NULL-SPACE ALIGNMENT FOR INVERSION-FREE MUSIC EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic music editing is inherently underdetermined: a text instruction specifies *what* should change, but leaves many musical degrees of freedom in *how* that change is realized. An effective editing system must therefore distinguish between what is explicitly prescribed by the instruction and what remains unspecified for the generative prior to determine. Motivated by this view, we investigate whether the conditional flow difference of a pretrained flow-matching model naturally exhibits such a separation. We discover that projecting this difference onto the local edit trajectory yields two complementary components, including a *parallel* component driving the prescribed change and an *orthogonal null-space* component. Crucially, increasing the strength of the null-space component consistently shifts the output from source-content fidelity toward stronger instruction alignment. This indicates that the null-space component provides a controllable degree of generative freedom, balancing source preservation and semantic adaptation. Building on this insight, we propose Null-Space Alignment (NSA), a training-free and inversion-free framework that gates the null-space component based on its directional consistency. Experiments on MelodiaEdit and ZoME-Bench demonstrate that NSA achieves consistent state-of-the-art performance across both quantitative metrics and human evaluations.

1 INTRODUCTION

Text-guided music editing Han et al. (2024); Zhang et al. (2024a) differs fundamentally from generative music synthesis Liu et al. (2023). While pure generation produces a plausible sample from scratch given a prompt, semantic editing begins with an existing audio and modifies only the aspects specified by a text instruction. Specifically, the source audio establishes the baseline melody, rhythm, timbre, and structural composition. Therefore, effective editing must precisely execute the target modification while faithfully preserving all uninstructed musical structures.

This raises a fundamental question: how can a generative model execute an edit instruction without corrupting the underlying structure of the source audio? Ideally, the instruction dictates what should change, while the source audio and pretrained prior determine how the modification is realized by providing structural constraints and generative knowledge. Recent flow-matching frameworks Kulikov et al. (2025); Lipman et al. (2022) show that the difference between conditional velocity fields under source and target conditions provides a natural editing direction. In this work, we hope to analyze how the instruction and the pretrained prior interact along these generative dynamics.

Specifically, given a local vector defining the edit trajectory in progress, we project the conditional flow difference onto a parallel component aligned with the edit direction and an orthogonal null-space component representing variations outside this direction. As illustrated in Figure 1, we observe a consistent relationship between the magnitude of the orthogonal component and the resulting editing behavior. Increasing the orthogonal component progressively shifts the output from high source-content fidelity toward stronger instruction alignment across diverse datasets. This observation suggests that the orthogonal component serves as a controllable degree of freedom, allowing the pretrained prior to introduce generative flexibility while maintaining the principal semantic trajectory of the edit.

Based on this, we propose Null-Space Alignment (NSA), instantiated within the latent space of pretrained flow-matching models, leveraging their conditional velocity fields to perform flow vector

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.

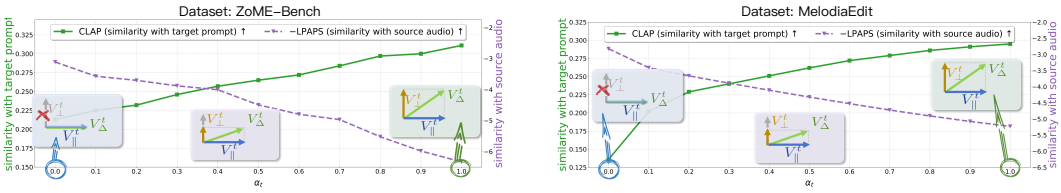

Figure 1: **Continuous control via the orthogonal component.** Holding the parallel component fixed while varying the magnitude of the orthogonal component reveals a smooth trade-off. As its strength increases, the edited audio gradually shifts from higher source-content similarity toward stronger instruction alignment. This consistent trend across datasets and prompts demonstrates that the null-space is not an uninformative residual, but a controllable degree of generative freedom.

decomposition. Starting from source and target prompts, we construct an inversion-free generative proposal derived directly from the conditional flow dynamics. We decompose this velocity proposal into a parallel component that drives the targeted semantic change and an orthogonal component that captures generative variation. To regulate this variation, we introduce an adaptive gate that selectively retains the orthogonal component based on its alignment with the main edit trajectory. The resulting framework grants precise, continuous control over edit intensity while maintaining source structural integrity without requiring fine-tuning. Our main contributions are as follows:

- We introduce a novel view of semantic music editing that coordinates explicit textual instructions and the implicit generative prior within the null-space.
- We develop Null-Space Alignment (NSA), a training-free framework that decomposes conditional flow velocity fields into parallel and orthogonal components, separating instruction-driven edit motion from prior-driven generative freedom.
- We propose a gating mechanism that filters the orthogonal component based on its directional alignment with the edit vector, enabling fine-grained control over semantic transformation while preserving underlying musical structures.

2 Related Work

Benefiting from recent advances in diffusion and flow matching models Hung et al. (2026); Zhu et al. (2026), music editing has witnessed rapid development in recent years. Early approaches mainly follow a training-based paradigm, using dedicated diffusion models Han et al. (2024); Wang et al. (2023); Lin et al. (2024) or instruction-tuning pretrained music generators Zhang et al. (2024a) to acquire controllable editing capabilities over instruments, styles, and local musical attributes. However, these methods typically require additional training data, task-specific adaptation, and computational overhead. Zero-shot approaches subsequently emerge to manipulate the generation trajectories of pretrained models without retraining Yang et al. (2025); Xu et al. (2024); Yang et al. (2026); Li et al. (2026); Jia et al. (2025), incorporating latent-space constraints Zhang et al. (2024b), DDPM inversion Manor & Michaeli (2024), or disentangled inversion control Liu et al. (2025) to improve editing consistency and controllability. Nevertheless, these methods still require mapping the source music to noise or intermediate latent states, where inversion can accumulate reconstruction errors and distort fine-grained musical structures. To mitigate such information loss, recent works construct editing updates directly from the difference between source- and target-conditioned denoising scores Niu et al. (2026); Ge et al. (2026), thereby bypassing the full forward–backward diffusion process. Flow-matching models Lipman et al. (2022); Liu et al. (2022); Lipman et al. (2024) provide a more direct formulation for inversion-free editing by describing distribution transport through continuous velocity fields. Recent studies have shown that source-to-target editing dynamics can be constructed directly from pretrained flow models Kulikov et al. (2025); Fu et al. (2026); Lei et al. (2026) without explicit inversion or per-instance optimization. However, the velocity difference between source and target conditions may still contain motion components irrelevant to the desired semantic transformation, leading to unnecessary trajectory deviation and degraded source-content preservation. In this paper, we aim to regulate the effective editing direction while retaining the advantages of inversion-free and training-free inference.

3 PRELIMINARY

3.1 INVERSION-FREE GENERATIVE PROPOSAL

We consider a pretrained flow-matching model whose latent dynamics $z_t \in \mathbb{R}^d$ for flow time $t \in [0, 1]$ are governed by the Ordinary Differential Equation (ODE):

$$\frac{dz_t}{dt} = v_\theta(z_t, t, \psi), \tag{1}$$

where v_θ denotes the velocity model and ψ represents conditioning text prompts.

Given a source audio x_s with source condition ψ^{src} and target condition ψ^{tar}, we adopt an inversion-free flow formulation Kulikov et al. (2025) and maintain an edit trajectory z_t^{edit} initialized at $z_0^{\text{edit}} = x_s$. At flow step t, the source- and target-conditioned states are constructed via shared Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$z_t^{\text{src}} = (1-t)x_s + t\epsilon, \qquad z_t^{\text{tar}} = z_t^{\text{edit}} + z_t^{\text{src}} - x_s. \tag{2}$$

Evaluating conditional velocity fields at these states yields $V_t^{\text{src}} = v_\theta(z_t^{\text{src}}, t, \psi^{\text{src}})$ and $V_t^{\text{tar}} = v_\theta(z_t^{\text{tar}}, t, \psi^{\text{tar}})$. The model-induced editing proposal is defined as their vector difference:

$$V_\Delta^t = V_t^{\text{tar}} - V_t^{\text{src}}. \tag{3}$$

3.2 THEORETICAL LIMITATIONS OF NAIVE PROPOSALS

To determine whether V_Δ^t is a theoretically appropriate update direction, we formulate a quantitative metric to measure local editing progress and analyze how integration along V_Δ^t affects trajectory fidelity.

Evaluating Proposal Quality via Local Discrepancy. Let z_t^{tar} denote the ideal target-conditioned latent state sharing the same source noise realization at flow time t. We define the *local discrepancy metric* $\mathcal{L}_t(z)$ as the squared Euclidean distance to this target landmark:

$$\mathcal{L}_t(z) = \frac{1}{2}\left\|z - z_t^{\text{tar}}\right\|^2. \tag{4}$$

The canonical direction of maximum local progress toward z_t^{tar} is given by the negative gradient of $\mathcal{L}_t(z)$ evaluated at the current source state z_t^{src}:

$$D_t \triangleq -\nabla_z \mathcal{L}_t(z_t^{\text{src}}) = z_t^{\text{tar}} - z_t^{\text{src}}. \tag{5}$$

Here, D_t represents the local transport vector required to shift the latent state from the source trajectory onto the target trajectory.

Decomposing the Proposal V_Δ^t. An appropriate editing proposal should align closely with D_t to reduce $\mathcal{L}_t$. To evaluate the naive velocity proposal V_Δ^t, we project it relative to the canonical edit direction D_t:

$$V_\Delta^t = V_\parallel^t + V_\perp^t, \quad \text{where} \quad V_\parallel^t = \gamma D_t \quad (\gamma \in \mathbb{R}), \quad \text{and} \quad (V_\perp^t)^\top D_t = 0. \tag{6}$$

Here, $V_\parallel^t$ is the effective transport component aligned with D_t (with alignment scale γ), while $V_\perp^t$ is the orthogonal velocity residual lying in the subspace orthogonal to D_t.

When advancing the current state z_t^{src} along the proposal V_Δ^t using a discrete integration step $\eta > 0$, the resulting change in local discrepancy $\Delta\mathcal{L}_t(V_\Delta^t) \triangleq \mathcal{L}_t(z_t^{\text{src}} + \eta V_\Delta^t) - \mathcal{L}_t(z_t^{\text{src}})$ expands as:

$$\begin{aligned}\Delta\mathcal{L}_t(V_\Delta^t) &= \frac{1}{2}\left\|(z_t^{\text{src}} - z_t^{\text{tar}}) + \eta(V_\parallel^t + V_\perp^t)\right\|^2 - \frac{1}{2}\left\|z_t^{\text{src}} - z_t^{\text{tar}}\right\|^2 \\ &= -\eta D_t^\top (V_\parallel^t + V_\perp^t) + \frac{\eta^2}{2}\|V_\parallel^t + V_\perp^t\|^2.\end{aligned} \tag{7}$$

Exploiting the orthogonality condition $(V_\perp^t)^\top D_t = 0$, this simplifies to:

$$\Delta\mathcal{L}_t(V_\Delta^t) = \underbrace{-\eta\gamma\|D_t\|^2 + \frac{\eta^2}{2}\gamma^2\|D_t\|^2}_{\text{Target Transport Progress}} + \underbrace{\frac{\eta^2}{2}\|V_\perp^t\|^2}_{\text{Orthogonal Trajectory Drift}}. \tag{8}$$

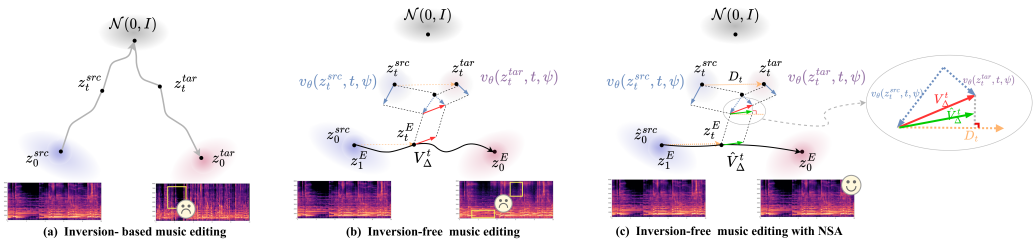

Figure 2: **Comparison of inversion-based and inversion-free music editing frameworks.** (a) *Inversion-based editing* maps the source audio latent z_0^{src} to noise space $\hat{z}_1$ via reverse ODE under source conditions, then integrates forward under target conditions. (b) *Standard inversion-free editing* traverses a direct path z_t^{E} without inversion, using source- and target-conditioned velocities to update the trajectory. (c) *Null-Space Alignment (NSA)* decomposes the editing velocity into direction-aligned ($V_{\parallel}^t$) and orthogonal ($V_{\perp}^t$) components, adaptively suppressing orthogonal motion while preserving target-directed transport for improved structural preservation.

Proposition 1 (When Does V_Δ^t Fail). *Suppose V_Δ^t has positive target alignment ($\gamma > 0$). A discrete step along the monolithic proposal V_Δ^t fails to reduce local discrepancy (i.e., $\Delta\mathcal{L}_t(V_\Delta^t) > 0$) whenever the energy ratio of the orthogonal residual exceeds:*

$$\frac{\|V_\perp^t\|^2}{\|D_t\|^2} > \frac{2\gamma}{\eta} - \gamma^2. \tag{9}$$

In practice, raw velocity differences $V_\Delta^t = V_\theta(z_t, \psi^{\text{tar}}) - V_\theta(z_t, \psi^{\text{src}})$ produced by neural networks are noisy and imperfectly aligned with D_t. When V_Δ^t contains significant off-target noise, $\|V_\perp^t\|^2$ easily breaches the threshold in Proposition 1. As a result, updating directly along the naive proposal V_Δ^t causes the latent path to drift away from the target manifold, explaining the severe structural artifacts and content degradation observed in monolithic inversion-free editing.

4 Methodology

Our theoretical analysis in Section 3.2 established that naive inversion-free editing updates (Figure 2(b)) accumulate larger editing deviations when orthogonal noise energy breaches the divergence threshold. To eliminate this vulnerability, we propose *Null-Space Alignment (NSA)*, a lightweight, inversion-free framework that explicitly rectifies the velocity proposal V_Δ^t during generation. As depicted in Figure 2(c), NSA isolates the target-directed velocity component $V_{\parallel}^t$ from the orthogonal residual $V_{\perp}^t$, adaptively suppressing the orthogonal motion that drives trajectory drift while preserving primary edit transport.

4.1 Null-Space Decomposition

As demonstrated in Figure 1, holding the target-aligned velocity component fixed while smoothly varying the magnitude of the orthogonal component reveals that the orthogonal velocity space encodes a controllable degree of generative freedom. Smaller orthogonal magnitudes tightly preserve source content, whereas larger magnitudes allow the pretrained prior to execute richer, instruction-aligned transformations. To systematically harness this trade-off without triggering the divergence bounds of Proposition 1, we explicitly disentangle these two functional roles.

For any non-zero edit displacement $\|D_t\| > 0$, we decompose the raw model proposal V_Δ^t into orthogonal vector fields aligned with and perpendicular to the canonical transport direction $D_t = z_t^{\text{tar}} - z_t^{\text{src}}$:

$$V_\Delta^t = V_{\parallel}^t + V_{\perp}^t. \tag{10}$$

Using the orthogonal projection operator onto D_t, the parallel component $V_{\parallel}^t$ and orthogonal complement $V_{\perp}^t$ are defined as:

$$V_{\parallel}^t = \frac{(V_\Delta^t)^\top D_t}{\|D_t\|^2} D_t, \qquad V_{\perp}^t = \left(I - \frac{D_t D_t^\top}{\|D_t\|^2}\right) V_\Delta^t. \tag{11}$$

By construction, $V_\perp^t$ lies in the kernel (null-space) of the local discrepancy gradient $\nabla_z \mathcal{L}_t(z_t^{\mathrm{src}})^\top = -D_t^\top$, satisfying the exact orthogonality condition:

$$(V_\perp^t)^\top D_t = 0. \tag{12}$$

Consequently, we refer to $V_\parallel^t$ as the target transport component and $V_\perp^t$ as the orthogonal null-space residual. To reduce unnecessary trajectory deviations while retaining the generative flexibility provided by the pretrained prior, NSA operates on a modulated velocity proposal $\widehat{V}_\Delta^t$. Formally, we parameterize $\widehat{V}_\Delta^t$ by scaling the orthogonal null-space residual $V_\perp^t$ by a modulation coefficient $\alpha_t \geq 0$ while leaving the target transport component $V_\parallel^t$ intact:

$$\widehat{V}_\Delta^t \triangleq V_\parallel^t + \alpha_t V_\perp^t, \qquad \alpha_t \geq 0. \tag{13}$$

When $\alpha_t = 1$, $\widehat{V}_\Delta^t$ recovers the standard monolithic proposal V_Δ^t; when $\alpha_t = 0$, $\widehat{V}_\Delta^t$ collapses to pure target-directed transport along $V_\parallel^t$. To understand how this modulation affects editing behavior, we analyze its influence from finite-step perspectives.

4.2 Theoretical Properties of the Modulated Proposal

We first evaluate how modulating α_t affects instantaneous progress toward the target landmark z_t^{tar} in continuous time ($\eta \to 0$).

Proposition 2 (First-Order Transport Invariance)**.** *For any modulation factor $\alpha_t \geq 0$, the directional derivative of the local discrepancy $\mathcal{L}_t$ along the modulated proposal $\widehat{V}_\Delta^t$ is strictly invariant to α_t and coincides with the linear progress of the raw proposal V_Δ^t:*

$$\frac{d}{d\epsilon}\mathcal{L}_t(z_t^{\mathrm{src}} + \epsilon \widehat{V}_\Delta^t)\Big|_{\epsilon=0} = -(V_\Delta^t)^\top D_t = -(V_\parallel^t)^\top D_t. \tag{14}$$

Geometric Interpretation. Proposition 2 establishes that $V_\perp^t$ is a genuine degree of freedom of the proposal, its magnitude can be changed without altering the first-order progress toward reducing $\mathcal{L}_t$. In contrast, $V_\parallel^t$ contains the component of V_Δ^t that determines its first-order discrepancy reduction. This decomposition therefore reveals the orthogonal component as a controllable degree of freedom that is locally neutral to the surrogate objective, enabling flexible regulation of generative variations.

While $V_\perp^t$ is transport-neutral in continuous time, actual numerical flow generation requires discrete steps ($\eta > 0$). We then analyze the local discrepancy change induced by an Euler update.

Proposition 3 (Quadratic Drift Effect)**.** *Consider advancing the state z_t^{src} by a finite Euler integration step size $\eta > 0$ along $\widehat{V}_\Delta^t$. The change in local discrepancy $\Delta\mathcal{L}_t(\widehat{V}_\Delta^t) \triangleq \mathcal{L}_t\left(z_t^{\mathrm{src}} + \eta\widehat{V}_\Delta^t\right) - \mathcal{L}_t\left(z_t^{\mathrm{src}}\right)$ expands as:*

$$\Delta\mathcal{L}_t(\widehat{V}_\Delta^t) = -\eta(V_\Delta^t)^\top D_t + \frac{\eta^2}{2}\|V_\parallel^t\|^2 + \underbrace{\frac{\eta^2}{2}\alpha_t^2\|V_\perp^t\|^2}_{\Delta\mathcal{L}_\perp(\alpha_t)}. \tag{15}$$

Thus, the discrete local discrepancy contribution $\Delta\mathcal{L}_\perp(\alpha_t)$ scales quadratically with α_t:

$$\Delta\mathcal{L}_\perp(\alpha_t) = \alpha_t^2 \cdot \Delta\mathcal{L}_\perp(1). \tag{16}$$

Controlling Orthogonal Drift. Propositions 2 and 3 provide the exact mathematical tool needed to prevent the failure mode established in Proposition 1.

Recall from Proposition 1 that a step along raw V_Δ^t diverges ($\Delta\mathcal{L}_t > 0$) whenever the orthogonal energy ratio $\frac{\|V_\perp^t\|^2}{\|D_t\|^2}$ exceeds $\frac{2\gamma}{\eta} - \gamma^2$ (where $V_\parallel^t = \gamma D_t$). By introducing α_t, the effective orthogonal energy becomes $\alpha_t^2\|V_\perp^t\|^2$. Thus, the non-divergence condition ($\Delta\mathcal{L}_t(\widehat{V}_\Delta^t) < 0$) under modulated updates transforms to:

$$\alpha_t^2 \frac{\|V_\perp^t\|^2}{\|D_t\|^2} < \frac{2\gamma}{\eta} - \gamma^2 \implies \alpha_t < \alpha_t^{\max} \triangleq \frac{\|D_t\|}{\|V_\perp^t\|}\sqrt{\frac{2\gamma}{\eta} - \gamma^2}. \tag{17}$$

Equation (17) establishes a profound theoretical result: no matter how noisy or ill-aligned the raw neural velocity proposal V_Δ^t is (i.e., arbitrarily large $\|V_\perp^t\|$), there always exists a positive scaling factor $\alpha_t < \alpha_t^{\max}$ that guarantees local trajectory convergence ($\Delta\mathcal{L}_t < 0$) without sacrificing linear transport.

- When V_Δ^t is noisy ($\|V_\perp^t\| \gg \|V_\parallel^t\|$): Suppressing $\alpha_t \to 0$ eliminates quadratic drift $\mathcal{O}(\alpha_t^2)$, pulling the update safely back inside the convergence bound of Eq. (17).
- When V_Δ^t is well-aligned ($\|V_\perp^t\| \ll \|V_\parallel^t\|$): α_t can safely approach 1, allowing the model to unleash rich generative details from its pretrained prior without risking off-axis drift.

This motivates the need for an automated, adaptive gating mechanism to dynamically compute α_t at every integration step, which we formulate in Section 4.3.

4.3 Null-Space Alignment

Suppressing the orthogonal component $V_\perp^t$ via a modulation factor $\alpha_t < \alpha_t^{\max}$ guarantees local trajectory convergence ($\Delta\mathcal{L}_t < 0$) without sacrificing linear target transport. However, computing $\alpha_t^{\max}$ directly requires knowing the exact alignment scale γ at each step. To operationalize this insight into an adaptive metric during generation, we introduce directional consistency c_t to measure the instantaneous signal-to-noise ratio of the model proposal V_Δ^t. At step t, we measure the geometric alignment between the raw proposal V_Δ^t and the canonical transport vector D_t via cosine similarity:

$$c_t \triangleq \cos\left(\angle(V_\Delta^t, D_t)\right) = \frac{(V_\Delta^t)^\top D_t}{\|V_\Delta^t\|\|D_t\|} \in [-1, 1]. \tag{18}$$

To understand why c_t serves as an ideal gating signal, we analyze its squared value c_t^2 under the orthogonal decomposition $V_\Delta^t = V_\parallel^t + V_\perp^t$.

Proposition 4. *The squared directional consistency c_t^2 equals the exact fraction of proposal kinetic energy concentrated in target-directed transport:*

$$c_t^2 = \frac{\|V_\parallel^t\|^2}{\|V_\parallel^t\|^2 + \|V_\perp^t\|^2} = \frac{\|V_\parallel^t\|^2}{\|V_\Delta^t\|^2}. \tag{19}$$

Proposition 4 provides a clean, bounded measure of proposal reliability $c_t^2 \in [0, 1]$:

- High Consistency ($c_t^2 \to 1 \implies \|V_\perp^t\| \ll \|V_\parallel^t\|$): The model proposal V_Δ^t is dominated by target-directed velocity. Orthogonal noise is negligible, placing the trajectory safely below the divergence boundary of Proposition 1. The pretrained prior is reliable and can be fully trusted.
- Low Consistency ($c_t^2 \to 0 \implies \|V_\perp^t\| \gg \|V_\parallel^t\|$): The proposal energy is overwhelmingly consumed by off-axis noise $V_\perp^t$. This signals high model uncertainty or prompt conflict, threatening to breach $\alpha_t^{\max}$ and trigger trajectory divergence.

To suppress $V_\perp^t$ proportionally when energy alignment is low, we formulate the Null-Space Alignment (NSA) gating coefficient $\alpha_t(c_t)$ based on three key design principles:

- Smooth Confidence Mapping ($\alpha_t \propto c_t^2$): Since Proposition 3 reveals a quadratic sensitivity of orthogonal drift to α_t, we use the squared directional consistency c_t^2 as a smooth confidence signal to modulate the orthogonal component.
- Opposing-Vector Filtering ($\mathrm{clamp}(c_t, 0, 1)$): If $c_t < 0$, the proposal moves strictly away from the target landmark ($V_\parallel^t$ points opposite to D_t). We apply $\mathrm{clamp}(c_t, 0, 1) \triangleq \max(0, c_t)$ to completely zero out generative prior updates during hostile steps.
- Generative Floor Boundary (ρ): To prevent over-constraining the model when subtle generative transitions are desired, we introduce a lower bound $\rho \in [0, 1]$.

Combining these principles yields the final NSA gating function:

$$\alpha_t = \rho + (1 - \rho) \cdot \mathrm{clamp}(c_t, 0, 1)^2, \qquad \rho \in [0, 1]. \tag{20}$$

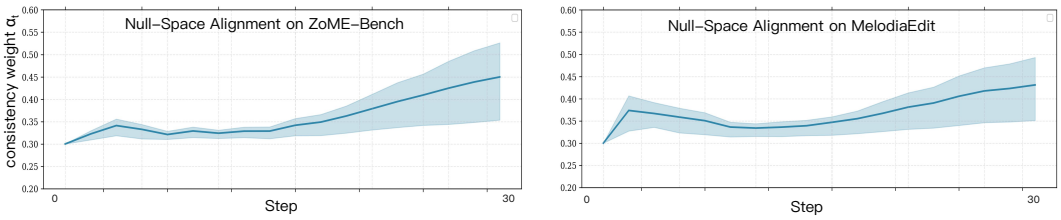

Figure 3: Trends of the coefficient α_t along the flow schedule on ZoME-Bench (left) and MelodiaEdit (right). On both datasets, α_t exhibits an overall increasing trend as generation step decreases, consistent with the design rationale of NSA. The gate stays conservative when the model proposal weakly aligns with the requested change, and opens up as the trajectory stabilizes. Solid lines denote the mean over runs, shaded regions one standard deviation.

Method	$\text{FAD}_{\text{CLAP}}\downarrow$	$\text{FAD}_{\text{Vggish}}\downarrow$	CQT1-PCC↑	LPAPS↓	CLAP↑	MOS-P↑	MOS-T↑	Overall↑
MusicMagus	0.593	7.631	0.338	5.243	0.238	3.125	2.547	2.673
ZETA	0.509	3.380	0.293	5.458	0.252	2.314	2.436	2.145
SteerMusic	<u>0.278</u>	<u>2.426</u>	<u>0.480</u>	<u>3.772</u>	0.259	<u>3.878</u>	3.030	<u>3.454</u>
RF-solver	0.352	6.171	0.324	5.379	0.278	2.654	3.117	2.899
FireFlow	0.302	6.101	0.330	5.354	<u>0.285</u>	2.750	<u>3.222</u>	3.006
Ours	**0.106**	**2.190**	**0.542**	**3.570**	**0.334**	**4.230**	**3.576**	**3.692**

Table 1: Comparison of text-guided music editing methods on the ZoME-Bench dataset.

At each Euler integration step, NSA computes the target transport component $V_{\parallel}^t$, isolates the null-space residual $V_{\perp}^t$, evaluates the directional consistency c_t, and updates the latent state along the recombined velocity field $\widehat{V}_{\Delta}^t$:

$$\widehat{V}_{\Delta}^t = V_{\parallel}^t + \left[\rho + (1-\rho)\cdot \text{clamp}\left(\frac{(V_{\Delta}^t)^\top D_t}{\|V_{\Delta}^t\|\|D_t\|}, 0, 1\right)^2\right] V_{\perp}^t. \tag{21}$$

Equation (21) completes the Null-Space Alignment framework. By dynamically dampening $V_{\perp}^t$ when proposals are noisy ($c_t \to 0$) and restoring $V_{\perp}^t$ when proposals are coherent ($c_t \to 1$), NSA achieves optimal structural preservation while preserving full generative editing power.

Empirical Verification of Gating Dynamics. Figure 3 tracks the trajectory evolution of α_t across generation runs on MelodiaEdit and ZoME-Bench. Matching our theoretical design, α_t remains conservatively suppressed near ρ at early flow steps where V_{Δ}^t is weakly aligned with D_t, effectively mitigating early off-axis drift $\mathcal{O}(\alpha_t^2)$. As the trajectory stabilizes, $c_t \to 1$ and α_t smoothly unclamps toward 1, enabling full generative detail synthesis from the pretrained prior without compromising structural alignment. Appendix A.5 shows the algorithms in detail.

5 Experiment

5.1 Experimental Setups

Implementation Details. We build our framework upon a pretrained flow-based text-to-audio generation model, Stable Audio 3.0 Evans et al. (2026), without any additional training or parameter optimization. All experiments are conducted using a single Nvidia L40 GPU.

Dataset. We evaluate on ZoME-Bench Liu et al. (2024b) and MelodiaEdit Yang et al. (2026). ZoME-Bench contains 1,000 ten-second audio samples from MusicCaps Agostinelli et al. (2023), each paired with source and target text prompts describing the original content and desired modification, providing diverse cases for assessing both instruction following and source audio preservation. MelodiaEdit targets different aspects of music editing and covers three categories, Instrument, Style, and Mood, with 1,090 synthetic and 925 real-world editing pairs in total.

Category	Metric	MusicMagus	ZETA	SteerMusic	RF-solver	FireFlow	Ours
Instrument	$FAD_{CLAP}\downarrow$	0.549	0.493	0.408	0.389	<u>0.360</u>	**0.252**
	$FAD_{Vggish}\downarrow$	<u>3.490</u>	3.793	4.732	6.244	4.311	**2.311**
	CQT1-PCC↑	0.205	<u>0.398</u>	0.289	0.179	0.172	**0.365**
	LPAPS↓	**3.124**	3.922	3.743	5.005	5.317	<u>3.556</u>
	CLAP↑	0.208	0.189	0.209	<u>0.245</u>	0.222	**0.283**
Style	$FAD_{CLAP}\downarrow$	0.548	<u>0.259</u>	<u>0.259</u>	0.328	0.338	**0.195**
	$FAD_{Vggish}\downarrow$	3.687	<u>2.759</u>	3.874	4.145	4.092	**2.551**
	CQT1-PCC↑	0.201	0.209	0.130	0.209	<u>0.216</u>	**0.537**
	LPAPS↓	4.717	3.310	<u>2.871</u>	4.911	4.817	**2.760**
	CLAP↑	0.199	0.240	<u>0.260</u>	0.236	0.245	**0.291**
Mood	$FAD_{CLAP}\downarrow$	0.259	0.276	<u>0.206</u>	0.319	0.312	**0.168**
	$FAD_{Vggish}\downarrow$	3.330	<u>1.657</u>	2.080	5.091	4.019	**1.629**
	CQT1-PCC↑	0.165	<u>0.241</u>	0.129	0.205	0.217	**0.636**
	LPAPS↓	4.105	3.942	**3.250**	5.323	5.321	<u>3.583</u>
	CLAP↑	0.257	0.236	0.214	0.266	<u>0.276</u>	**0.306**

Table 2: Comparison of semantic music editing methods on different categories of MelodiaEdit.

Evaluation Metrics. Following SteerMusic Niu et al. (2026), we adopt four objective metrics: Frechet Audio Distance (FAD) Kilgour et al. (2019), Learned Perceptual Audio Patch Similarity (LPAPS) Iashin & Rahtu (2021), Top-1 Constant-Q Transform Pearson Correlation Coefficient (CQT1-PCC, PCC) Brown (1991), and CLAP Score Elizalde et al. (2023). FAD measures the distributional distance between edited audio and reference audio distributions in the embedding space. LPAPS evaluates perceptual similarity for source preservation. CQT1-PCC measures melody consistency between source and edited music. CLAP Score assesses semantic consistency between edited audio and the target instruction.

Baselines. We compare our method against representative inversion-based music editing methods, including ZETA Manor & Michaeli (2024), MusicMagus Zhang et al. (2024b), and SteerMusic Niu et al. (2026), which are based on the AudioLDM2 backbone Liu et al. (2024a). To ensure a fair comparison with flow-based models, We further adapt the inversion methods Rf-solver Wang et al. (2024) and FireFlow Deng et al. (2025) to the Stable Audio 3 Medium backbone and include them as additional baselines.

5.2 Experimental results

Quantitative comparison on ZoME-Bench. Table 1 compares our method with representative music editing approaches. As shown in Table 1, our method consistently achieves the best performance across all five metrics, demonstrating superior audio quality, source music preservation, and target instruction alignment. In particular, compared with SteerMusic, our method reduces LPAPS from 3.772 to 3.570 and improves CLAP from 0.259 to 0.334. Moreover, our method outperforms both RF-solver and FireFlow, suggesting that the proposed directional alignment effectively improves flow-based inversion-free music editing.

Quantitative comparison on MelodiaEdit. Table 2 reports results on MelodiaEdit across Instrument, Style, and Mood categories. Our method consistently achieves competitive performance against all baselines across diverse editing scenarios, demonstrating its ability to handle different types of musical transformations. By adaptively regulating the contribution of the orthogonal component, our approach strikes a better balance between preserving the original musical structure and following the target instruction. These results verify the effectiveness of NSA as a general training-free and inversion-free strategy for music editing across various musical attributes.

Subjective comparison. Figure 4 shows spectrogram comparisons with SOTA methods. We further conduct a Mean Opinion Score (MOS) study, reported in Table 1. Twenty participants rated 100 randomly selected edited results on a 5-point scale (1: Bad, 5: Excellent) along three criteria: preservation of the original audio (MOS-P), adherence to the target instruction (MOS-T), and overall perceptual quality (Overall). Participants were presented with randomly shuffled anonymous audio

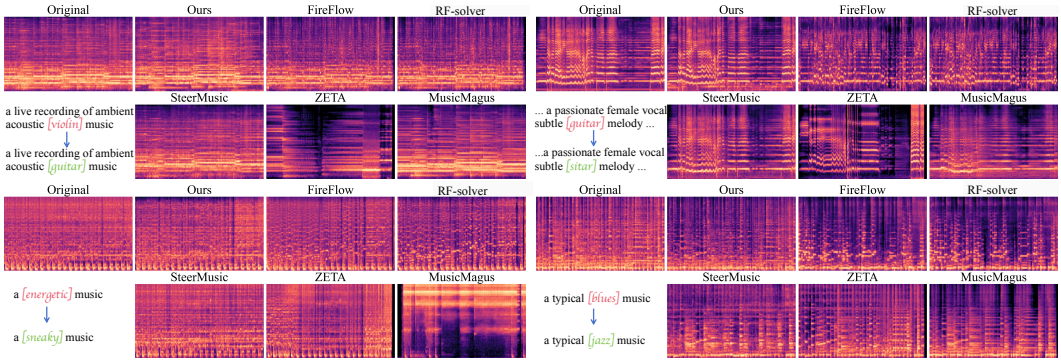

Figure 4: Visual comparison of edited audio. In addition to faithfully following the editing prompts, our method better preserves the overall spectral structure of the input audio.

Method	$FAD_C\downarrow$	$FAD_V\downarrow$	PCC↑	LPAPS↓	CLAP↑
FlowEdit ($\alpha = 1$)	0.142	3.328	0.149	5.517	**0.336**
w/ $\alpha = 0.3$	0.112	2.362	0.460	3.880	0.246
w/ NSA ($\rho = 0.3$)	**0.106**	**2.190**	**0.542**	**3.570**	0.334

Table 3: Ablation study of orthogonal component regulation in NSA on ZoME-Bench.

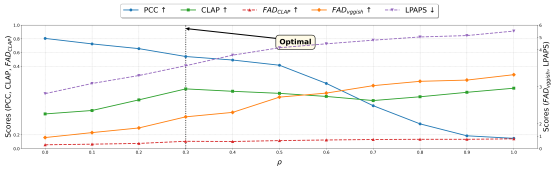

Figure 5: Ablation study on the effect of ρ in NSA for text-guided music editing on ZoME-Bench.

pairs without method identifiers. Our method achieves the best performance on all three criteria, indicating improved editing fidelity and more perceptually convincing results.

Ablation study. We first study the effect of the minimum null-space retention weight ρ. As shown in Figure 5, a smaller ρ makes NSA more aggressively suppress the null-space component when the model proposal conflicts with the editing objective, improving source preservation but reducing generative flexibility. A larger ρ retains more prior-driven variation, enhancing semantic diversity but potentially deviating from the source. We find that $\rho = 0.3$ achieves a favorable trade-off between source preservation and editing effectiveness, maintaining content consistency while allowing sufficient semantic modification.

We further compare different strategies for regulating the orthogonal component in Table 3. FlowEdit ($\alpha = 1$) directly preserves the full orthogonal component without explicit regulation, which provides good generative flexibility but may introduce unnecessary deviations from the source content. Moreover, to demonstrate the effect of the adaptive gating mechanism, we compare NSA with a fixed-weight strategy ($\alpha = 0.3$) under the same adaptive setting ($\rho = 0.3$). The fixed strategy reduces unnecessary deviations by uniformly suppressing the orthogonal contribution throughout the generation process, but lacks the ability to adapt to different editing states. In contrast, NSA dynamically adjusts the orthogonal component according to directional consistency, retaining useful prior-driven variations while suppressing conflicting deviations. This adaptive regulation achieves a better balance between source preservation and generative flexibility.

6 Conclusion

In this paper, we propose the Null-Space Alignment for inversion-free text-guided music editing. The difference between source- and target-conditioned flow fields is a model-induced generative proposal, not a guaranteed editing direction. NSA decomposes it into an objective-effective component and a first-order null-space component, regulating the latter by directional agreement between the proposal and the current transition, preserving generative freedom while reducing unsupported deviations. The procedure requires no additional training and provides a principled way to combine local objective guidance with the prior of pretrained flow models.

AI USE STATEMENT

Generative AI tools were used for two purposes: language polishing and literature search assistance. AI-assisted text was reviewed and edited by the authors to preserve meaning and avoid fabricated or misattributed content. All references cited in this paper were independently verified by the authors against the original publications. AI tools were not used to generate bibliographic entries or to attribute claims without verification. AI tools were not used for research ideation, problem formulation, experimental design, data collection, data analysis, result interpretation, or figure or algorithm design.

The authors take full responsibility for the final content of this work, including all claims, results, figures, code, and artifacts, irrespective of whether any portion of the text or references was produced or revised with the aid of AI.

REPRODUCIBILITY STATEMENT

We provide a reproducibility statement to clarify the efforts made to ensure the reproducibility of our work. The source code is anonymously available in the supplementary materials. All datasets used in our experiments are publicly available. Complete proofs of all theoretical results are provided in Appendix.

REFERENCES

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267, 2025.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zach Evans, Julian D Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3. *arXiv preprint arXiv:2605.17991*, 2026.

Zhongyuan Fu, Yuhang Jia, Hui Wang, Pengjun Chen, Jian Gao, Cun Liu, Wenjia Zeng, Yong Chen, and Yong Qin. Echoedit: Stabilizing inversion-free audio editing via optimal transport geometry. *arXiv preprint arXiv:2606.15149*, 2026.

Zhengkun Ge, Xiaoqian Liu, Haoran Zhang, Yuan Ge, Junxiang Zhang, Zhengtao Yu, Jingbo Zhu, and Tong Xiao. Directaudioedit: Inversion-free text-guided audio editing via diffusion prediction contrast. *arXiv preprint arXiv:2606.07356*, 2026.

Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. Instructme: an instruction guided music edit framework with latent diffusion models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5835–5843, 2024.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Vladimir Iashin and Esa Rahtu. Taming Visually Guided Sound Generation. *arXiv e-prints*, art. arXiv:2110.08791, October 2021. doi: 10.48550/arXiv.2110.08791.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech 2019*, pp. 2350–2354, 2019.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Yue Lei, Siqi Yang, Ting Zhong, and Fan Zhou. Harmonic canvas: Inversion-free editing for visually-guided music style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29727–29737, June 2026.

Haowen Li, Tianxiang Li, Yi Yang, Boyu Cao, and Qi Liu. Polyphonia: Zero-shot timbre transfer in polyphonic music with acoustic-informed attention calibration. *arXiv preprint arXiv:2605.10203*, 2026.

Liwei Lin, Gus Xia, Yixiao Zhang, and Junyan Jiang. Arrange, inpaint, and refine: Steerable long-term music audio generation and editing via content-based controls. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7690–7698, 8 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024a.

Huadai Liu, Jialei Wang, Xiangtai Li, Rongjie Huang, Yang Liu, Jiayang Xu, and Zhou Zhao. Medic: Zero-shot music editing with disentangled inversion control. *arXiv preprint arXiv:2407.13220*, 2024b.

Huadai Liu, Jialei Wang, Xiangtai Li, Wen Wang, Qian Chen, Rongjie Huang, Yang Liu, Jiayang Xu, Zhou Zhao, and Wei Xue. Melodyedit: Zero-shot music editing with disentangled inversion control. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10083–10092, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, pp. 34603–34629. JMLR.org, 2024.

Xinlei Niu, Kin Wai Cheuk, Jing Zhang, Naoki Murata, Chieh-Hsin Lai, Michele Mancusi, Woosung Choi, Giorgio Fabbro, Wei-Hsiang Liao, Charles Patrick Martin, et al. Steermusic: Enhanced musical consistency for zero-shot text-guided and personalized music editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2000–2010, 2026.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Manjie Xu, Chenxing Li, Duzhen Zhang, Dan Su, Wei Liang, and Dong Yu. Prompt-guided precise audio editing with diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 55126–55143, 2024.

Chenyu Yang, Shuai Wang, Hangting Chen, Jianwei Yu, Wei Tan, Rongzhi Gu, Yaoxun Xu, Yizhi Zhou, Haina Zhu, and Haizhou Li. Songeditor: Adapting zero-shot song generation language model as a multi-task editor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 25597–25605, 2025.

Y Yang, H Li, T Li, B Cao, X Zhang, L Chen, and Q Liu. Melodia: Training-Free music editing guided by attention probing in diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40:2209–2217, 2026.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning. *arXiv preprint arXiv:2405.18386*, 2024a.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*, 2024b.

Ge Zhu, Yutong Wen, and Zhiyao Duan. Audio generation through score-based generative modeling: design principles and implementation. *Foundations and Trends® in Signal Processing*, 20(1):1–82, 2026.

A Appendix

A.1 Proof of Proposition 1 (When Does V_Δ^t Fail)

Proof. Recall the local discrepancy metric $\mathcal{L}_t(z) \triangleq \frac{1}{2}\|z - z_t^{\mathrm{tar}}\|^2$ and let $D_t \triangleq z_t^{\mathrm{tar}} - z_t^{\mathrm{src}}$ denote the target transport vector. Advancing the state z_t^{src} by an Euler integration step $\eta > 0$ along the proposal V_Δ^t yields $z_{t+\eta} = z_t^{\mathrm{src}} + \eta V_\Delta^t$. The resulting change in discrepancy expands as:

$$\begin{aligned}
\Delta\mathcal{L}_t(V_\Delta^t) &\triangleq \mathcal{L}_t(z_{t+\eta}) - \mathcal{L}_t(z_t^{\mathrm{src}}) \\
&= \frac{1}{2}\left\|(z_t^{\mathrm{src}} + \eta V_\Delta^t) - z_t^{\mathrm{tar}}\right\|^2 - \frac{1}{2}\|z_t^{\mathrm{src}} - z_t^{\mathrm{tar}}\|^2 \\
&= \frac{1}{2}\left\|\eta V_\Delta^t - D_t\right\|^2 - \frac{1}{2}\|D_t\|^2 \\
&= -\eta (V_\Delta^t)^\top D_t + \frac{\eta^2}{2}\|V_\Delta^t\|^2. \qquad (22)
\end{aligned}$$

Under the orthogonal decomposition $V_\Delta^t = V_\parallel^t + V_\perp^t = \gamma D_t + V_\perp^t$ with $(V_\perp^t)^\top D_t = 0$, the scalar components in Eq. (22) satisfy:

$$(V_\Delta^t)^\top D_t = (\gamma D_t + V_\perp^t)^\top D_t = \gamma\|D_t\|^2, \qquad (23)$$

$$\|V_\Delta^t\|^2 = \|\gamma D_t + V_\perp^t\|^2 = \gamma^2\|D_t\|^2 + \|V_\perp^t\|^2. \qquad (24)$$

Substituting these identities into Eq. (22) yields the exact discrete discrepancy change:

$$\Delta\mathcal{L}_t(V_\Delta^t) = -\eta\gamma\|D_t\|^2 + \frac{\eta^2}{2}\left(\gamma^2\|D_t\|^2 + \|V_\perp^t\|^2\right). \qquad (25)$$

Local trajectory failure occurs when a discrete step fails to decrease discrepancy ($\Delta\mathcal{L}_t(V_\Delta^t) > 0$), which requires:

$$\frac{\eta^2}{2}\|V_\perp^t\|^2 > \eta\gamma\|D_t\|^2 - \frac{\eta^2}{2}\gamma^2\|D_t\|^2 = \left(\eta\gamma - \frac{\eta^2}{2}\gamma^2\right)\|D_t\|^2. \tag{26}$$

Dividing both sides by $\frac{\eta^2}{2}\|D_t\|^2 > 0$ isolates the relative orthogonal energy ratio:

$$\frac{\|V_\perp^t\|^2}{\|D_t\|^2} > \frac{\eta\gamma - \frac{\eta^2}{2}\gamma^2}{\frac{\eta^2}{2}} = \frac{2\gamma}{\eta} - \gamma^2, \tag{27}$$

which completes the proof. □

A.2 Proof of Proposition 2 (First-Order Transport Invariance)

Proof. Recall the local discrepancy metric $\mathcal{L}_t(z) \triangleq \frac{1}{2}\|z - z_t^{\mathrm{tar}}\|^2$ evaluated at state $z \in \mathbb{R}^d$, with target transport vector $D_t \triangleq z_t^{\mathrm{tar}} - z_t^{\mathrm{src}}$. The spatial gradient of $\mathcal{L}_t$ evaluated at the current state z_t^{src} is:

$$\nabla_z\mathcal{L}_t(z_t^{\mathrm{src}}) = z_t^{\mathrm{src}} - z_t^{\mathrm{tar}} = -D_t. \tag{28}$$

By applying the chain rule, the instantaneous directional derivative of $\mathcal{L}_t$ along the modulated velocity proposal $\widehat{V}_\Delta^t \triangleq V_\parallel^t + \alpha_t V_\perp^t$ is given by the inner product of the gradient with $\widehat{V}_\Delta^t$:

$$\begin{aligned}\left.\frac{d}{d\epsilon}\mathcal{L}_t(z_t^{\mathrm{src}} + \epsilon\widehat{V}_\Delta^t)\right|_{\epsilon=0} &= \nabla_z\mathcal{L}_t(z_t^{\mathrm{src}})^\top\widehat{V}_\Delta^t \\ &= -D_t^\top\left(V_\parallel^t + \alpha_t V_\perp^t\right) \\ &= -D_t^\top V_\parallel^t - \alpha_t\left(D_t^\top V_\perp^t\right).\end{aligned} \tag{29}$$

By definition of the orthogonal decomposition $V_\Delta^t = V_\parallel^t + V_\perp^t$, the residual component $V_\perp^t$ satisfies the strict orthogonality condition $(V_\perp^t)^\top D_t = 0$. Substituting this identity into Eq. (29) causes the term dependent on α_t to vanish identically:

$$\left.\frac{d}{d\epsilon}\mathcal{L}_t(z_t^{\mathrm{src}} + \epsilon\widehat{V}_\Delta^t)\right|_{\epsilon=0} = -D_t^\top V_\parallel^t - \alpha_t(0) = -(V_\parallel^t)^\top D_t. \tag{30}$$

Finally, evaluating the linear inner product of the raw unmodulated proposal V_Δ^t with D_t yields:

$$(V_\Delta^t)^\top D_t = (V_\parallel^t + V_\perp^t)^\top D_t = (V_\parallel^t)^\top D_t + (V_\perp^t)^\top D_t = (V_\parallel^t)^\top D_t. \tag{31}$$

Thus, for any modulation factor $\alpha_t \geq 0$:

$$\left.\frac{d}{d\epsilon}\mathcal{L}_t(z_t^{\mathrm{src}} + \epsilon\widehat{V}_\Delta^t)\right|_{\epsilon=0} = -(V_\Delta^t)^\top D_t = -(V_\parallel^t)^\top D_t, \tag{32}$$

which proves that instantaneous target transport progress is strictly invariant to α_t. □

A.3 Proof of Proposition 3 (Quadratic Drift Control)

Proof. Let $z_t^{\mathrm{src}} \in \mathbb{R}^d$ be the current state, $z_t^{\mathrm{tar}} \in \mathbb{R}^d$ be the target landmark, and $D_t \triangleq z_t^{\mathrm{tar}} - z_t^{\mathrm{src}}$ denote the target transport vector. Advancing z_t^{src} by a finite Euler integration step size $\eta > 0$ along the gated proposal $\widehat{V}_\Delta^t \triangleq V_\parallel^t + \alpha_t V_\perp^t$ yields the updated state $z_{t+\eta} = z_t^{\mathrm{src}} + \eta\widehat{V}_\Delta^t$.

Evaluating the local discrepancy metric $\mathcal{L}_t(z) \triangleq \frac{1}{2}\|z - z_t^{\mathrm{tar}}\|^2$ at $z_{t+\eta}$ gives:

$$\begin{aligned}\mathcal{L}_t(z_{t+\eta}) &= \frac{1}{2}\left\|(z_t^{\mathrm{src}} + \eta\widehat{V}_\Delta^t) - z_t^{\mathrm{tar}}\right\|^2 \\ &= \frac{1}{2}\left\|\eta\widehat{V}_\Delta^t - D_t\right\|^2 \\ &= \frac{1}{2}\|D_t\|^2 - \eta D_t^\top\widehat{V}_\Delta^t + \frac{\eta^2}{2}\|\widehat{V}_\Delta^t\|^2.\end{aligned} \tag{33}$$

Subtracting $\mathcal{L}_t(z_t^{\mathrm{src}}) = \frac{1}{2}\|D_t\|^2$ yields the net change in local discrepancy:

$$\Delta\mathcal{L}_t(\widehat{V}_\Delta^t) \triangleq \mathcal{L}_t(z_{t+\eta}) - \mathcal{L}_t(z_t^{\mathrm{src}}) = -\eta D_t^\top \widehat{V}_\Delta^t + \frac{\eta^2}{2}\|\widehat{V}_\Delta^t\|^2. \tag{34}$$

We simplify the two terms on the right-hand side of Eq. (34) as follows:

1. By Proposition 2, the linear transport contribution is invariant to α_t:
$$D_t^\top \widehat{V}_\Delta^t = D_t^\top (V_\parallel^t + \alpha_t V_\perp^t) = D_t^\top V_\parallel^t + \alpha_t (D_t^\top V_\perp^t) = D_t^\top V_\Delta^t. \tag{35}$$

2. By definition of the orthogonal decomposition, $V_\parallel^t$ and $V_\perp^t$ are orthogonal, i.e., $(V_\parallel^t)^\top V_\perp^t = 0$. By the Pythagorean theorem, the squared norm expands to:
$$\|\widehat{V}_\Delta^t\|^2 = \|V_\parallel^t + \alpha_t V_\perp^t\|^2 = \|V_\parallel^t\|^2 + \alpha_t^2 \|V_\perp^t\|^2. \tag{36}$$

Substituting these two identities back into Eq. (34) yields the exact discrete discrepancy expansion:

$$\Delta\mathcal{L}_t(\widehat{V}_\Delta^t) = -\eta (V_\Delta^t)^\top D_t + \frac{\eta^2}{2}\|V_\parallel^t\|^2 + \frac{\eta^2}{2}\alpha_t^2 \|V_\perp^t\|^2, \tag{37}$$

which establishes Eq. (15).

Isolating the orthogonal drift component $\Delta\mathcal{L}_\perp(\alpha_t) \triangleq \frac{\eta^2}{2}\alpha_t^2\|V_\perp^t\|^2$, we observe that for the unmodulated monolithic proposal ($\alpha_t = 1$), the baseline drift is $\Delta\mathcal{L}_\perp(1) = \frac{\eta^2}{2}\|V_\perp^t\|^2$. Consequently:

$$\Delta\mathcal{L}_\perp(\alpha_t) = \alpha_t^2 \cdot \left(\frac{\eta^2}{2}\|V_\perp^t\|^2\right) = \alpha_t^2 \cdot \Delta\mathcal{L}_\perp(1), \tag{38}$$

which proves that discrete off-axis trajectory drift scales quadratically with the gating factor α_t. □

A.4 Proof of Proposition 4 (Energy Ratio Interpretation of Directional Consistency)

Proof. Recall from Eq. (18) that the directional consistency c_t between the monolithic proposal V_Δ^t and the target transport vector $D_t \triangleq z_t^{\mathrm{tar}} - z_t^{\mathrm{src}}$ is defined as the cosine similarity:

$$c_t \triangleq \frac{(V_\Delta^t)^\top D_t}{\|V_\Delta^t\|\|D_t\|}. \tag{39}$$

Squaring both sides of Eq. (39) yields:

$$c_t^2 = \frac{\left((V_\Delta^t)^\top D_t\right)^2}{\|V_\Delta^t\|^2\|D_t\|^2}. \tag{40}$$

We analyze the numerator and denominator of Eq. (40) separately using the orthogonal decomposition $V_\Delta^t = V_\parallel^t + V_\perp^t$:

1. **Numerator Simplification:** Substituting the orthogonal decomposition into the inner product and exploiting the orthogonality condition $(V_\perp^t)^\top D_t = 0$ gives:
$$(V_\Delta^t)^\top D_t = (V_\parallel^t + V_\perp^t)^\top D_t = (V_\parallel^t)^\top D_t + (V_\perp^t)^\top D_t = (V_\parallel^t)^\top D_t. \tag{41}$$
By definition of the parallel projection $V_\parallel^t \triangleq \frac{(V_\Delta^t)^\top D_t}{\|D_t\|^2} D_t$, the vector $V_\parallel^t$ is collinear with D_t. Therefore, the absolute inner product satisfies $|(V_\parallel^t)^\top D_t| = \|V_\parallel^t\|\|D_t\|$. Squaring this quantity yields:
$$\left((V_\Delta^t)^\top D_t\right)^2 = \left((V_\parallel^t)^\top D_t\right)^2 = \|V_\parallel^t\|^2\|D_t\|^2. \tag{42}$$

2. **Denominator Simplification:** Since $V_{\parallel}^t$ and $V_{\perp}^t$ are orthogonal (i.e., $(V_{\parallel}^t)^\top V_{\perp}^t = 0$), the Pythagorean theorem gives:

$$\|V_{\Delta}^t\|^2 = \|V_{\parallel}^t + V_{\perp}^t\|^2 = \|V_{\parallel}^t\|^2 + \|V_{\perp}^t\|^2. \tag{43}$$

Thus, the denominator of Eq. (40) expands to:

$$\|V_{\Delta}^t\|^2\|D_t\|^2 = \left(\|V_{\parallel}^t\|^2 + \|V_{\perp}^t\|^2\right)\|D_t\|^2. \tag{44}$$

Substituting Eq. (42) and Eq. (44) back into Eq. (40) yields:

$$c_t^2 = \frac{\|V_{\parallel}^t\|^2\|D_t\|^2}{\left(\|V_{\parallel}^t\|^2 + \|V_{\perp}^t\|^2\right)\|D_t\|^2}. \tag{45}$$

Canceling the non-zero norm term $\|D_t\|^2 > 0$ from the numerator and denominator gives:

$$c_t^2 = \frac{\|V_{\parallel}^t\|^2}{\|V_{\parallel}^t\|^2 + \|V_{\perp}^t\|^2}. \tag{46}$$

Finally, substituting identity Eq. (43) into the denominator yields the equivalent form:

$$c_t^2 = \frac{\|V_{\parallel}^t\|^2}{\|V_{\Delta}^t\|^2}, \tag{47}$$

which proves that c_t^2 measures the exact proportion of proposal kinetic energy allocated to target-directed motion. □

A.5 Algorithm

The schedule is divided into three phases by $n_{\max}$ and $n_{\min}$. Steps with $N - i > n_{\max}$ are skipped because both conditional velocities are dominated by noise and their difference is uninformative. Steps with $N - n_{\max} \le N - i < N - n_{\min}$ perform the velocity-difference update, where n_{avg} noise draws are averaged to reduce Monte Carlo variance and separate CFG scales s_{src}, s_{tar} are applied to the two passes. The final $n_{\min}$ steps revert to plain target-conditioned sampling, initialised once from the coupled state, to clean up residual noise and place the output near the target-conditioned data manifold. The tolerance ϵ guards the decomposition when $\|d_z\|$ is numerically small. In all experiments we use $\epsilon = 10^{-8}$, $n_{\text{avg}} = 1$, $n_{\max} = 20$, $n_{\min} = 2$, $s_{\text{src}} = 1.5$, $s_{\text{tar}} = 4.5$, $\rho = 0.3$, and $\gamma = 2.0$.

Algorithm 1 Null-Space Alignment for Semantic Music Editing

Require: x_s: source audio latent; $\psi^{\mathrm{src}}, \psi^{\mathrm{tar}}$: source / target conditions; v_θ: flow-matching model, callable as $v_\theta(z, t, \psi, \mathrm{cfg})$; $\{\sigma_i\}_{i=0}^{N}$: descending schedule $[1.0, \ldots, 0.0]$; hyper-parameters $n_{\mathrm{avg}}, n_{\max}, n_{\min}, s_{\mathrm{src}}, s_{\mathrm{tar}}, \rho, \gamma, \epsilon$

Ensure: $\hat{x}$: edited audio latent

$z \leftarrow x_s$ // edit trajectory, initialised at the source
for $i = 0, 1, \ldots, N-1$ **do**
  if $N - i > n_{\max}$ **then**
    continue // skip highest-sigma steps
  end if
  $t \leftarrow \sigma_i$; $\Delta t \leftarrow \sigma_{i+1} - \sigma_i$ // $\Delta t < 0$
  if $N - i > n_{\min}$ **then**
    $\varepsilon \sim \mathcal{N}(0, I)$; $z^{\mathrm{src}} \leftarrow (1-t)\, x_s + t\, \varepsilon$; $z^{\mathrm{tar}} \leftarrow z + z^{\mathrm{src}} - x_s$ // forward-noise source, couple edit to it
    $V_\Delta \leftarrow \frac{1}{n_{\mathrm{avg}}} \sum_k \big(v_\theta(z^{\mathrm{tar}}, t, \psi^{\mathrm{tar}}, s_{\mathrm{tar}}) - v_\theta(z^{\mathrm{src}}, t, \psi^{\mathrm{src}}, s_{\mathrm{src}}) \big)$ // averaged velocity difference
    $d_z \leftarrow z^{\mathrm{tar}} - z^{\mathrm{src}}$ // accumulated transport direction
    if $\|d_z\| > \epsilon$ **then**
      $u \leftarrow d_z / \|d_z\|$; $c \leftarrow (V_\Delta \cdot u) / \|V_\Delta\|$
      $V_\Delta \leftarrow (V_\Delta \cdot u)\, u + \big[\rho + (1-\rho)\, \mathrm{clamp}(c, 0, 1)^2\big] \big(V_\Delta - (V_\Delta \cdot u)\, u\big)$ // keep parallel, gate perpendicular
    end if
    $z \leftarrow z + \Delta t\, V_\Delta$ // integrate the edit
  else
    if $N - i = n_{\min}$ **then**
      $\varepsilon \sim \mathcal{N}(0, I)$; $z^{\mathrm{src}} \leftarrow (1-t)\, x_s + t\, \varepsilon$; $z \leftarrow z + z^{\mathrm{src}} - x_s$ // couple once at handoff to clean-up
    end if
    $z \leftarrow z + \Delta t\, v_\theta(z, t, \psi^{\mathrm{tar}}, s_{\mathrm{tar}})$ // final steps: plain target sampling
  end if
end for
return z