differentiable but fixed across blocks. R_0 differs from the scoring descriptor below (Appendix B).

4.2 Relation-conditioned evidence scoring

Each boundary–candidate pair (i, j) has four evidence tokens $\mathcal{E}_{ij}$: fused candidate e_j^{pre}, contextual audio a_j^{ctx}, local rhythm $(R/B)_j$, and relative allocation ϕ_{ij}.

Candidate-local rhythm $(R/B)_j$. Map madmom beat/downbeat/onset detections (Böck et al., 2016) to overlapping labels $B_j, D_j, O_j \in \{0, 1\}$. Interpolating beat indices gives $b(t)$ in beats, not seconds. At τ_j, beat phase is $u_j = b(\tau_j) \bmod 1$; bar phase is $v_j = ((b(\tau_j) - b(d_j))/M) \bmod 1$, for preceding downbeat d_j and estimated beats per bar M. Sine/cosine pairs encode periodic continuity; q_j indicates available bar information:

$$\begin{aligned} \mathbf{s}_j &= [\sin(2\pi u_j), \cos(2\pi u_j), q_j \sin(2\pi v_j), q_j \cos(2\pi v_j), B_j, D_j, O_j, q_j], \\ \mathbf{c}_j &= [D_j, B_j, O_j \, \mathbf{1}\{\min_{a \in \{1/4, 1/2, 3/4\}} |u_j - a| \le 0.10\}, O_j], \\ (R/B)_j &= [\mathbf{s}_j; \mathbf{c}_j] \in \mathbb{R}^{12}. \end{aligned} \tag{2}$$

Here $\mathbf{1}$ is an indicator. Channels mark downbeats, beats, onsets near quarter-beat subdivisions, and all onsets; they may overlap, with $D_j = 1 \Rightarrow B_j = 1$. Missing bar information sets $q_j = 0$. $(R/B)_j$ stays fixed during generation.

Anchor-relative allocation ϕ_{ij}. For slot i, let t_L, t_R be the nearest known left/right boundaries and n_L, n_R the intervening clip counts, known from slot order even with masked times. Choosing τ_j gives

$$d^{\mathrm{l}}_{ij} = \frac{b(\tau_j) - b(t_L)}{n_L}, \qquad d^{\mathrm{r}}_{ij} = \frac{b(t_R) - b(\tau_j)}{n_R}, \qquad G_{ij} = \frac{\kappa(\tau_j) - \kappa(t_L)}{n_L}, \tag{3}$$

Here $\kappa(t)$ is the bar index: d measures available *beats per clip* and G bar boundaries per clip, not enforced equal durations. S_0 supplies the left anchor when needed; m_i flags a known right anchor. Scaled/log-compressed d^{l}_{ij}, scaled d^{r}_{ij}, their log ratio, G_{ij}, and m_i form $\phi_{ij} \in \mathbb{R}^6$ (Appendix B). Unlike $(R/B)_j$, it changes with revealed boundaries, never hidden references.

Query-conditioned readout. The query from h_i^{post} attends to separately normalized/projected keys and values for the four evidence types, yielding read_{ij}. An MLP scores their normalized concatenation $[h_i^{\mathrm{post}}; e_j^{\mathrm{pre}}; \mathrm{read}_{ij}]$ (Appendix B). Softmax mixes *types, not candidate times*; invalid candidates are masked.

4.3 Differentiable planning and masked training

Scores $S = [s_{ij}]$ define a distribution over feasible paths, not independent rows (Mensch & Blondel, 2018; Kim et al., 2017):

$$p_\theta(\mathbf{z} \mid V, A, \mathcal{O}, \mathcal{C}) \propto \exp\left(T^{-1} \sum_i s_{i, z_i}\right), \qquad \mathbf{z} \in \Omega(\mathcal{O}, \mathcal{C}). \tag{4}$$

At temperature T, forward–backward DP yields marginals $B_{ij} = p_\theta(z_i = j \mid V, A, \mathcal{O}, \mathcal{C})$. Their expected candidate representation is fed into the next boundary-state input:

$$f_i = \mathrm{Proj}_{\mathrm{fb}}\left(\sum_j B_{ij} e_j^{\mathrm{pre}}\right), \qquad x_i^{(2)} = x_i^{\mathrm{base}} + \lambda_{\mathrm{fb}} f_i. \tag{5}$$

Passes share parameters; the second loss backpropagates through soft DP into the first, but not through grid construction, hard rules, or discrete decoding.

Training objective. We train on partial reveals and all-MASK inputs (Ghazvininejad et al., 2019; Chang et al., 2022); both passes keep the same mask, changing only feedback. The objective averages masked-slot cross-entropy and weighted auxiliary losses (Appendix B). Adapters, trajectory network, and feedback train jointly on frozen features, without path NLL or metric optimization. GT-plus adds complete augmentation views without mixing paths (Section 6.4).

4.4 Decoding, conditional completion, and test-time scaling

Two-pass baseline (P2). Predict from masked boundaries, compute soft DP feedback, then predict again (Figure 3b). The input mask stays fixed: predicted cuts do not become known anchors between passes. Reuse second-pass scores to commit positions in confidence order and update DP beliefs, without extra neural passes or per-boundary argmax. Anchors condition both network and DP.

Progressive-reveal test-time scaling. TTS repeats this prediction-and-feedback module ($P = 2$ here) over D reveal rounds, feeding sampled boundaries into later rounds to generate K paths (Figure 3c). Select the valid path with highest summed conditional log probability under the DP-guided sampler. *All metrics use this path; neither GT nor evaluation scores rank candidates.* Nominal D, P, K are not actual costs, nor a guarantee of monotonic quality gains.

5 Evaluation Framework

We evaluate *relative time allocation* and *acoustic change correspondence*, which absolute cut errors alone miss. JointH4 summarizes the model-selected path.

5.1 Relative time allocation

For reference grid times g_i and predicted times p_i at internal slots $i = 1, \ldots, M$, set $g_0 = p_0 = S_0$ and $\Delta_i^x = b(x_i) - b(x_{i-1})$ for $x \in \{g, p\}$, using the beat coordinate in Section 4.2. All four metrics exclude END; its error is separate.

RelOrder (↑). Duration-order agreement uses all reference-distinguishable interval pairs:

$$\mathcal{P} = \{(i,j) : 1 \le i < j \le M, |\log_2(\Delta_i^g / \Delta_j^g)| \ge 0.25\},$$
$$\mathrm{RelOrder} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbf{1}\{(\Delta_i^p - \Delta_j^p)(\Delta_i^g - \Delta_j^g) > 0\}. \tag{6}$$

The threshold is in log-ratio units, not beats. Predicted ties are errors; an empty reference pair set makes the score undefined.

RelRatio (↓). We compare only *adjacent* interval ratios, without thresholding:

$$\mathrm{RelRatio} = \frac{1}{M-1} \sum_{i=1}^{M-1} \left| \log_2 \frac{\Delta_{i+1}^p}{\Delta_i^p} - \log_2 \frac{\Delta_{i+1}^g}{\Delta_i^g} \right|. \tag{7}$$

Both metrics require $M \ge 2$ and are invariant to uniform positive scaling of predicted beat spans.

5.2 ACOUSTIC CHANGE CORRESPONDENCE

Here g_i uses original reference seconds. Independently extracted log energy, log-mel shape, and chroma give normalized, Gaussian-smoothed features $a_{f,\sigma}(t)$ (Appendix C). Without offset search or warping, compare direction and magnitude by $C(r, v) = 2\langle r, v\rangle/(\|r\|^2 + \|v\|^2) \in [-1, 1]$.

LocalTrend (↑). With $a = a_{f,\sigma}$ and $Q = \{-2, -1, 0, 1, 2\}$, center observations *separately around each cut*:

$$r_{iq} = a(g_i + q\sigma) - \tfrac{1}{5}\sum_{u\in Q} a(g_i + u\sigma),$$
$$v_{iq} = a(p_i + q\sigma) - \tfrac{1}{5}\sum_{u\in Q} a(p_i + u\sigma),$$
$$\mathrm{LocalTrend}_{f,\sigma} = C(\mathrm{vec}[r_{iq}], \mathrm{vec}[v_{iq}]). \tag{8}$$

Vectors concatenate corresponding cuts, offsets, and features: pool dot products and energies before division, not per-cut scores. One cut suffices.

SeqTrend (↑): across-cut changes. Sample once per cut and center *across the cut sequence*:

$$r_i = a(g_i) - \tfrac{1}{M}\sum_j a(g_j), \qquad v_i = a(p_i) - \tfrac{1}{M}\sum_j a(p_j),$$
$$\mathrm{SeqTrend}_{f,\sigma} = C(\mathrm{vec}[r_i], \mathrm{vec}[v_i]). \tag{9}$$

Two cuts are required. Both metrics average valid scales per family, then valid families equally. Eligibility uses reference variation only; flat predictions remain eligible (Appendix C).

JointH4 and reporting. Map LocalTrend and SeqTrend by $(x+1)/2$, RelRatio by 2^{-x}, and leave RelOrder unchanged; take the harmonic mean per combo, then macro-average (Appendix C). Zero components and invalid predictions give JointH4 zero on fixed reference-eligible supports. Labeled *common-valid* comparisons instead condition on all systems' valid outputs and disclose validity separately. These reference-correspondence scores do not measure human preference or aesthetics.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Data and constraints. We use 16,341 training combos from 5,827 sources and 343 validation combos from 215 disjoint sources. Inputs are music, ordered clips, and GT grid S_0; internal GT cuts appear only in the anchor study. DP uses a uniform 5-second *virtual source-clip budget*, not sample-specific source durations. This constrains playback, not music length or a fixed output duration; END is predicted rather than fixed to the GT or music endpoint.

Training and selection. Our 32.09M-trainable-parameter network (configuration: Appendix B) trains from scratch for 20 epochs with four GPUs, 32 original examples/GPU (global batch 128), seed 42, and P2: full masking with probability 0.1, random masking otherwise. GT accompanies enabled augmentation views (Section 6.4). Two-pass-baseline validation JointH4 selects checkpoints before TTS (Table 3).

Foundation-model baseline. We fine-tune Qwen3-Omni-30B-A3B-Instruct on GT plus independent starts, using four uniformly sampled frames per full clip in training and inference. Full music, rhythm-event times, S_0, and planning limits condition timestamp generation without DP repair or snapping. Parameters, pretraining, input tokens, and compute are unmatched (Appendix D.3).

Evaluation protocol. Table 1 uses common-valid samples; within-model analyses retain supports 343/324/302/324/302. Our larger-K runs reuse nested candidates; only model-ranked top-1 is evaluated, never GT-ranked (Section 4.4). The two-pass baseline starts without feedback, TTS with zero feedback: they differ beyond sampling budget. These are development *validation* results, with

one training seed and sampling-seed family unless stated. Source-cluster intervals are descriptive without multiplicity correction; free-S_0 prediction is not evaluated.

6.2 Main results and test-time scaling

Table 1: **Main comparison and complete TTS sweep.** Metric supports are 322, 304, 282, 304, 282 on common-valid samples. Cutalio uses GT + independent starts and P2. Calls: mean trajectory-network predictions over all K draws on 322 examples, not latency or FLOPs. Qwen's autoregressive calls are not comparable. Geometric validity on all 343: Qwen 93.9%, Cutalio 100% (planning violations: Appendix D.3).

Model / inference	Calls	Local↑	Seq↑	Order↑	Ratio↓	JointH4↑
Qwen3-Omni SFT (4 frames)	–	0.2526	0.3572	**0.7405**	**0.6813**	**0.6292**
Cutalio: Two-pass baseline	2.00	0.2815	0.4299	0.6588	0.9715	0.5829
D4/P2/K1	7.65	0.2307	0.3907	0.6764	0.9057	0.5870
D4/P2/K2	15.30	0.2531	0.3957	0.6801	0.8610	0.5931
D4/P2/K4	30.61	0.2657	0.4058	0.6961	0.8091	0.6081
D4/P2/K8	61.22	0.2506	0.4119	0.6793	0.7992	0.5985
D4/P2/K16	122.43	0.2783	**0.4375**	0.6823	0.7932	0.6051
D8/P2/K1	12.83	0.2254	0.3759	0.6917	0.8436	0.5900
D8/P2/K2	25.65	0.2530	0.3808	0.6788	0.8244	0.5910
D8/P2/K4	51.30	0.2636	0.3984	0.6887	0.8103	0.6026
D8/P2/K8	102.61	0.2580	0.4079	0.6740	0.7790	0.5971
D8/P2/K16	205.22	**0.2842**	0.4249	0.6858	0.7533	0.6080

Foundation-model comparison. Qwen-SFT leads on relative duration allocation and JointH4; two-pass Cutalio and both displayed K16 settings exceed it on both acoustic metrics. D4/P2/K4 reduces the JointH4 gap from 0.0463 to 0.0211. Paired 95% intervals for the two previously reported TTS comparators cross zero, which does not establish equivalence (Appendix D.3). These conditional-quality results do not remove the input and resource differences in Section 6.1.

Sampling breadth. All ten TTS settings exceed baseline JointH4 on this common support. At D4, increasing K from 1 to 4 raises JointH4 from 0.5870 to 0.6081, but K8 falls to 0.5985 and K16 only recovers to 0.6051. Ratio error decreases with K at both depths, while duration ordering and acoustic correspondence fluctuate. Nested draws and model-only ranking mean more candidates do not guarantee a better evaluated edit; separating generation from ranking requires further controls. Appendix E shows individual gains and failures.

Depth and cost. At K4, D8 uses 51.30 mean predictions versus D4's 30.61, yet has lower JointH4 (0.6026 versus 0.6081). D8/K16 reaches 0.6080 with about $6.7\times$ as many predictions as D4/K4: nearly unchanged JointH4 but improved Ratio and Local. Seq instead peaks at D4/K16. Thus no budget wins on every metric, and extra depth is not uniformly beneficial. The unchanged full-cohort sweep remains in Appendix D.2; neither table identifies an untouched-test optimum.

6.3 Modality and planning controls

We independently train V+A and V+R on the full model's GT+independent-start contract (Table 2). Both retain video, grid, timing, DP, and supervision. V+A zeros neural rhythm/relative inputs; V+R zeros neural audio content. V+A is not a rhythm-free system.

V+A+R achieves the highest observed JointH4 across these three variants' complete sweeps (0.6075, D8/K16). Baseline comparisons show rhythm improves JointH4 over V+A, while audio improves both acoustic trends over V+R. All gain at D8/K16, but paired JointH4 gap intervals cross zero (Appendix D.4). TTS gains vary across historical relation/fusion controls (Appendix D).

Table 2: **Independent modality training: JointH4** ($n = 302$). P2 and equal inference budgets; baseline validation selects checkpoints before TTS.

Neural inputs	Baseline	D4/K4	D8/K16
Cutalio: V+A+R	**0.5809**	**0.6054**	**0.6075**
V+A	0.5513	0.5792	0.6018
V+R	0.5779	0.5895	0.5997

6.4 Data augmentation ablations

We retain GT and add one complete independent-start path per example when available: $\mathcal{L} = \mathcal{L}_{\text{GT}} + \mathcal{L}_{\text{independent}}$. Each view uses its own S_0; validation always uses GT grid S_0. Appendix D.1 preserves the full four-way controls.

Table 3: **Data augmentation: JointH4.** P2 with matched original-example updates, not FLOPs; checkpoints selected by baseline validation before TTS.

Training data	Baseline	D4/P2/K4	D8/P2/K16
GT only	**0.5896**	0.5984	0.5844
GT + independent start	0.5809	**0.6054**	**0.6075**

Augmentation helps TTS, not the two-pass baseline. Independent-start augmentation trails GT-only under the baseline but leads in JointH4 at all ten TTS settings. At D8/P2/K16, it gains 0.0267 over its baseline score versus -0.0052 for GT-only; its advantage over GT-only is 0.0231 (paired 95% interval $[0.0023, 0.0448]$; Table 3). This supports multi-trajectory supervision as complementary to TTS, not necessary: GT-only also benefits at smaller budgets. Generation and ranking effects remain unisolated (Appendix D.1).

6.5 Conditional completion with known anchors

Historical Cutalio uses S_0 plus one/all GT-grid script anchors (*oracle completion*), P2/K1, and identical decoding. Network+DP conditions both forwards and all DP; DP-only constrains final DP with hash-verified baseline-identical logits.

Table 4: **Oracle-anchor completion.** Local: withheld LocalTrend; MAE/END: mean absolute raw-time cut/endpoint errors (s); Grid exact: GT-grid match fraction (definitions/rationale: Appendix C). All script-anchor slots are excluded across settings; common-valid cuts/END: 30/31 of 32 combos.

Conditioning	Local↑	MAE (s)↓	Grid exact↑	END (s)↓
Two-pass, no internal anchors	0.1670	1.8333	0.1069	3.5386
DP-only, one anchor	0.1984	1.7232	0.1303	3.3463
Network+DP, one anchor	0.2016	1.2260	0.1417	1.6875
DP-only, all anchors	0.1685	1.4209	0.1220	3.1141
Network+DP, all anchors	**0.3753**	**0.6924**	**0.2402**	**1.0613**

Network+DP improves withheld cuts and END beyond DP-only, without isolating individual components. Common-valid results need failure audits and larger-cohort validation.

7 Conclusion

Cutalio combines masked boundary prediction, differentiable planning feedback, and relational evaluation for music-driven arrangement. Validation supports anchor completion and complementarity between independent-start augmentation and test-time scaling, despite lower two-pass-baseline JointH4. Independent evaluation, matched-compute controls, and human assessment are needed.

AI USE STATEMENT

We used generative AI tools, including Codex, to assist with discussions of the method and evaluation framework. Following research plans proposed by the authors, these tools assisted with method implementation, debugging, and the organization of experimental results. We also used them for manuscript translation and language editing, literature searches, figure preparation, and demonstration-page development. The uses of vision-language and multimodal models within the research pipeline, including data construction and baseline experiments, are described in the corresponding sections. The research plans and final conclusions were determined by the authors, who take responsibility for the final manuscript, code, experimental results, and research conclusions.

ETHICS STATEMENT

This study examines edit timing for video clips in a fixed order. Source videos are drawn from publicly accessible samples on the Internet, but we will not release source media files or sample-level data. Automated screening by multimodal models and metadata-based categorization may introduce selection bias; our data collection should not be regarded as representative of all video-editing domains. All AI-generated images used for illustration are clearly labeled. Automatically adjusting edit timing may change viewers' interpretation of the footage; downstream use should respect the original context of the material, privacy, and applicable permission requirements.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe data construction and audit notes. Section 4 and Appendix B specify the model and inference procedures; Section 5 and Appendix C define the metrics. Training settings, checkpoint selection, and inference budgets are reported in Section 6, with additional analyses in Appendix D. We plan to release the data-processing workflow together with model and evaluation code, but not the underlying videos, audio, or sample-level data. This supports inspection and reuse of the pipeline on separately obtained data; exact reproduction of our reported scores remains limited by access to the original dataset.

REFERENCES

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022. doi: 10.48550/arXiv.2207.14255. URL https://arxiv.org/abs/2207.14255.

Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. madmom: a new Python audio and music signal processing library. In *Proceedings of the 24th ACM International Conference on Multimedia*, 2016. URL https://arxiv.org/abs/1605.07008.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chang_MaskGIT_Masked_Generative_Image_Transformer_CVPR_2022_paper.html.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903, 2017. URL https://proceedings.mlr.press/v70/cuturi17a.html.

Meihua Dang, Jiaqi Han, Minkai Xu, Kai Xu, Akash Srivastava, and Stefano Ermon. Inference-time scaling of diffusion language models via trajectory refinement. *arXiv preprint arXiv:2507.08390*, 2025. doi: 10.48550/arXiv.2507.08390. URL https://arxiv.org/abs/2507.08390v5. Version 5, revised August 2026.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-Predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pp. 6112–6121, 2019. doi: 10.18653/v1/D19-1633. URL https://aclanthology.org/D19-1633/.

Daniel Gonzálbez-Biosca, Josep Cabacas-Maso, Carles Ventura, and Ismael Benito-Altamirano. When and how to cut classical concerts? a multimodal automated video editing approach. In *Proceedings of the 3rd International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, 2025. doi: 10.1145/3746278.3759387. URL https://arxiv.org/abs/2510.05661.

Zheng Huang, Kiran Ramnath, Yueyan Chen, Aosong Feng, Sangmin Woo, Balasubramaniam Srinivasan, Zhichao Xu, Kang Zhou, Shuai Wang, Haibo Ding, and Lin Lee Cheong. Diffusion language model inference with Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3493–3512, 2026. doi: 10.18653/v1/2026.findings-eacl.180. URL https://aclanthology.org/2026.findings-eacl.180/.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Taejun Kim and Juhan Nam. All-In-One metrical and functional structure analysis with neighborhood attentions on demixed audio. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023. URL https://arxiv.org/abs/2307.16425.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1702.00887.

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*, 2026. doi: 10.48550/arXiv.2604.04875. URL https://arxiv.org/abs/2604.04875v1.

Zihao Lin, Haibo Wang, Zhiyang Xu, Siyao Dai, Huanjie Dong, Xiaohan Wang, Yolo Y. Tang, Yixin Wang, Qifan Wang, and Lifu Huang. GLANCE: A global-local coordination multi-agent framework for music-grounded non-linear video editing. *arXiv preprint arXiv:2604.05076*, 2026. doi: 10.48550/arXiv.2604.05076. URL https://arxiv.org/abs/2604.05076v1.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2523–2534, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_Scaling_Inference_Time_Compute_for_Diffusion_Models_CVPR_2025_paper.html.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3462–3471, 2018. URL https://proceedings.mlr.press/v80/mensch18a.html.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html.

Karen Pearlman. *Cutting Rhythms: Creative Film Editing*. Routledge, 3 edition, 2025. ISBN 9781041024088. URL https://www.routledge.com/Cutting-Rhythms-Creative-Film-Editing/Pearlman/p/book/9781041024088.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. doi: 10.48550/arXiv.2604.15804. URL https://arxiv.org/abs/2604.15804v2.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.07524.

Marcelo Sandoval-Castaneda, Bryan Russell, Josef Sivic, Gregory Shakhnarovich, and Fabian Caba Heilbron. EditDuet: A multi-agent system for video non-linear editing. In *SIGGRAPH Conference Papers '25*, 2025. doi: 10.1145/3721238.3730761. URL https://arxiv.org/abs/2509.10761.

Tomáš Souček and Jakub Lokoč. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. URL https://arxiv.org/abs/2008.04838.

Thorbjörn Swenberg and Simon Carlgren. On-beat/off-beat: Visual responses to audio-visual asynchrony in music videos. *Projections*, 15(1):28–54, 2021. doi: 10.3167/proj.2021.150103. URL https://doi.org/10.3167/proj.2021.150103.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/d877ea0dd78bdbe54830670618c1de09-Abstract-Conference.html.

Yutong Wang, Sidan Zhu, Hongteng Xu, and Dixin Luo. An inverse partial optimal transport framework for music-guided movie trailer generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9739–9748, 2024. doi: 10.1145/3664647.3680751. URL https://dixin-lab.github.io/project-page-trailer-generator/.

Yutong Wang, Yunke Wang, Xinyuan Chen, and Chang Xu. BEAT: Rhythm-elastic alignment for agentic music-guided movie trailer generation. *arXiv preprint arXiv:2605.27067*, 2026. doi: 10.48550/arXiv.2605.27067. URL https://arxiv.org/abs/2605.27067v1.

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, Zhijian Liu, Pavlo Molchanov, Ping Luo, Song Han, and Enze Xie. Fast-dLLM v2: Efficient block-diffusion LLM. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d0865cbe51d35ec322f9af9db7806fc7-Abstract-Conference.html.

Zijie Xin, Minquan Wang, Jingyu Liu, Quan Chen, Ye Ma, Peng Jiang, and Xirong Li. Music grounding by short video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22285–22293, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xin_Music_Grounding_by_Short_Video_ICCV_2025_paper.html.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Ryo Yonetani, Tatsunori Taniai, Mohammadamin Barekatain, Mai Nishimura, and Asako Kanezaki. Path planning using Neural A* search. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12029–12039, 2021. URL https://proceedings.mlr.press/v139/yonetani21a.html.

Xinyu Zhang, Dong Gong, Zicheng Duan, Anton van den Hengel, and Lingqiao Liu. Let your video listen to your music! *arXiv preprint arXiv:2506.18881*, 2025. doi: 10.48550/arXiv.2506.18881. URL https://arxiv.org/abs/2506.18881.

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. doi: 10.48550/arXiv.2603.29664. URL https://arxiv.org/abs/2603.29664v1.

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. doi: 10.48550/arXiv.2606.23327. URL https://arxiv.org/abs/2606.23327v2.

Sidan Zhu, Yutong Wang, Hongteng Xu, and Dixin Luo. Weakly-supervised movie trailer generation driven by multi-modal semantic consistency. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10234–10242, 2025a. doi: 10.24963/ijcai.2025/1137. URL https://www.ijcai.org/proceedings/2025/1137.

Sidan Zhu, Hongteng Xu, and Dixin Luo. Self-paced and self-corrective masked prediction for movie trailer generation. *arXiv preprint arXiv:2512.04426*, 2025b. doi: 10.48550/arXiv.2512.04426. URL https://arxiv.org/abs/2512.04426v3. Version 3, revised March 2026.

A Data Audit Notes

Figure 2's overlapping clip ranges are illustrative, not a contiguity restriction. Every combo retains its parent segment's full audio. Grid checks include S_0, internal cuts, and END; END is not silently moved to restore feasibility. Augmentations are rechecked and deduplicated after mapping.

Source tags use metadata rules, not visual verification: one match assigns a category, multiple matches give Multi-category, and no match gives Mix (1,237 sources). Mix examples summarize a random 50-source title/tag review, not corpus-wide relabeling. Other combines pets/animals, travel/nature, and food/cooking, each below 2%. Counts weight unique sources equally, not their combos, duration, or training frequency. Automated screening and parsing can introduce selection bias.

B Model Configuration and Evidence Encoding

Configuration and parameters. The main 512-dimensional Cutalio network has **32,088,144 trainable parameters**. Its 36,850,768 registered parameters also include 4,762,624 in a frozen, bypassed legacy audio-query stack. Counts include adapters, backbone, scoring/auxiliary heads, and feedback, but exclude the frozen offline Qwen3-Omni encoders. Table 5 matches the saved configuration.

Table 5: **Main-model configuration.**

Component	Setting
Input representation	2048-D visual/audio features projected to 512; 4 pooled tokens/clip
Shared / audio blocks	8 / 1 blocks; 8 / 4 heads; FFN widths 1536 / 1024
Anchors / regularization	Audio and rhythm at blocks 5–8; zero-initialized scalar gates; dropout 0.1
Scoring / context	4 evidence types; 12-D rhythm, 6-D relation; all retained grid candidates
Input cap	At most 1500 audio tokens; uniform selection, not global averaging
Training	20 epochs; 4 GPUs; 32 original combos/GPU; seed 42; bfloat16
Optimization	AdamW; LR 10^{-4}; decay 0.01; gradient-norm clip 1; 50-step warmup, then cosine decay
Masking / feedback	All-MASK probability 0.1, otherwise random masking (minimum ratio 0.05); P2; feedback weight 1
Sampling	Temperature 0.8; top-p 0.95; model-ranked top-1

Video and rhythm coordinates. Four pooled tokens are not four raw frames. Within a clip of M sampled visual tokens, token k has relative position $(k + 1/2)/M$; source duration and token count are not embedded. This removes an explicit seconds-valued video clock, not indirect duration cues from content/motion. Between detected beats, $b(t) = k + (t - t_k^{\text{beat}})/(t_{k+1}^{\text{beat}} - t_k^{\text{beat}})$; outside their range, the nearest interval is extrapolated. Beats per bar is the rounded median downbeat-to-downbeat beat distance, clipped to $[1, 16]$; this is not a variable-time-signature model. $\kappa(t)$ counts downbeat intervals, distinct from event-channel vector $\mathbf{c}_j$.

Exact relation descriptor. For $C_a(x) = \min(a, \max(-a, x))$ and $\epsilon = 10^{-4}$,

$$\phi_{ij} = \Big[C_2(d_{ij}^{\text{l}}/8),\ \log(1 + \max(d_{ij}^{\text{l}}, 0))/\log 17,\ m_i C_2(d_{ij}^{\text{r}}/8),\ m_i C_3\left(\log \frac{\max(d_{ij}^{\text{l}}, \epsilon)}{\max(d_{ij}^{\text{r}}, \epsilon)}\right),\ C_2(G_{ij}/4),\ m_i\Big]. \tag{10}$$

No right anchor zeros entries 3, 4, and 6; unavailable bar indices zero entry 5. Invalid candidates are masked. Left-only log compression and bar reference are implementation choices, not a claim that symmetric encoding is unnecessary.

Residual sources and scorer. R_0 sums the multi-label event embedding and a projection of interval, beat/bar phase, and scaled relative beat position, subtracting the projection of zero input. It excludes audio, S_0, and modality embeddings and differs from scoring $(R/B)_j$. A_0 precedes time

encoding, the audio Transformer, and rhythm fusion. Both remain differentiable and fixed across shared blocks. Each evidence type has separate normalization and key/value projection:

$$\begin{aligned} \mathbf{q}_i &= W_q\,\mathrm{LN}(h_i^{\mathrm{post}}), \quad (\mathbf{k}_{ijr}, \mathbf{v}_{ijr}) = \mathrm{Proj}_r(\mathrm{LN}(\mathcal{E}_{ijr})), \\ w_{ijr} &= \mathrm{softmax}_r(\mathbf{q}_i^\top \mathbf{k}_{ijr}/\sqrt{d}), \quad \mathrm{read}_{ij} = \sum_r w_{ijr}\mathbf{v}_{ijr}, \\ s_{ij} &= \mathrm{MLP}([\mathrm{LN}(h_i^{\mathrm{post}}); \mathrm{LN}(e_j^{\mathrm{pre}}); \mathrm{LN}(\mathrm{read}_{ij})]). \end{aligned} \tag{11}$$

Softmax is over types, not candidate times; the SiLU MLP maps $3d \to d \to 1$.

Training losses. On valid masked slots, each pass uses the saved weights

$$\begin{aligned} \mathcal{L}^{(p)} = \mathcal{L}^{(p)}_{\mathrm{cand}} + 0.25\mathcal{L}^{(p)}_{\mathrm{state}} + 0.25\mathcal{L}^{(p)}_{\mathrm{route}} + 0.2\mathcal{L}^{(p)}_{\mathrm{hier}} \\ + 0.2\mathcal{L}^{(p)}_{\mathrm{hard}} + 0.1\mathcal{L}^{(p)}_{\mathrm{aux}}, \qquad \mathcal{L}_{\mathrm{train}} = \tfrac{1}{2}(\mathcal{L}^{(1)} + \mathcal{L}^{(2)}). \end{aligned} \tag{12}$$

State supervision uses smooth L1 on phase pairs and binary cross-entropy on event/availability flags; unavailable bar phases are masked. Routing uses cross-entropy on channel distributions. Hierarchy ranks distance-to-target buckets (0.25/0.75/1.5/3 beats; log-mean-exp bucket scores; margin 0.15). The hard-negative hinge compares GT with the highest-scoring non-target within 1.25 beats (margin 0.25). Relation auxiliaries are span/type classification (weights 1 each), bar-span/phase (0.5 each), density (0.25), and beat-span regression (0.5), with unavailable targets masked. Losses are reduced per original combo; GT and available augmentation views are summed as in Section 6.4.

Initialization caveat. The deterministic two-pass baseline starts without feedback; the historical TTS sampler starts with a zero feedback tensor. Its projection has learnable normalization parameters, so these are not equivalent initializations and the comparison does not isolate computation alone.

C Evaluation Details

Acoustic feature extraction. Original music is decoded to 16 kHz mono; STFT uses a 1,024-point Hann window, hop 256, and no centered padding. Log-mel shape subtracts the within-frame mean over 48 bins. Shape and log energy use whole-track per-dimension median centering and $1.4826\,\mathrm{MAD}$ scaling, with standard-deviation fallback and a 3 dB scale floor. Twelve-dimensional chroma is L1-normalized and square-rooted. No statistics are fitted to individual GT cuts. For frames f_m at times τ_m,

$$a_{f,\sigma}(t) = \frac{\sum_m \exp[-(\tau_m - t)^2/(2\sigma^2)] f_m}{\sum_m \exp[-(\tau_m - t)^2/(2\sigma^2)]}, \qquad \sigma \in \{0.16, 0.50\}\ \mathrm{s}. \tag{13}$$

Smoothing normalizes over available frames, including track edges; there is no zero-padding or reflection. The five LocalTrend offsets are observation centers, not crop boundaries. Cuts correspond by slot, without rematching, offset search, stretching, or dynamic time warping.

Eligibility and aggregation. For each feature-family/scale component, reference mean-square energy of the pooled centered observations must exceed 10^{-8}. A locally flat GT cut is not individually excluded if other cuts vary; flat predictions cannot remove a reference-eligible component. Each acoustic metric averages valid scales per family, then valid families equally. JointH4 uses

$$\begin{aligned} (q_1, q_2, q_3, q_4) = \left(\frac{\mathrm{LocalTrend} + 1}{2}, \frac{\mathrm{SeqTrend} + 1}{2}, \mathrm{RelOrder}, 2^{-\mathrm{RelRatio}}\right), \\ \mathrm{JointH4} = \frac{4}{\sum_k 1/q_k}. \end{aligned} \tag{14}$$

A zero component gives zero. Aggregate per combo before macro-averaging; column means cannot reconstruct JointH4. Fixed-support reporting assigns invalid predictions zero on reference-eligible support. Table 1 instead conditions on all models' geometrically valid outputs and discloses failures separately (Appendix D.3). For continuous Qwen times, $b(p_i)$ interpolates frozen grid-time/beat pairs, with nearest-interval extension inside waveform edges; no output is snapped or repaired. Acoustic evaluation uses original seconds.

Interpretation examples. For GT beat spans $[2, 4, 8]$ and prediction $[3, 9, 6]$, RelOrder is $2/3$ and RelRatio ≈ 1.085; prediction $[3, 6, 12]$ gives 1 and 0 despite a different total span. A ratio 2 changed to 4 or 1 incurs the same log-ratio error 1. END is excluded. The acoustic agreement is not cosine similarity: $C(r, 2r) = 0.8$ for nonzero r; a flat prediction against a varying reference scores zero. SeqTrend gives 1 to scalar sequences $[1, 3, 2]$ and $[6, 8, 7]$, but -1 to $[1, 3, 2]$ versus $[8, 6, 7]$, after across-cut centering. Raw scores $(0.2, 0.4, 0.8, 1)$ give JointH4 ≈ 0.6304; an acoustic zero maps to 0.5, so JointH4 is not an accuracy or percentage. These scores quantify reference correspondence, not human preference. Means weight combos equally; uncertainty resamples source-video groups. Path selection never uses evaluation scores.

Known-anchor completion diagnostics. For Table 4, let $\mathcal{W}$ exclude the union of all script-selected internal anchor slots in *every* setting, including the no-anchor baseline; neither S_0 nor END belongs to $\mathcal{W}$. With predicted seconds p_i, original seconds g_i^{raw}, and grid indices z_i, z_i^{GT}, we use

$$\mathrm{MAE}_{\mathcal{W}} = |\mathcal{W}|^{-1} \sum_{i \in \mathcal{W}} |p_i - g_i^{\mathrm{raw}}|, \qquad \mathrm{GridExact}_{\mathcal{W}} = |\mathcal{W}|^{-1} \sum_{i \in \mathcal{W}} \mathbf{1}\{z_i = z_i^{\mathrm{GT}}\},$$
$$\mathrm{ENDerror} = |p_{\mathrm{END}} - g_{\mathrm{END}}^{\mathrm{raw}}|. \tag{15}$$

LocalTrend pools centered observations over $\mathcal{W}$ before normalization. Common-valid supports are 30 combos for withheld cuts and 31 for END, out of 32; empty $\mathcal{W}$ or insufficient reference variation leaves the corresponding metric undefined. MAE/END are in seconds; GridExact is a grid-dependent fraction. Removing anchors before interval differencing would merge original clip intervals, so completion uses these pointwise/local diagnostics rather than redefining full-path relative metrics. GT-grid anchors make this oracle completion, not automatic script evaluation.

D Supplementary Experiments

Historical architecture diagnostics. Table 6 uses an older training mixture, not matched final-model ablations. Checkpoints are selected by baseline validation JointH4 before TTS; all use GT-grid S_0 and model-only ranking. The first three TTS rows average five sampling seeds (42–46), not retraining. Gated fusion jointly removes neural candidate–state relations, changes the fusion operator, and removes separate local-audio and scorer R/B branches. Its decline cannot be attributed to ϕ_{ij} alone.

Table 6: **Historical architecture diagnostics: JointH4.** Shared support: 302 combos, 197 sources. $\pm$: sampling-seed SD; last two rows use one sampling seed.

Historical model	Baseline	TTS	Budget
No A/R residuals	0.5703	0.5781 ± 0.0043	D4/P2/K8
Cutalio (historical)	0.5722	0.5880 ± 0.0040	D4/P2/K8
Gated fusion	0.5708	0.5405 ± 0.0072	D4/P2/K8
No A/R residuals	0.5703	0.5793	D4/P2/K1
w/o scorer relations	0.5618	0.5487	D4/P2/K1

Crossed source/sampling-seed 95% gain intervals for the first three rows are $[-0.0080, 0.0225]$, $[-0.0002, 0.0310]$, and $[-0.0455, -0.0146]$. The scorer-only ablation removes its relation projection/evidence slot but retains backbone relations and DP. Its D4/K1 deficit against the reference is -0.0306 ($[-0.0467, -0.0140]$). At K4/K8 it scores 0.5696/0.5628 versus 0.5759/0.5718; between-model intervals cross zero and the ablation improves over its own baseline. This supports relations at the smallest tested budget, not their necessity for every TTS gain. Baseline/TTS initial feedback differs; intervals are descriptive, without multiplicity correction.

D.1 Additional augmentation controls

Each enabled family supplies one available complete path alongside GT. With both families, $\mathcal{L} = \mathcal{L}_{\mathrm{GT}} + \frac{1}{2}\mathcal{L}_{\mathrm{fixed}} + \frac{1}{2}\mathcal{L}_{\mathrm{independent}}$; missing families contribute zero without weight reallocation. Each view uses its own S_0; evaluation uses GT-grid S_0. Fixed-start preserves its generation-time start, not necessarily the calibrated GT cell.

Table 7: **Four-way augmentation controls: JointH4.** P2; matched original-example updates, not FLOPs; baseline-selected epochs 16/19/20/17 in row order.

Training data	Baseline	D4/P2/K4	D8/P2/K16
GT only	**0.5896**	0.5984	0.5844
GT + fixed start	0.5793	0.5971	0.5932
GT + independent start	0.5809	**0.6054**	**0.6075**
GT + both	0.5654	0.5982	0.5838

Independent-start minus GT-only *TTS gain over each baseline* is 0.0318 at D8/P2/K16 (paired source-cluster 95% interval $[0.0057, 0.0592]$) and 0.0156 at D4/P2/K4 ($[-0.0064, 0.0396]$), on 302 combos/197 sources. Training FLOPs and selected epochs differ. These single-seed exploratory interactions do not isolate whether augmentation helps generation, ranking, or both.

D.2 Complete test-time scaling sweep

Table 8 retains the full-cohort results; all 343 examples and requested draws are valid. Supports are 343/324/302/324/302, unlike the common-valid main comparison. Fwd. sums predictions across all K draws, then averages over examples; early completion lowers the nominal DPK count. Ranking reuses generation scores. Counts exclude preprocessing and non-neural DP/metrics and are neither FLOPs nor latency.

Table 8: **Complete mainline TTS sweep.** Same checkpoint, GT-grid S_0, nested seeds, and model-ranked top-1 as Section 6.1.

Inference	Fwd.	Local↑	Seq↑	Order↑	Ratio↓	JointH4↑
Two-pass baseline	2.00	**0.2797**	0.4337	0.6536	0.9776	0.5809
D4/P2/K1	7.65	0.2273	0.3917	0.6715	0.9073	0.5864
D4/P2/K2	15.30	0.2505	0.3968	0.6792	0.8609	0.5942
D4/P2/K4	30.60	0.2652	0.4058	**0.6898**	0.8218	0.6054
D4/P2/K8	61.20	0.2500	0.4137	0.6724	0.8114	0.5957
D4/P2/K16	122.40	0.2736	**0.4377**	0.6774	0.7997	0.6036
D8/P2/K1	12.91	0.2239	0.3798	0.6894	0.8440	0.5918
D8/P2/K2	25.82	0.2508	0.3818	0.6760	0.8220	0.5917
D8/P2/K4	51.64	0.2620	0.3977	0.6848	0.8193	0.6013
D8/P2/K8	103.28	0.2550	0.4075	0.6699	0.7885	0.5957
D8/P2/K16	206.55	0.2790	0.4233	0.6830	**0.7588**	**0.6075**

D.3 Task-adapted Qwen3-Omni baseline

Inputs and training. Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) receives ordered clips, full music, clip count, S_0, virtual 5-second source budgets, planning limits, and grid-aligned beat/downbeat/onset timestamps in a Chinese prompt, but no GT cuts/END, dense candidate grid, or Cutalio features. Output is JSON `start_points` and `end` in seconds; non-event times are allowed. Four frames span each full clip including endpoints; short clips repeat indices. Original timestamps/FPS are omitted; processor-default timing remains. The video pixel budget is 200,704; audio is uncapped. Four spatial-token frames are not equivalent to four pooled Cutalio tokens.

SFT starts from the pretrained checkpoint with 16,340 GT and 4,965 independent views before bad-media filtering. One epoch uses response-only loss, frozen vision/audio encoders, Adam at 10^{-5}, constant LR after 0.03 warmup ratio, gradient clipping 1, and seed 42. Microbatch/accumulation are 1/4 on 32 workers, with tensor/pipeline/expert/context parallelism 2 each and context limit 98,304. Final step 1329 is not validation-selected. Inference is one greedy generation, at most 2,048 new tokens, without thinking/audio output, retries, snapping, DP repair, or GT ranking.

Validity and uncertainty. Of 343 outputs, 322 are geometrically valid; failures are malformed JSON (2), wrong counts (4), non-increasing times (2), and out-of-waveform times (13). Among

valid outputs, 22 paths have 26 sub-0.2-second intervals; none exceeds the audited 10-second/16-beat maxima. These planning violations remain in the quality comparison. Common-valid supports are 322/304/282/304/282. On 282 combos/189 sources, Qwen-minus-Cutalio JointH4 differences are 0.0463 $[0.0202, 0.0741]$ for baseline, 0.0211 $[-0.0010, 0.0434]$ for D4/K4, and 0.0212 $[-0.0036, 0.0465]$ for D8/K16 (descriptive 95% source-cluster intervals). Crossing zero does not establish equivalence. Pretraining, capacity, input compression, exposures, checkpoint selection, and decoding costs are unmatched; no architecture-only or compute-efficiency claim follows.

D.4 Independently trained modality controls

V+A and V+R use the main model's GT+independent-start training contract and selection rule; selected epochs are 20/17, versus 20 for V+A+R. V+A zeros explicit neural rhythm and derived relations in backbone, anchors, and scorer. V+R zeros neural audio content. Both retain video, time, candidate slots, GT-grid S_0, DP/feedback, targets, auxiliary supervision, and parameter shapes. Thus V+A is not rhythm-free, nor an isolated ϕ_{ij} ablation. All 343 outputs/draws are valid; seeds and model-only top-1 ranking match.

Table 9: **All-budget modality comparison: JointH4** ($n = 302$). Same P2 and nested draws; bold denotes each variant's maximum, not significance.

Inference	V+A+R	V+A	V+R
Two-pass baseline	0.5809	0.5513	0.5779
D4/K1	0.5864	0.5700	0.5736
D4/K2	0.5942	0.5706	0.5803
D4/K4	0.6054	0.5792	0.5895
D4/K8	0.5957	0.5849	0.5870
D4/K16	0.6036	0.5873	0.5912
D8/K1	0.5918	0.5833	0.5875
D8/K2	0.5917	0.5819	0.5981
D8/K4	0.6013	0.5906	0.5936
D8/K8	0.5957	0.5865	0.5960
D8/K16	**0.6075**	**0.6018**	**0.5997**

V+A+R does not win every budget (Table 9). At baseline, its JointH4 advantage over V+A is 0.0296 (paired 95% interval: $[0.0105, 0.0483]$). Relative to V+R, gains are 0.0314 in Local ($[0.0089, 0.0528]$), 0.0218 in Seq ($[0.0038, 0.0407]$), and 0.0029 in JointH4 ($[-0.0136, 0.0192]$). At D8/K16, the JointH4 gap over V+A is 0.0058 ($[-0.0157, 0.0278]$); over V+R it is 0.0079 ($[-0.0082, 0.0238]$). V+A has the largest observed TTS gain; gain-difference intervals also cross zero. These descriptive, single-seed results do not establish significant composite-score superiority or that explicit neural rhythm is necessary for TTS.

E Qualitative Trajectory Examples

Figure 4 compares saved two-pass and D4/P2/K4 outputs of the main GT+independent-start model. Among 150 readable 4–9-clip cases, 20 have higher JointH4, lower RelRatio, no lower Local/Seq/Order, no greater internal-cut MAE, and closer full-duration proportions (L1 distance after normalizing each duration vector by its total span, including END). A is the lower-median RelRatio reduction in this subset; B illustrates ratio degradation despite a higher composite score. These are *post-hoc illustrations*, not prevalence estimates. Each path remains the original model-ranked top-1; GT never reranks the four draws.

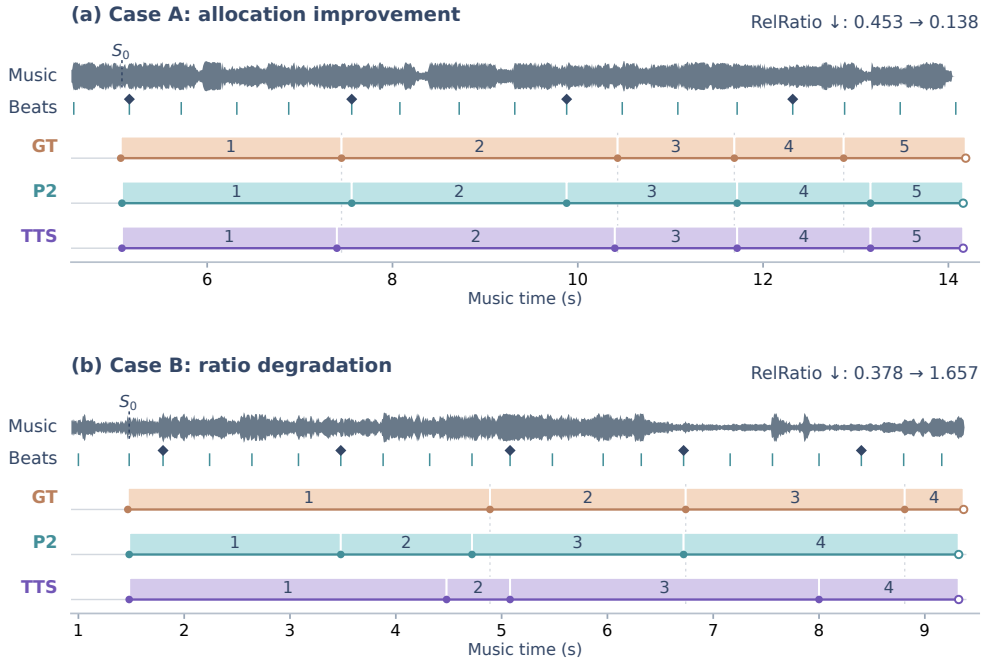

Figure 4: **Real model outputs: improved allocation and a metric trade-off.** Waveforms are from the original music; ticks denote grid-aligned detected beats, diamonds downbeats. Numbered intervals are ordered clips, not video thumbnails. Rows show the raw reference (GT), two-pass baseline (P2), and D4/P2/K4 (TTS); open circles mark END. Both model rows share the given GT-grid S_0, which can differ slightly from raw GT. Axes retain absolute music time and are cropped around the paths. END is displayed but excluded from the four component metrics; RelRatio measures beat-span ratios, not seconds.

Table 10: **Per-example scores for Figure 4.** Acoustic and duration-allocation changes need not agree.

Case	Inference	Local↑	Seq↑	Order↑	Ratio↓	JointH4↑
A	Two-pass	0.4787	0.7929	0.8000	0.4525	0.7864
A	D4/P2/K4	0.6920	0.9223	1.0000	0.1376	0.9254
B	Two-pass	−0.3763	−0.0447	0.5000	0.3780	0.4651
B	D4/P2/K4	0.1441	0.2531	1.0000	1.6570	0.5335

What changes. In A, TTS moves the second cut from 9.88 to 10.40 s (GT: 10.43 s), better recovering the long second and short third intervals. All four component metrics improve. In B, the second interval shrinks from 1.24 to 0.60 s, against 1.85 s in GT. Ratio error worsens even though acoustic correspondence, duration ordering, and JointH4 improve. B is a failure of ratio preservation, not an overall-score decline: a higher composite score does not imply uniformly better rhythm. Aggregate results remain in Section 6.2.