CHANGANYA NDIMI: CODE-SWITCHED SPEECH GENERATION WITH A DIFFUSION PRIOR AND LINGUISTIC CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We generate code-switched speech by minimally editing monolingual utterances with a pretrained diffusion model guided by two differentiable modules: a multilingual LID classifier (where/how much to switch) and a contrastive segment encoder (what to insert). The method requires no parallel code-switched data—constraints steer retrieval and replacement of foreign-language segments while iterative denoising preserves fluency and speaker identity. On a semantically aligned corpus spanning five African languages, it achieves strong scores (segment-level COMET 0.815, LaBSE 0.880) and robust speaker consistency (EER 6.7%). The system reproduces natural code-switching patterns—frequency, placement, alternation—without explicit supervision and, to our knowledge, is the first to support controlled multilingual infusion within a single utterance. These results position guided diffusion as a flexible, plug-and-play approach for multilingual, low-resource speech generation. Audio samples: github.com/codeSwitchLugha/CodeSwitch.

1 INTRODUCTION

Code-switching (CS)—the alternation of languages within an utterance—is pervasive in African speech communities Biswas et al. (2022); Sitaram et al. (2019). Yet modern speech systems (ASR, S2ST, SLU, speech LLMs) remain largely monolingual because high-quality CS corpora are scarce; collecting spontaneous CS audio is costly and often yields unnatural interactions Tarunesh et al. (2021); Hsu et al. (2023). Speech-level *synthesis* of CS is especially under-explored.

We propose a retrieval-augmented, constrained denoising diffusion model (DDPM) that transforms monolingual audio into realistic code-switched utterances via minimal, semantically coherent edits. Starting from noise, the model iteratively denoises while two differentiable controls guide *where* and *how* to infuse foreign segments:

1. $c_1(x, y)$: a Language Identification (LID)-based controller that decides *where* and *how much* to switch;
2. $c_2(x, y)$: a retrieval-based controller that uses a multilingual encoder to find semantically matched foreign-language segments and *blend* them into selected spans with a time-ramped coefficient (blend-and-write-back), subject to prosody/context checks.

This realizes sampling from

$$p(x|y) \;\propto\; p(x)\, c_1(x, y)\, c_2(x, y), \tag{1}$$

where $p(x)$ is a pretrained DDPM prior and c_1, c_2 provide differentiable guidance during *mode-seeking* inference (see §2, App. A). The infusion specification is $y = (\mathcal{L}_{\text{inf}}, \pi_{\text{switch}}, \phi_{\text{src}}, \mathcal{R})$, encoding the infusion language set, a prior over switch placement/amount, source semantics (text or embeddings), and the retrieval index. We apply guidance by minimizing a free-energy bound in which $-\log C(x, y) = -\log c_1 - \log c_2$ is time-ramped during inference (§2.2).

Contributions. (1) A plug-and-play, retrieval-augmented diffusion framework for CS *speech* that requires no parallel CS data. (2) Two complementary controls—LID-based switching (c_1) and semantic, prosody-aware segment infusion (c_2)—with explicit blend-and-write-back and late-commit

schedules. (3) An evaluation suite for semantic fidelity, speaker consistency, and CS structure, plus human judgments. (4) Empirical results on multiple African languages showing fluent, semantically aligned, and speaker-consistent synthesis, enabling downstream use in low-resource ASR, S2ST, and speech LLM training.

2 Problem Formulation

We aim to sample from a *constrained* posterior $p(x \mid y)$ for code-switched speech, where x is a high-dimensional utterance (waveform) and y is the *infusion specification*:

$$y = \big(\mathcal{L}_{\text{inf}},\, \pi_{\text{switch}},\, \phi_{\text{src}},\, \mathcal{R}\big),$$

with $\mathcal{L}_{\text{inf}}$ the infusion language(s), π_{switch} a prior over *where/how much* to switch, ϕ_{src} source semantics (text or embeddings), and $\mathcal{R}$ a retrieval index. We impose two differentiable soft constraints: $c_1(x, y)$ (switch plausibility: *where/how much*) and $c_2(x, y)$ (semantic consistency: *what* to infuse), and write $C(x, y) = c_1(x, y)\, c_2(x, y)$. For clarity, the constraints consume (overlapping) subsets of y:

$$c_1(x, y) \equiv c_1\big(x, \mathcal{L}_{\text{inf}}, \pi_{\text{switch}}\big), \qquad c_2(x, y) \equiv c_2\big(x, \mathcal{L}_{\text{inf}}, \phi_{\text{src}}, \mathcal{R}\big).$$

We assume $C(x, y) > 0$ and $\mathbb{E}_{q(x)}[|\log C(x, y)|] < \infty$.

Introducing diffusion latents h (noise variables and timestep indices) and a variational distribution $q(x, h)$, we optimize the free-energy bound

$$F \;=\; \mathrm{KL}\big(q(x, h) \,\|\, p(x, h)\big) \;-\; \mathbb{E}_{q(x)}[\log C(x, y)]\,, \qquad p(x, h) = p(h)\, p(x \mid h). \tag{2}$$

The first term encourages samples likely under the pretrained diffusion prior; the second provides plug-and-play guidance for satisfying the constraints. In practice (see §2.1), we adopt a mode-seeking choice $q(x) = \delta(x - \eta)$, so $-\log C(\eta, y)$ enters DDPM updates as a simple additive guidance signal. A full derivation from the ELBO, including the role of h and the reduction to equation 2, appears in Appendix A.

2.1 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs learn to reverse a fixed forward noising process that transforms a clean sample x_0 into Gaussian noise $x_T \sim \mathcal{N}(0, I)$. The forward marginals satisfy

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\, (1 - \bar{\alpha}_t) I\big), \qquad \bar{\alpha}_t = \textstyle\prod_{s=1}^{t} \alpha_s.$$

The reverse process is a Markov chain $p_\theta(x_{t-1} \mid x_t)$; in the noise-prediction parameterization the model $\hat{\epsilon}_\theta(x_t, t)$ is trained with an MSE denoising loss that is *equivalent (up to constants)* to maximizing the variational bound Ho et al. (2020).

Following Chung et al. (2023), we adopt a *mode-seeking* variational family in which the marginal collapses to a point, $q(x) = \delta(x - \eta)$. This avoids integration over latent trajectories and enables plug-and-play use of pretrained DDPMs with external constraint guidance; in our case, $-\log C(\eta, y)$ enters the denoising updates as an additive guidance term (see App. B).

2.2 Free-Energy Instantiation for DDPM

Under the mode-seeking approximation $q(x) = \delta(x - \eta)$, the free-energy bound (Eq. 2) becomes the standard DDPM loss minus a time-ramped log-constraint term:

$$\begin{aligned} F(\eta; \theta) = \mathbb{E}_{\substack{t \sim \mathrm{Unif}\{1,\dots,T\} \\ \epsilon_0 \sim \mathcal{N}(0,I)}} \Big[\big\| \hat{\epsilon}_\theta(x_t, t) - \epsilon_0 \big\|_2^2 \Big] - \mathbb{E}_t\big[w(t)\, \log C(\eta, y)\big], \\ \text{where } x_t = \sqrt{\bar{\alpha}_t}\, \eta + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_0,\; C(\eta, y) = c_1(\eta, y)\, c_2(\eta, y). \end{aligned} \tag{3}$$

If separate control weights are desired, $-\log C(\eta, y) = -\lambda_1 \log c_1(\eta, y) - \lambda_2 \log c_2(\eta, y)$ with $\lambda_1, \lambda_2 > 0$. Because c_1, c_2 are differentiable, we update the clean point η by backpropagating through $-\log C(\eta, y)$ while keeping $\hat{\epsilon}_\theta$ fixed (pretrained). See App. C for details.

3 DIFFUSION-BASED CODE-SWITCHING MODEL (DCSM)

The DCSM transforms a monolingual utterance x into a code-switched one while preserving fluency and coherence. It approximates the constrained posterior in Eq. equation 1 by iteratively refining x with a pretrained DDPM, guided by two differentiable constraints:

- $c_1(x, y)$ — **LID-based switch controller**: decides *where* and *how much* to switch.
- $c_2(x, y)$ — **Semantic segment infusion**: retrieves segments in $\mathcal{L}_{\text{inf}}$, then *blends and writes back* them with a time ramp $w(t)$ (prosody/context-aware).

Together with the DDPM prior, these constraints reshape x toward natural code-switching under the free-energy objective in Eq. equation 3. We detail c_1 in §3.1 and c_2 in §3.2.

3.1 CONSTRAINT $c_1(x, y)$: LANGUAGE IDENTIFICATION AND INFUSION DECISION

The first constraint decides whether a *monolingual* utterance x_j should undergo foreign-language infusion, and by how much. We segment x_j into n spans $\{s^{(i)}\}_{i=1}^{n}$ (e.g., short log-Mel windows) and run a frozen multilingual LID classifier f_{cl} that outputs posteriors over $\mathcal{L}_{\text{all}} = \{\ell_{\text{mono}}\} \cup \mathcal{L}_{\text{inf}}(y)$ with $\mathcal{L}_{\text{inf}}(y) = \{\ell_1, \ldots, \ell_m\}$:

$$p(\ell \mid s^{(i)}) \;=\; f_{\text{cl}}(s^{(i)})_\ell, \qquad \sum_{\ell \in \mathcal{L}_{\text{all}}} p(\ell \mid s^{(i)}) = 1.$$

(Here the "monolingual language" is the language of x_j; languages eligible for infusion come from y via $\mathcal{L}_{\text{inf}}$.)

Per-segment foreignness and global presence.

$$u^{(i)} \;=\; \sum_{\ell \in \mathcal{L}_{\text{inf}}} p(\ell \mid s^{(i)}) \in [0, 1], \qquad P_{\text{foreign}}(x_j) \;=\; \tfrac{1}{n} \sum_{i=1}^{n} u^{(i)} \in [0, 1].$$

Global controller (scalar c_1). Given the desired infusion rate $\pi_{\text{switch}} \in [0, 1]$,

$$c_1(x_j, y) \;=\; \sigma\Big(\alpha_1 \,[\, P_{\text{foreign}}(x_j) - \pi_{\text{switch}} \,]\Big),$$

with sharpness $\alpha_1 > 0$. This scalar enters the free-energy objective via $-\log c_1$.

Local soft gates (for segment selection).

$$g^{(i)} \;=\; \sigma\Big(\alpha_1 \,[\, u^{(i)} - \pi_{\text{switch}} \,]\Big),$$

used as masks to prioritize segments during retrieval/blend in c_2. Optionally, derive straight-through hard gates $z^{(i)} \in \{0, 1\}$ from $g^{(i)}$ at inference while keeping gradients through the sigmoid.

Loss from c_1 (scalar).

$$\mathcal{L}_{c_1}(x_j, y) = -\log c_1(x_j, y) = -\log \sigma\Big(\alpha_1 \,[\, P_{\text{foreign}}(x_j) - \pi_{\text{switch}} \,]\Big). \tag{4}$$

During sampling, $\mathcal{L}_{c_1}$ is added to the free-energy with a (possibly time-ramped) weight $w(t)$.

3.2 SECOND CONSTRAINT: $c_2(x, y)$ — FOREIGN SEGMENT INFUSION

At each denoising step we edit a *single* span selected by c_1. Let $i^\star = \arg\max_i g^{(i)}$ be the chosen source segment in utterance x_j. Constraint c_2 replaces $s_{x_j}^{(i^\star)}$ with a semantically matched, prosodically compatible segment from an infusion language.

Notation. f_{enc}: pretrained, frozen multilingual segment encoder (outputs ℓ_2-normalized embeddings); $\mathrm{Sim}(a,b) = a^\top b$ (cosine similarity); $\mathcal{D}$: FAISS index of candidate segments $s_{y_k}^{(m)}$ with language labels $\ell(\cdot)$ and metadata; $\mathcal{L}_{\text{inf}}(y)$: infusion language set specified by y; $g^{(i)}$: per-segment gate from c_1; $i^\star = \arg\max_i g^{(i)}$; $d_{x_j}^{(i)}$, $\mathcal{O}_{x_j}^{(i)}$: duration and onset of source segment $s_{x_j}^{(i)}$; $\bar{d}_{x_j}$: mean segment duration in x_j; $d_{y_k}^{(m)}$, $\mathcal{O}_{y_k}^{(m)}$: duration and onset of candidate $s_{y_k}^{(m)}$; R_{x_j}, P_{x_j} and R_{y_k}, P_{y_k}: speech-rate and prosody proxies (e.g., syllables/s, median F0/energy) for the source and candidate utterances; w: neighbor window size (segments); $\lambda_d \in [0,1)$: duration tolerance; $\Delta\tau = |\bar{d}_{x_j} - \bar{d}_{y_k}|$: onset tolerance; α, β, γ: weights in the infusion loss (defaults $\alpha = \beta = 1$, $\gamma = 0.1$); α_t: time-ramped blend coefficient; M: (optional) soft retrieval beam; $\tau_{\text{ret}} > 0$: retrieval temperature.

Retrieval (top-1 replacement). Encode the query $q = f_{\text{enc}}(s_{x_j}^{(i^\star)})$ (frozen f_{enc}) and query $\mathcal{D}$ restricted to $\ell(s_{y_k}^{(m)}) \in \mathcal{L}_{\text{inf}}(y)$. Select

$$m^\star = \arg\max_m \mathrm{Sim}\big(q,\ f_{\text{enc}}(s_{y_k}^{(m)})\big), \qquad s^\star \equiv s_{y_k}^{(m^\star)}.$$

Optional (early steps): use a soft top-M mixture and anneal $\tau_{\text{ret}} \downarrow 0$ to hard top-1 late.

Prosodic compatibility. Normalize the target duration by tempo/prosody:

$$S_{\text{ratio}} = \sqrt{\frac{R_{x_j}}{R_{y_k}} \cdot \frac{P_{x_j}}{P_{y_k}}}, \qquad \hat{d} \ = \ d_{x_j}^{(i^\star)}\, S_{\text{ratio}}.$$

Define hinge penalties:

$$\mathcal{L}_{\text{dur}} = \max\Big(0, \frac{|d_{y_k}^{(m^\star)} - \hat{d}| - \lambda_d\, \hat{d}}{\hat{d}}\Big), \quad \mathcal{L}_{\text{on}} = \max\Big(0, \frac{|\mathcal{O}_{y_k}^{(m^\star)} - \mathcal{O}_{x_j}^{(i^\star)}| - \Delta\tau}{\Delta\tau}\Big).$$

Semantic and contextual consistency. Let $\mathcal{N}(i^\star) = \{s_{x_j}^{(t)} : |t - i^\star| \le w,\ t \ne i^\star\}$. Define

$$\mathcal{L}_{\text{sem}} = -\ \mathrm{Sim}\big(f_{\text{enc}}(s_{x_j}^{(i^\star)}),\ f_{\text{enc}}(s^\star)\big), \qquad \mathcal{L}_{\text{ctx}} = -\ \frac{1}{|\mathcal{N}(i^\star)|} \sum_{s' \in \mathcal{N}(i^\star)} \mathrm{Sim}\big(f_{\text{enc}}(s^\star),\ f_{\text{enc}}(s')\big).$$

Per-step loss and write-back. Gate by $g^{(i^\star)}$ and form the infusion loss:

$$\mathcal{L}_{c_2} = g^{(i^\star)} [\alpha\, \mathcal{L}_{\text{sem}} + \beta\, \mathcal{L}_{\text{ctx}} + \gamma\, (\mathcal{L}_{\text{dur}} + \mathcal{L}_{\text{on}})]\,.$$

Perform blend-and-write-back with ramp α_t:

$$s_{x_j}^{(i^\star)} \ \leftarrow \ (1 - \alpha_t)\, s_{x_j}^{(i^\star)} + \alpha_t\, s^\star,$$

then reassemble the utterance and continue DDPM denoising. Gradients flow through the inputs to f_{enc} and the edited segment (hence to η), while f_{enc} weights and the FAISS index remain fixed.

3.3 End-to-End Objective and Inference

Pretrained components. We reuse a *pretrained, frozen* DDPM prior $\hat{\epsilon}_\theta$ and frozen controllers f_{cl} (LID) and f_{enc} (segment encoder). No DDPM retraining is required.

Free-energy objective (mode-seeking). Under the mode-seeking approximation $q(x) = \delta(x - \eta)$, we optimize the clean point η at inference by minimizing

$$\mathcal{J}(\eta) = \mathbb{E}_{t \sim \mathrm{Unif}\{1:T\},\, \epsilon_0 \sim \mathcal{N}(0,I)} \big[\|\hat{\epsilon}_\theta(x_t, t \,|\, M) - \epsilon_0\|_2^2\big] \ + \ \lambda_1\, \mathcal{L}_{c_1}(x_j, y) \ + \ \lambda_2(t)\, \mathcal{L}_{c_2}(x_j, y), \quad (5)$$

where $x_t = \sqrt{\bar{\alpha}_t}\, \eta + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_0$, M denotes optional acoustic conditioning (e.g., Mel), $\lambda_2(t)$ is a *time-ramped* weight, and $i^\star$ is the single edited span selected by c_1 at step t (minimal CS policy, K=1). The DDPM parameters θ are fixed; gradients flow to η only.

Schedules (stability and late commit). We use a blend ramp $\alpha_t \uparrow$ (small early, larger late) when writing the retrieved segment back, and a guidance ramp $\lambda_2(t) \uparrow$ to delay the effect of c_2 until the sample is less noisy. This stabilizes generation and encourages gradual linguistic modulation.

Inference loop (one-segment-per-step). At each denoising step $t = T \to 1$: (i) segment η and compute gates $g^{(i)}$ with c_1; pick $i^\star = \arg\max_i g^{(i)}$; (ii) retrieve one candidate segment in $\mathcal{L}_{\text{inf}}(y)$, compute $\mathcal{L}_{c_2}$, and blend-and-write-back with α_t; (iii) form x_t, predict $\hat{\epsilon}_\theta(x_t, t \mid M)$, and update

$$\eta \;\leftarrow\; \eta \;-\; \gamma_t \, \nabla_\eta \, \mathcal{J}(\eta).$$

Repeat until $t{=}1$ to obtain the code-switched sample.

Notes. We optionally anneal retrieval from soft (top-M) early to hard top-1 late, keep λ_1 fixed, and apply c_1/c_2 only in the edited window (plus a small temporal neighborhood) to avoid global drift. See Appendix 1 for the full inference procedure.

3.4 Speaker Identity Harmonization

To standardize timbre, pitch, and rhythm across edited spans *without changing content*, we apply a short identity-harmonization pass with the *pretrained, frozen* DDPM prior. We condition on a reference speaker/prosody descriptor extracted from the source utterance and/or the current sample:

$$\phi_{\text{spk}} = \text{ECAPA}(x_{\text{mono}}), \qquad \phi_{\text{mel}} = \text{MelStats}(x),$$

and run a shallow denoising refinement:

$$x_{\text{final}} \;=\; \text{Refine}_\theta\big(x \,\big|\, \phi_{\text{spk}}, \, \phi_{\text{mel}}\big), \tag{6}$$

using $T_{\text{ref}}{=}150$ diffusion steps with a low-noise schedule (late timesteps only).

During refinement we disable segment edits ($\lambda_2{=}0$) and keep the LID rate term weak (small λ_1) to preserve semantics and switch structure. In practice, conditioning can be implemented via feature concatenation or FiLM-style modulation in the DDPM U-Net, with θ kept fixed. Alternative variants (e.g., adding a cosine speaker-embedding penalty $\mathcal{L}_{\text{id}}$ computed by a frozen ECAPA on sliding windows) yield similar improvements; see Appendix E for details.

4 Evaluation

5 Dataset

We curate a proprietary speech corpus from the Kenya Broadcasting Corporation (KBC), which operates 11 radio stations delivering aligned news content across multiple Kenyan languages. News bulletins are authored in English and translated into local languages, then read by native presenters—yielding semantically aligned *monolingual* utterances across languages.

Scope and acquisition. The collection spans 2018–2023 and focuses on the daily 7 pm bulletins, which are typically the most content-rich. To preserve real-world variability, we retain advertisements, presenter intros, and ambient segments. Each bulletin is segmented with an over-segmentation VAD pipeline Duquenne et al. (2021), producing silence-bounded units of 3–20 s.

Languages. We target five major languages—Swahili, Luo, Kikuyu, Nandi, and English—chosen for geographic and typological coverage within Kenya. Swahili and Kikuyu are Bantu (Niger–Congo), while Luo and Nandi are Nilotic (often grouped under Nilo–Saharan). English serves as a lingua franca across stations.

Speakers and variability. Bulletins are voiced by different presenters across dates and stations, capturing diverse accents, pitch ranges, and prosodic styles. This variation supports generalization for synthesis, translation, and recognition tasks.

Splits and standardization. We apply a 70/30 split *by segment within language* for training/evaluation and zero-pad segments to a fixed 20 s window for batching. (In practice, we group by bulletin ID to avoid near-duplicate leakage across splits.)

Table 1: Monolingual speech dataset summary (2018–2023, 7 pm bulletins).

Language	Family	Bulletins (count)	Duration (hrs)
Swahili	Niger–Congo	2190	1353
Luo	Nilotic	2190	1284
Kikuyu	Niger–Congo	2190	1304
Nandi	Nilotic	2190	1256
English	–	2190	1206

Table 2: Segment statistics per language (VAD units).

Language	Avg. seg. length (s)	Total segments
Swahili	15.2	320,400
Luo	16.5	280,100
Kikuyu	15.6	300,900
Nandi	17.1	264,600
English	15.0	289,400

Code-switched field recordings. Beyond news, we include 4,044 Swahili and 3,211 naturally occurring code-switched utterances collected from morning call-in shows. For controlled evaluation, selected items were re-recorded in monolingual Swahili and Luo by consenting speakers to create paired counterparts.

5.1 Tools and Resources

Tools and Models. We use standard components for classification, encoding, and synthesis: a multilingual segment-level language classifier, a contrastive speech encoder, the SegUniDiff model for conditional speech generation, and lightweight ASR/MT systems for evaluation. Full architectural and training details, along with performance metrics, are provided in Appendix F.

5.2 Evaluation Metrics and Framework

We evaluate segment- and utterance-level fidelity with four text-based metrics that jointly capture surface and semantic adequacy: SacreBLEU Post (2018), BERTScore Zhang et al. (2019), COMET Rei et al. (2020), and LaBSE cosine similarity Feng et al. (2022). Text proxies are obtained via a *fixed* ASR→MT pipeline used uniformly across all conditions; the intent is metric comparability, not absolute ASR/MT quality. We report means with 95% confidence intervals from 1,000 bootstrap samples (language-stratified, percentile intervals unless noted). Metric definitions and the resampling protocol are in Appendix G.

5.3 Segment-Level Evaluation with Confidence Intervals

We assess semantic fidelity on 8,500 synthetic utterances and 7,255 naturally code-switched utterances (Swahili, Luo) from radio call-ins (Sec. 5). For each utterance x, we extract VAD segments, run the frozen LID f_{cl} to identify foreign spans, and pair each detected foreign segment $s_{x_c}^{(k)}$ with its monolingual counterpart $s_x^{(k)}$: for *synthetic* data, indices are known from the edit log; for *natural* data, we align to the monolingual re-recording via DTW over f_{enc} embeddings. Both segments are transcribed (ASR), translated to English (MT), and scored by SacreBLEU, BERTScore, COMET, and LaBSE. We report language-wise scores with 95% CIs.

Table 3: Segment-level semantic fidelity for synthetic and natural CS (higher is better ↑). Means with 95% CIs from 1,000 bootstrap samples.

Source	**SacreBLEU ↑**	**BERTScore ↑**	**COMET ↑**	**LaBSE ↑**
Swahili (Synthetic)	38.4 [36.9, 39.6]	0.814 [0.809, 0.818]	0.831 [0.822, 0.843]	0.890 [0.882, 0.897]
Luo (Synthetic)	35.7 [34.2, 37.1]	0.805 [0.801, 0.810]	0.809 [0.798, 0.819]	0.876 [0.869, 0.884]
Kikuyu (Synthetic)	36.8 [35.0, 38.4]	0.808 [0.802, 0.813]	0.817 [0.806, 0.828]	0.881 [0.874, 0.888]
Nandi (Synthetic)	34.5 [33.1, 35.7]	0.801 [0.795, 0.806]	0.804 [0.794, 0.814]	0.872 [0.865, 0.880]
Luo (Natural)	36.2 [35.4, 37.0]	0.822 [0.818, 0.841]	0.833 [0.827, 0.843]	0.885 [0.879, 0.891]
Swahili (Natural)	39.0 [38.2, 39.7]	0.836 [0.831, 0.841]	0.845 [0.841, 0.864]	0.898 [0.892, 0.904]
Average (Synthetic)	**36.4 [35.2, 37.4]**	**0.807 [0.804, 0.810]**	**0.815 [0.807, 0.823]**	**0.880 [0.873, 0.886]**

Our model achieves strong semantic fidelity on synthetic utterances; Swahili consistently leads, likely due to stronger ASR/MT coverage. Synthetic scores closely track natural CS: e.g., COMET gaps are +0.014 (Swahili) and +0.024 (Luo); LaBSE gaps are +0.008 and +0.009, respectively. These results indicate that minimal, guided segment infusion can recover cross-lingual semantics without access to parallel CS training data. Full confidence-interval methodology and sensitivity to ASR/MT choice are detailed in Appendix G.

5.4 Utterance-Level Evaluation with Masked and Full Variants

We assess fluency, cross-span coherence, and long-context consistency at the utterance level, complementing the segment results in Sec. 5.3. We evaluate both **synthetic** code-switched utterances and **natural** ones from Swahili/Luo radio call-ins (Sec. 5). A frozen LID f_{cl} identifies foreign-language spans in each code-switched utterance x_c. We then evaluate two text-based variants against a clean monolingual reference x_{ref}:

- **Full reconstruction (unmasked).** For each foreign span in x_c, we run ASR (fixed model) to obtain text in the foreign language, translate it with a fixed MT system into the monolingual language, and *reinsert* the translation into the original transcript positions to form a fully monolingual hypothesis x_{full}.
- **Masked evaluation.** We *remove* all foreign spans from x_c, concatenating the remaining monolingual spans to form x_{mask}. This probes preservation of unedited content and boundary effects.

Both x_{full} and x_{mask} are compared to x_{ref} using SacreBLEU, BERTScore, COMET, and LaBSE. ASR→MT systems are identical across conditions to ensure comparability (the aim is *relative* scoring, not absolute ASR/MT quality). We report means with 95% CIs from 1,000 bootstrap samples (language-stratified percentile intervals). Table 4 summarizes results (higher is better ↑).

Table 4: Utterance-level evaluation of code-switched speech. Full = foreign spans translated back and reinserted; Masked = foreign spans removed. Means with 95% CIs from 1,000 bootstrap samples; higher is better ↑.

Source	Type	SacreBLEU ↑	BERTScore ↑	COMET ↑	LaBSE ↑
Swahili (Synthetic)	Full	36.6 [35.3, 37.8]	0.762 [0.757, 0.766]	0.669 [0.660, 0.681]	0.882 [0.875, 0.888]
	Masked	34.9 [33.5, 36.0]	0.737 [0.732, 0.737]	0.642 [0.631, 0.655]	0.854 [0.846, 0.860]
Luo (Synthetic)	Full	33.9 [32.6, 35.3]	0.753 [0.747, 0.758]	0.647 [0.637, 0.657]	0.871 [0.864, 0.878]
	Masked	32.2 [30.9, 33.6]	0.728 [0.722, 0.733]	0.620 [0.609, 0.631]	0.844 [0.837, 0.854]
Kikuyu (Synthetic)	Full	35.0 [33.4, 36.4]	0.756 [0.750, 0.761]	0.655 [0.644, 0.666]	0.876 [0.870, 0.882]
	Masked	33.3 [31.8, 34.6]	0.731 [0.725, 0.736]	0.628 [0.617, 0.639]	0.850 [0.843, 0.857]
Nandi (Synthetic)	Full	32.7 [31.5, 33.9]	0.749 [0.743, 0.756]	0.643 [0.633, 0.653]	0.869 [0.861, 0.875]
	Masked	31.2 [29.9, 32.4]	0.724 [0.717, 0.731]	0.615 [0.604, 0.625]	0.841 [0.834, 0.849]
Swahili (Natural)	Full	37.3 [36.5, 38.1]	0.785 [0.780, 0.791]	0.701 [0.692, 0.710]	0.896 [0.890, 0.902]
	Masked	35.5 [34.7, 36.3]	0.760 [0.755, 0.765]	0.667 [0.658, 0.676]	0.867 [0.860, 0.873]
Luo (Natural)	Full	35.2 [34.1, 37.3]	0.772 [0.767, 0.778]	0.682 [0.673, 0.691]	0.884 [0.878, 0.891]
	Masked	33.6 [32.5, 35.6]	0.745 [0.740, 0.751]	0.654 [0.645, 0.662]	0.856 [0.850, 0.862]
Average (Synthetic)	Full	**34.5 [33.5, 35.6]**	**0.755 [0.751, 0.759]**	**0.653 [0.645, 0.661]**	**0.874 [0.870, 0.878]**
	Masked	**32.9 [31.9, 34.0]**	**0.730 [0.726, 0.734]**	**0.626 [0.617, 0.635]**	**0.847 [0.843, 0.852]**

While utterance-level scores are lower than segment-level (Table 3), the model retains strong fluency and coherence. Swahili again leads, likely reflecting stronger ASR/MT support. The masked variant indicates that unedited content is largely preserved (small gap vs. Full), and the modest synthetic–natural differences (e.g., Swahili COMET ≈ +0.014) suggest that minimal, guided segment infusion reproduces key properties of real code-switching without parallel CS training data. Sensitivity to the ASR/MT choice and the bootstrap details are in Appendix G.

6 Speaker Identity Verification

We evaluate whether code-switched speech preserves speaker identity using fixed, pretrained ECAPA-TDNN embeddings Desplanques et al. (2020) (frozen weights). For each generated code-switched utterance x_c, we apply VAD (Sec. 5) and extract embeddings $e^{(i)} = f_{\text{spk}}(s_{x_c}^{(i)})$ from non-overlapping segments of 1.5–3.0 s (segments shorter than 1.0 s are merged or discarded). To avoid trivial positives, genuine trials exclude overlapping/adjacent segment pairs within the same utterance.

Trials and scoring. *Genuine* pairs are all $(i \neq j)$ within the same x_c (non-overlapping). *Impostor* pairs are across different utterances with different speaker prompts/ids. Scores are cosine similarities in $[-1, 1]$; we optionally apply s-norm with a cohort of 2,000 embeddings. EER is computed from balanced genuine/impostor trial sets per language; 95% CIs are from 1,000 bootstrap samples.

Results indicate strong identity preservation in synthetic code-switched speech (EER ≈ 6.5–7.6%), with natural CS forming an upper bound (e.g., Swahili: 0.903 similarity, 3.6% EER). The relatively

Table 5: Speaker verification on code-switched utterances (higher is better ↑, lower is better ↓). Means with 95% CIs.

Language	Avg. Cosine (Genuine) ↑	Avg. Cosine (Impostor) ↓	EER % ↓
Swahili (Synthetic)	0.872 [0.867, 0.877]	0.432 [0.427, 0.437]	6.5 [6.1, 6.9]
Luo (Synthetic)	0.861 [0.856, 0.866]	0.418 [0.412, 0.424]	7.2 [6.8, 7.7]
Kikuyu (Synthetic)	0.868 [0.862, 0.874]	0.427 [0.421, 0.433]	6.8 [6.4, 7.2]
Nandi (Synthetic)	0.854 [0.848, 0.860]	0.411 [0.405, 0.417]	7.6 [7.1, 8.1]
Swahili (Natural)	0.903 [0.898, 0.908]	0.391 [0.386, 0.396]	3.6 [3.3, 3.9]
Luo (Natural)	0.870 [0.865, 0.875]	0.430 [0.425, 0.435]	5.1 [4.8, 5.5]
Average (Synthetic)	**0.864**	**0.422**	**7.0**

small gap suggests that segment-level infusion plus DDPM refinement (Sec. 3.4) maintains speaker traits across substitutions.

Alternation Points. We define alternation rate as the proportion of segment boundaries where the language label changes:

$$\text{Alternation Rate} = \frac{\#\text{Alternations}}{\#\text{Segments}}$$

Table 6 shows alternation is rare (3–5%), with synthetic and natural values well-aligned. Swahili (Natural) alternates most, likely due to short insertions. Luo shows the opposite trend—longer insertions, fewer switches.

Table 6: Average alternation rate: % of boundaries where language changes.

Source Language	Alternation Rate (%)
Swahili (Synthetic)	4.7%
Luo (Synthetic)	3.6%
Kikuyu (Synthetic)	3.1%
Nandi (Synthetic)	4.1%
Swahili (Natural)	5.3%
Luo (Natural)	2.1%
Avg. (Synthetic)	**3.88%**

Across four structural dimensions, our model exhibits realistic code-switching behavior. It avoids over-insertion, respects language constraints, mirrors natural switch placement, and captures alternation rates—all without explicit rules. This suggests it internalizes structural code-switching patterns via training on monolingual segments and guided diffusion alone.

6.1 Human Preference Evaluation: Fluency and Acceptability

We complement automatic metrics with a human study targeting perceptual qualities that automatic scores miss. We recruited $N = 638$ undergraduate raters (consented; uncompensated/compensated per IRB/ethics approval) and sampled $1{,}437$ utterances (balanced across source languages and synthetic/natural). Raters were *language-matched* (self-reported proficiency in the utterance's monolingual language); each rater evaluated exactly six utterances, and each utterance received ≥ 4 independent ratings.

Protocol. Stimuli were presented in randomized order and *blind* to condition (synthetic vs. natural). We enforced headphone use (HINT-style check) and included two attention checks per rater. Very short/long items (< 3 s or > 20 s) were excluded for consistency. Raters used 5-point Likert scales on three dimensions with brief anchors: **Fluency** (smooth, natural delivery), **Coherence** (semantic consistency and perceived single-speaker identity), **Realism** (resemblance to naturally occurring multilingual speech).

Analysis. We report means with 95% CIs via 1,000 bootstrap resamples (clustered by utterance). Inter-rater reliability (average-measures ICC) was $\text{ICC}_{\text{fluency}} = 0.79$, $\text{ICC}_{\text{coherence}} = 0.76$, $\text{ICC}_{\text{realism}} = 0.74$. Group differences were tested with a mixed-effects model (fixed: language, condition; random: utterance, rater).

Synthetic utterances score ≥ 4.0 on all dimensions with tight CIs, indicating high perceived quality. Natural utterances remain a ceiling, especially on realism (e.g., $+0.5$ for Swahili), but the gap is

Table 7: Human ratings (5=best). Means with 95% CIs; SD shown for per-utterance scores.

Source Language	Fluency ↑	Coherence ↑	Realism ↑	Std. Dev.
Swahili (Synthetic)	4.10 [4.05, 4.15]	4.20 [4.15, 4.25]	4.00 [3.95, 4.05]	0.42
Luo (Synthetic)	4.08 [4.03, 4.13]	4.02 [3.97, 4.07]	4.07 [4.02, 4.12]	0.46
Kikuyu (Synthetic)	4.18 [4.12, 4.23]	4.10 [4.05, 4.16]	4.01 [3.96, 4.06]	0.44
Nandi (Synthetic)	3.94 [3.89, 3.99]	4.01 [3.96, 4.06]	3.92 [3.87, 3.97]	0.48
Luo (Natural)	4.73 [4.68, 4.77]	4.42 [4.37, 4.47]	4.60 [4.55, 4.65]	0.46
Swahili (Natural)	4.59 [4.54, 4.63]	4.82 [4.78, 4.86]	4.48 [4.43, 4.53]	0.42
Avg. (Synthetic)	**4.08 [4.05, 4.11]**	**4.08 [4.05, 4.11]**	**4.00 [3.97, 4.03]**	**0.45**

modest. Mixed-effects analysis confirms a significant main effect of condition (natural>synthetic, $p < 0.01$) and a language effect (Swahili>others), with no significant interaction, suggesting consistent synthetic quality across languages.

7 Conclusion

We presented a diffusion-based framework for generating fluent, coherent, and sociolinguistically realistic code-switched speech without relying on parallel training data. By guiding a pre-trained monolingual generative prior with differentiable linguistic constraints—including a multilingual language classifier and a contrastive segment encoder—our method performs targeted segment replacements that preserve fluency, speaker identity, and semantic coherence. Extensive evaluation across five African languages demonstrates that the proposed system closely matches natural code-switching behavior in frequency, structure, and placement, while achieving strong segment-level semantic fidelity (COMET 0.815, LaBSE 0.880) and speaker consistency (EER 6.7%). Human listeners rated the generated utterances favorably across fluency, coherence, and realism. To our knowledge, this is the first method to enable plug-and-play multi-language infusion within a single utterance, offering a new paradigm for cross-lingual speech generation in low-resource settings. Future work will explore integrating prosodic control, expanding to more languages, and applying the approach to spontaneous conversational domains.

References

Peter Auer. *Code-switching in conversation: Language, interaction and identity*. Routledge, 1998.

Astik Biswas, Emre Yılmaz, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. Code-switched automatic speech recognition in five south african languages. *Computer Speech & Language*, 71:101262, 2022.

Yuewen Cao, Songxiang Liu, Xixin Wu, Shiyin Kang, Peng Liu, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng. Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7619–7623. IEEE, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Jie Chi, Brian Lu, Jason Eisner, Peter Bell, Preethi Jyothi, and Ahmed M Ali. Unsupervised code-switched text generation from parallel text. In *Proc. INTERSPEECH*, volume 2023, pp. 1419–1423, 2023.

Hyungjin Chung, Jonathan Ho, Tim Salimans, Jong Chul Lee, and Diederik P. Kingma. Diffusion models as plug-and-play priors. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=PlKWVd2yBkY.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34:15748–15761, 2021.

Fuli Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891. Association for Computational Linguistics, 2022.

Bryan Gregorius and Takeshi Okadome. Generating code-switched text from monolingual text with dependency tree. In *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*, pp. 90–97, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

I Hsu, Avik Ray, Shubham Garg, Nanyun Peng, Jing Huang, et al. Code-switched text synthesis in unseen language pairs. *arXiv preprint arXiv:2305.16724*, 2023.

Sehoon Kim, Amir Gholami, Albert Shaw, Nicholas Lee, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, and Kurt Keutzer. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*, 35:9361–9373, 2022.

Carol Myers-Scotton. *Social motivations for code-switching: Evidence from Africa*. Oxford University Press, 1993.

Oriol Nieto, Zeyu Jin, Franck Dernoncourt, and Justin Salamon. Efficient spoken language recognition via multilabel classification. *arXiv preprint arXiv:2306.01945*, 2023.

Peter Ochieng and Dennis Kaburu. Phonology-guided speech-to-speech translation for african languages. *Speech Communication*, 174:103287, 2025. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2025.103287. URL https://www.sciencedirect.com/science/article/pii/S0167639325001025.

Shana Poplack. Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618, 1980.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*, 2020.

S Sitaram, KR Chandu, SK Rallabandi, and AW Black. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*, 2019.

Sarah Slabbert and Rosalie Finlayson. A socio-historical overview of codeswitching studies in the african languages. *South African Journal of African Languages*, 19(1):60–72, 1999.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. From machine translation to code-switching: Generating high-quality code-switched text. *arXiv preprint arXiv:2107.06483*, 2021.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

8 APPENDIX / SUPPLEMENTAL MATERIAL

A FULL DERIVATION OF THE FREE-ENERGY OBJECTIVE

This appendix derives the free-energy objective used in §2. We aim to approximate the constrained posterior $p(x|y)$, where x is a high-dimensional utterance (waveform) and y is the *infusion specification* (infusion language set, switch prior, source semantics, retrieval index). Let $c_1(x, y)$ and $c_2(x, y)$ be differentiable soft constraints for switch plausibility and semantic consistency, and define $C(x, y) = c_1(x, y)\, c_2(x, y)$.

Assumptions. We assume $C(x, y) > 0$ and $\mathbb{E}_{q(x)}[\,|\log C(x, y)|\,] < \infty$ so expectations are well-defined. We also assume absolute continuity of the variational joint $q(x, h)$ with respect to the model joint $p(x, h)$ on the support of q, so that $\mathrm{KL}(q\|p)$ is finite. The variational marginal is $q(x) = \int q(x, h)\, dh$. For clarity, the constraints consume (possibly overlapping) subsets of y.

Free-energy objective. Consider the (negative) variational free energy

$$F \;=\; -\,\mathbb{E}_{q(x)}\big[\log p(x) + \log C(x, y) - \log q(x)\big], \tag{7}$$

which differs from the negative log evidence by the additive constraint term $\log C(x, y)$.

Latent decomposition and Jensen. Assume $p(x)$ admits a latent factorization with hidden variable h: $p(x) = \int p(x, h)\, dh$. Insert the identity $q(h|x)/q(h|x)$ inside the integral:

$$F = -\,\mathbb{E}_{q(x)}\left[\log \int q(h|x)\, \frac{p(x, h)}{q(h|x)}\, dh + \log C(x, y) - \log q(x)\right]. \tag{8}$$

Applying Jensen's inequality to the concave $\log$ yields

$$\log \int q(h|x)\, \frac{p(x, h)}{q(h|x)}\, dh \;\geq\; \mathbb{E}_{q(h|x)}\left[\log \frac{p(x, h)}{q(h|x)}\right], \tag{9}$$

and thus an *upper bound* on F:

$$F \leq -\,\mathbb{E}_{q(x)q(h|x)}\left[\log \frac{p(x, h)}{q(h|x)}\right] \;-\; \mathbb{E}_{q(x)}\big[\log q(x)\big] \;+\; \mathbb{E}_{q(x)}\big[\log C(x, y)\big]. \tag{10}$$

The bound is *tight* when $q(h \mid x) = p(h \mid x)$.

Joint form. Writing $q(x, h) = q(x)\, q(h|x)$ and rearranging,

$$F \leq -\,\mathbb{E}_{q(x,h)}\left[\log \frac{p(x, h)}{q(x, h)}\right] \;+\; \mathbb{E}_{q(x)}\big[\log C(x, y)\big]. \tag{11}$$

Define the *free-energy bound* we optimize as

$$\widetilde{F} \;\stackrel{\text{def}}{=}\; \underbrace{\mathrm{KL}(q(x, h)\,\|\,p(x, h))}_{\text{prior-matching}} \;-\; \underbrace{\mathbb{E}_{q(x)}\big[\log C(x, y)\big]}_{\text{constraint reward}}, \tag{12}$$

so $F \leq \widetilde{F}$ by Jensen, and we minimize $\widetilde{F}$ in practice. This is the free-energy form reported in the main text (Eq. 2).

Mode-seeking limit (used later). In §2.1 we employ a mode-seeking variational family for $q(x)$ via the narrow-Gaussian limit $q_\sigma(x) = \mathcal{N}(x; \eta, \sigma^2 I) \xrightarrow[\sigma\to 0]{} \delta(x - \eta)$, which preserves the derivation above and yields a tractable inference objective with $-\log C(\eta, y)$ entering as an additive guidance term.

B DDPM AND MODE-SEEKING APPROXIMATION

We justify the mode-seeking approximation used in our constrained diffusion framework. This section builds on denoising diffusion probabilistic models (DDPMs) Ho et al. (2020) and shows how inference under guidance is simplified by collapsing the variational posterior to a point mass.

DDPM recap. DDPMs generate samples by learning to reverse a fixed forward noising process. The forward process gradually adds Gaussian noise to data over T steps:

$$q(h = \{x_1, \dots, x_T\} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \qquad q(x_t \mid x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{\alpha_t}\, x_{t-1}, (1-\alpha_t)I\big), \tag{13}$$

with a fixed variance schedule $\{\alpha_t\}_{t=1}^{T}$. From the forward marginal,

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)I\big), \qquad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s. \tag{14}$$

The reverse process is learned as a Markov chain $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big)$, and, in the noise-prediction parameterization, the network $\hat{\epsilon}_\theta(x_t, t)$ is trained with a simple MSE loss that is *equivalent (up to constants)* to maximizing the ELBO Ho et al. (2020).

Constrained posterior and mode seeking. In our constrained setting we wish to sample from

$$p(x \mid y) \ \propto\ p(x)\, c_1(x, y)\, c_2(x, y), \tag{15}$$

where c_1, c_2 are differentiable soft constraints (see main text for y, the infusion specification). Direct estimation is intractable under a diffusion prior $p(x)$. Following Chung et al. (2023), we adopt a *mode-seeking* variational family via the narrow-Gaussian limit:

$$q_\sigma(x) = \mathcal{N}(x; \eta, \sigma^2 I) \xrightarrow[\sigma \to 0]{} q(x) = \delta(x - \eta), \tag{16}$$

where $\eta \in \mathbb{R}^d$ is an *inference-time* optimization variable (the clean point). This collapses inference to optimization over η and avoids evaluating constraints along full diffusion trajectories.

Given η, the noised input at timestep t is sampled from the forward marginal:

$$x_t = \sqrt{\bar{\alpha}_t}\, \eta + \sqrt{1 - \bar{\alpha}_t}\, \epsilon_0, \qquad \epsilon_0 \sim \mathcal{N}(0, I). \tag{17}$$

We then feed x_t to the noise predictor $\hat{\epsilon}_\theta(x_t, t)$. We *reuse* the standard denoising objective

$$\mathcal{L}_{\text{DDPM}} \ = \ \mathbb{E}_{t \sim \text{Unif}\{1,\dots,T\},\ \epsilon_0 \sim \mathcal{N}(0,I)} \left[\left\| \hat{\epsilon}_\theta(x_t, t) - \epsilon_0 \right\|_2^2 \right], \tag{18}$$

to define gradients with respect to η during guided sampling (the network parameters θ are fixed).

Why mode seeking helps for guidance. Because x_t is a differentiable function of η, the constraint terms $c_1(x, y), c_2(x, y)$ contribute $\nabla_\eta[-\log C(\eta, y)]$ that can be combined with the denoising gradients. Practically, this: (i) enables direct gradients $\nabla_\eta \log c_i(\eta, y)$, (ii) reduces variance and computation, and (iii) allows plug-and-play reuse of a pretrained diffusion prior p_θ without retraining under new constraints. Empirically, we find the mode-seeking approximation preserves sample quality while enabling fine-grained control over linguistic attributes during generation.

C Derivation of Free-Energy Objective under DDPM

This section expands §2.2, instantiating the free-energy bound (Eq. 2) for DDPMs under a mode-seeking variational family.

Variational family. We take the narrow-Gaussian limit around a clean point η:

$$q_\sigma(x) = \mathcal{N}(x; \eta, \sigma^2 I) \xrightarrow[\sigma \to 0]{} q(x) = \delta(x - \eta),$$

and define the forward trajectory $q(x_{1:T} \mid \eta) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$ with $x_0 = \eta$. Thus the joint variational density factorizes as $q(x, h) = q(x)\, q(h \mid x) = \delta(x - \eta)\, q(x_{1:T} \mid \eta)$ with latent path $h \equiv x_{1:T}$.

Bound decomposition. Starting from Appendix A,

$$\widetilde{F} \;=\; \mathrm{KL}\big(q(x,h)\,\|\,p(x,h)\big) \;-\; \mathbb{E}_{q(x)}[\log C(x,y)],$$

plug in the factorization and marginalize x to obtain

$$\begin{aligned}\widetilde{F}(\eta;\theta) &= \underbrace{\mathbb{E}_{q(x_{1:T}|\eta)}\Big[\log \tfrac{q(x_{1:T}|\eta)}{p_\theta(\eta,x_{1:T})}\Big]}_{\text{DDPM ELBO at clean point } \eta} \;-\; \log C(\eta,y),\\ &\text{where } p_\theta(\eta,x_{1:T}) \equiv p(x_T)\prod_{t=T}^{2} p_\theta(x_{t-1}\mid x_t).\end{aligned} \tag{19}$$

Using the standard DDPM ELBO decomposition Ho et al. (2020), this becomes

$$\begin{aligned}\widetilde{F}(\eta;\theta) = D_{\mathrm{KL}}\big(q(x_T\mid\eta)\,\|\,p(x_T)\big)\\ +\sum_{t=2}^{T}\mathbb{E}_{q(x_t|\eta)}\Big[D_{\mathrm{KL}}\big(q(x_{t-1}\mid x_t,\eta)\,\|\,p_\theta(x_{t-1}\mid x_t)\big)\Big] \;-\; \log C(\eta,y),\end{aligned} \tag{20}$$

where $q(x_{t-1}\mid x_t,\eta)$ is the closed-form forward posterior and $p_\theta(\cdot\mid\cdot)$ is the learned reverse kernel.

MSE form. Optimizing equation 20 is *equivalent (up to constants)* to minimizing the denoising MSE Ho et al. (2020):

$$\begin{aligned}\mathcal{L}_{\mathrm{DDPM}}(\eta;\theta) = \mathbb{E}_{t,\epsilon_0}\big[\|\hat\epsilon_\theta(x_t,t)-\epsilon_0\|_2^2\big],\\ \text{where } x_t=\sqrt{\bar\alpha_t}\,\eta+\sqrt{1-\bar\alpha_t}\,\epsilon_0,\quad \bar\alpha_t=\prod_{s=1}^{t}\alpha_s,\\ t\sim\mathrm{Unif}\{1{:}T\},\ \epsilon_0\sim\mathcal{N}(0,I).\end{aligned} \tag{21}$$

Final objective and optimization. Combining terms, the bound we optimize is

$$\widetilde{F}(\eta;\theta) = \mathcal{L}_{\mathrm{DDPM}}(\eta;\theta) \;-\; \log C(\eta,y) \quad \big(\text{or time-ramped } -\mathbb{E}_t[w(t)\log C(\eta,y)]\big). \tag{22}$$

In our plug-and-play setting, θ is *fixed* (pretrained prior) and we optimize η at inference by gradient steps:

$$\eta \leftarrow \eta \;-\; \gamma_t\,\nabla_\eta\big(\mathcal{L}_{\mathrm{DDPM}}(\eta;\theta)-\log C(\eta,y)\big),$$

optionally interleaving with DDPM sampling updates. When c_2 involves retrieval $\mathcal{R}$, candidate indices from the FAISS search are treated as constants (stop-gradient or straight-through), and we backpropagate through the differentiable scoring. Equation equation 22 corresponds to Eq. 3 in the main text.

D Inference Objective and Schedules (Details)

We keep the DDPM prior $\hat\epsilon_\theta$ and auxiliaries $f_{\mathrm{cl}}, f_{\mathrm{enc}}$ *fixed*. At inference we optimize the clean point η:

$$\min_{\eta}\ \mathbb{E}_{t\sim\mathrm{Unif}\{1:T\},\,\epsilon_0\sim\mathcal{N}(0,I)}\big[\|\hat\epsilon_\theta(x_t,t)-\epsilon_0\|_2^2\big]+\lambda_1\,\mathcal{L}_{c_1}(\eta,y)+\mathbb{E}_t[w(t)]\,\lambda_2\,\mathcal{L}_{c_2}(\eta,y),$$

where $x_t=\sqrt{\bar\alpha_t}\,\eta+\sqrt{1-\bar\alpha_t}\,\epsilon_0$ and $\bar\alpha_t=\prod_{s=1}^{t}\alpha_s$. We use Adam (lr $=2\times10^{-3}$, β_1=0.9, β_2=0.999), T=1000 steps, and update exactly one span per step: $i^\star=\arg\max_i g^{(i)}$.

Schedules. Blend/write-back coefficient:

$$\alpha_t = 1-\exp\Big(-\frac{T-t}{\beta T}\Big),\quad \beta=0.25.$$

Constraint ramp for c_2 ("late commit"):

$$w(t)=t/T,\quad \text{while } \lambda_1 \text{ is fixed.}$$

Differentiability notes. We backprop through inputs to f_{cl} and f_{enc} (frozen weights). Retrieval is top-1 with straight-through (early soft top-M optional); FAISS indices are stop-grad. Gradients are clipped to 1.0.

D.1 Constraint Weight Selection

We tune the guidance weights $\lambda_1, \lambda_2 \in [0.1, 5.0]$ with a Gaussian–process Bayesian optimizer (100 trials, log-uniform priors). The objective is a scalarized validation score:

$$J = \alpha_{\text{sem}}\big(\text{COMET} + \text{BERTScore}\big) - \alpha_{\text{pros}}\big(\Delta_{\text{onset}} + \Delta_{\text{dur}}\big) + \alpha_{\text{spk}}\ \text{cos\_sim}_{\text{ECAPA}},$$

with $\alpha_{\text{sem}}{=}1,\ \alpha_{\text{pros}}{=}1,\ \alpha_{\text{spk}}{=}1$ (normalized terms). We obtain $\lambda_1{=}0.35,\ \lambda_2{=}0.65$ on a held-out validation split. (Section 5.2 defines all metrics; details in App. G.)

D.2 Segment-Wise Gradient Update (Inference)

At each denoising step t, we edit exactly one span chosen by the c_1 gate: $i^\star = \arg\max_i g^{(i)}$. Let $m^{(i^\star)} \in \{0,1\}^{T_x}$ be a binary mask for that span (upsampled to the waveform), and let η be the current clean point. We compute

$$x_t = \sqrt{\bar\alpha_t}\,\eta + \sqrt{1-\bar\alpha_t}\,\epsilon_0, \quad \epsilon_0 \sim \mathcal{N}(0, I),$$

and take one masked gradient step on the free-energy objective:

$$\eta \leftarrow \eta \ - \ \gamma\,\Big(\nabla_\eta \underbrace{\|\hat\epsilon_\theta(x_t, t) - \epsilon_0\|_2^2}_{\mathcal{L}_{\text{DDPM}}} \ + \lambda_1\,\mathcal{L}_{c_1}(\eta, y) \ + \ w(t)\,\lambda_2\,\mathcal{L}_{c_2}(\eta, y)\Big) \ \odot \ m^{(i^\star)},$$

where γ is the step size (we use $\gamma = 5 \times 10^{-2}$), $w(t) = t/T$ is the late-commit ramp, and $\odot$ applies the update only within the selected span. After the gradient step, we *blend-and-write-back* the retrieved foreign segment with coefficient α_t (§3.2) and proceed to the next $t{-}1$ step. All networks $\hat\epsilon_\theta, f_{\text{cl}}, f_{\text{enc}}$ remain frozen; gradients flow through inputs only (FAISS indices are stop-grad; straight-through is used if top-1 selection is required).

D.3 Full Inference Algorithm

Algorithm 1 Inference for Minimal-Edit Code-Switched Speech (one segment per step)

Require: Frozen DDPM noise predictor $\hat\epsilon_\theta$; frozen LID f_{cl}; frozen segment encoder f_{enc}; FAISS index $\mathcal{D}$; infusion set $\mathcal{L}_{\text{inf}}(y)$; step sizes $\{\gamma_t\}_{t=1}^T$; blend ramp $\{\alpha_t\}_{t=1}^T$; weights λ_1, $\{\lambda_2(t)\}_{t=1}^T$; noise schedule $\{\bar\alpha_t\}_{t=1}^T$; optional conditioning M

Initialize clean point $\eta \sim \mathcal{N}(0, I)$ // mode-seeking start
for $t = T$ **down to** 1 **do**
  Sample $\epsilon_0 \sim \mathcal{N}(0, I)$; set $x_t = \sqrt{\bar\alpha_t}\,\eta + \sqrt{1-\bar\alpha_t}\,\epsilon_0$
  Segment & gate (Sec. 3.1): segment $\eta \to \{s_{x_j}^{(i)}\}_{i=1}^n$; compute gates $g^{(i)}$
  Compute $\mathcal{L}_{c_1}(x_j, y) \leftarrow -\log c_1(x_j, y)$ (Eq. equation 4)
  Select span $i^\star \leftarrow \arg\max_i g^{(i)}$ // minimal CS policy ($K{=}1$)
  Retrieval (Sec. 3.2):
    $q \leftarrow f_{\text{enc}}(s_{x_j}^{(i^\star)})$
    $m^\star \leftarrow \arg \max_{m:\, \ell(s_{y_k}^{(m)}) \in \mathcal{L}_{\text{inf}}(y)} \text{Sim}\big(q, f_{\text{enc}}(s_{y_k}^{(m)})\big)$; $s^\star \leftarrow s_{y_k}^{(m^\star)}$
  Prosody & semantics (Sec. 3.2): compute $\mathcal{L}_{\text{dur}},\ \mathcal{L}_{\text{on}},\ \mathcal{L}_{\text{sem}},\ \mathcal{L}_{\text{ctx}}$
  $\mathcal{L}_{c_2}^{(i^\star)} \leftarrow g^{(i^\star)}\big[\alpha\,\mathcal{L}_{\text{sem}} + \beta\,\mathcal{L}_{\text{ctx}} + \gamma(\mathcal{L}_{\text{dur}} + \mathcal{L}_{\text{on}})\big]$
  Blend & write-back: $s_{x_j}^{(i^\star)} \leftarrow (1-\alpha_t)\,s_{x_j}^{(i^\star)} + \alpha_t\,s^\star$; reassemble η
  DDPM loss: $\mathcal{L}_{\text{DDPM}} \leftarrow \big\|\hat\epsilon_\theta(x_t, t \mid M) - \epsilon_0\big\|_2^2$
  Objective: $\mathcal{J}(\eta) \leftarrow \mathcal{L}_{\text{DDPM}} + \lambda_1\,\mathcal{L}_{c_1}(x_j, y) + \lambda_2(t)\,\mathcal{L}_{c_2}^{(i^\star)}$
  Update: $\eta \leftarrow \eta - \gamma_t\,\nabla_\eta\,\mathcal{J}(\eta)$
end for
(Optional) harmonize speaker identity (Sec. 3.4)
return x_{final} (assembled from η)

E Speaker Identity Harmonization Details

Local, segment-level edits and cross-lingual substitutions can leave small inconsistencies in voice quality, prosody, or timbre. We therefore apply a short, *content-preserving* refinement with the same *pretrained, frozen* diffusion prior.

Reference conditioning. We extract a global reference from the utterance to standardize speaker characteristics:

$$\phi_{\text{mel}} = \text{Mel}(x) \quad \text{(e.g., 80-dim log-Mel, 25 ms window, 10 ms hop)},$$

and, when available, a speaker embedding from the original monolingual audio,

$$\phi_{\text{spk}} = \text{ECAPA}(x_{\text{mono}}).$$

We condition the refinement U-Net via feature concatenation or FiLM modulation.

Refinement pass. We run T_{ref}=150 late diffusion steps (low-noise regime) with DDPM parameters fixed and constraint guidance disabled (λ_1=0, λ_2=0):

$$x_{\text{final}} \;=\; \text{DDPM}_{\text{refine}}\big(x \,\big|\, \phi_{\text{mel}},\, \phi_{\text{spk}}\big).$$

This standardizes timbre and prosody across edited spans without altering content or switch placement.

Observed effects. (i) **Timbre smoothing**: reduces spectral discontinuities near edit boundaries. (ii) **Prosodic coherence**: aligns pitch and rhythm across spans. (iii) **Single-speaker percept**: mitigates residual cross-speaker cues from retrieved segments.

Alternatives and stability. We experimented with adding a speaker-consistency penalty (cosine distance between ECAPA embeddings of input/output) inside the main inference objective, but found optimization conflicts with semantic/timing terms; convergence degraded. Post-hoc refinement offered better stability and control with negligible overhead.

Sanity checks. We verify identity harmonization via: ECAPA cosine similarity (pre vs. post, higher is better), F0 CV across segment boundaries (lower is better), and MOS-like human ratings for voice consistency.

F Tools and Resources

Multilingual Language Classifier f_{cl}. We use a lightweight ECAPA-TDNN variant (LECAPAT Nieto et al. (2023)) as our segment-level LID model. Inputs are 64-bin log-Mel spectrograms (25 ms window, 10 ms hop; audio at 24 kHz). We train with cross-entropy for 50 epochs (Adam, lr $= 10^{-4}$, batch 64), early stopping (patience 5) and a 10% validation split. No augmentation is applied. The model attains 92.4% validation accuracy averaged across the five languages on a single A100. At inference f_{cl} is *frozen*; gradients flow through inputs only.

Multilingual Segment Encoder f_{enc}. f_{enc} maps segments to a shared space for retrieval. We train a contrastive encoder with a SimCLR-style objective Chen et al. (2020): positives are two augmentations of the same segment; negatives are other segments in the batch. (50% of segments are augmented with additive Gaussian noise and time masking.) The network uses a 1D conv front-end (256 filters, kernel 16, stride 8), an EfficientNet-B0 backbone Tan & Le (2019), global max pooling, and a projection head to a 720-dim space. Training runs 1M steps with AdamW (β_1=0.9, β_2=0.999, batch 512). At inference we discard the projection head and use the EfficientNet embeddings, ℓ_2-normalized; f_{enc} is *pretrained and frozen*. Cosine similarity is the dot product of unit-norm embeddings.

Pretrained Diffusion Prior (SegUniDiff). For synthesis we employ SegUniDiff Ochieng & Kaburu (2025), a segment-aware DDPM conditioned on Mel features. Unless stated otherwise, we use a *single multilingual* SegUniDiff and condition on a language code; the model is *frozen at inference* and reused as the prior in our free-energy objective. See Ochieng & Kaburu (2025) for architecture and training details.

Machine Translation and ASR. To support automatic evaluation, we built parallel text corpora by aligning semantically equivalent sentences across our language pairs and trained Transformer-base MT models. For ASR, we trained Squeezeformer Kim et al. (2022) for Nandi, Luo, and Kikuyu, and Whisper-small Radford et al. (2023) for Swahili and English.

Table 8: Parallel MT data and ASR performance.

Language Pair	Paired Sentences	SacreBLEU (↑)	Language (ASR)	WER (%)
Luo–Nandi	1.76M	32.2	Luo	14.2
Luo–Kikuyu	1.18M	31.8	Nandi	13.6
Nandi–Kikuyu	1.32M	27.3	Kikuyu	14.4
Kikuyu–Swahili	1.29M	30.4	Swahili	9.8
Kikuyu–English	1.71M	24.9	English	5.3
Swahili–English	1.52M	25.4	—	—
Luo–Swahili	1.43M	27.4	—	—
Luo–English	1.34M	28.1	—	—
Nandi–Swahili	1.44M	27.1	—	—
Nandi–English	1.37M	28.6	—	—

G Evaluation Metric Details and Resampling Protocol

Text used for scoring. All metrics operate on text produced by an ASR→MT pipeline (see §5.1): each utterance is transcribed by ASR and translated to English before scoring. For the *masked* variant (§5.4), foreign spans are removed prior to ASR; for the *full* variant, foreign spans are kept and translated back into English for comparison to the monolingual reference.

Metrics.

- **SacreBLEU** Post (2018): corpus-level BLEU with standardized tokenization; we report the SacreBLEU signature and case-sensitive scores.
- **BERTScore** Zhang et al. (2019): token-level alignment via contextual embeddings; we use the multilingual model, report F1, and apply inverse document frequency (IDF) weighting.
- **COMET** Rei et al. (2020): a reference-based learned metric trained on human judgments (adequacy/fluency). We use a publicly released COMET checkpoint and report the mean segment score.
- **LaBSE similarity** Feng et al. (2022): cosine similarity between sentence embeddings from the LaBSE encoder; applied on English translations to capture discourse-level semantic alignment beyond n-grams.

Aggregation. Scores are computed per segment (for §5.3) or per utterance (for §5.4) and then averaged across the evaluation set. Corpus BLEU is reported via SacreBLEU; all other metrics are macro-averaged over items.

Bootstrap resampling and confidence intervals. We estimate 95% confidence intervals (CIs) via nonparametric bootstrap with $B{=}1000$ resamples:

1. Sample with replacement the same number of items as the original set (segment- or utterance-level, matching the evaluation).
2. Recompute the aggregate metric on each bootstrap sample.
3. Form the CI using the percentile method (2.5% / 97.5% quantiles).

When multiple languages are pooled, we use a stratified bootstrap (sampling within each language and recombining) to preserve the original language mix.

Caveats. Because scoring uses ASR→MT text, absolute values may reflect downstream model bias (e.g., stronger Swahili ASR/MT yields higher scores). Cross-condition comparisons are still informative because the same pipeline is applied to all systems and variants.

G.1 Tolerance Selection for Cross-Lingual Segment Substitution

To set the duration tolerance used in the c_2 hinge penalty (cf. §3.2), we estimate a language-pair–specific threshold $\lambda_d(\ell_x, \ell_y)$ that reflects typical tempo/prosody differences between a *monolingual* language ℓ_x and an *infusion* language ℓ_y.

Base tolerance. Let $\tilde{d}_\ell$ denote a robust average segment duration for language ℓ (we use the median over segments; trimmed mean also works). Define

$$\lambda_{\text{base}}(\ell_x, \ell_y) \;=\; \frac{\big|\tilde{d}_{\ell_x} - \tilde{d}_{\ell_y}\big|}{\tilde{d}_{\ell_x}} \,.$$

Final tolerance. We add a small safety margin and enforce a minimum window:

$$\lambda_d(\ell_x, \ell_y) \;=\; \max\Big(\lambda_{\text{base}}(\ell_x, \ell_y) + \epsilon,\; \lambda_{\min}\Big), \qquad \epsilon = 0.05,\;\; \lambda_{\min} = 0.10.$$

This λ_d is used in the duration hinge of §3.2: $\mathcal{L}_{\text{dur}} = \max\big(0,\; \frac{|d_c - \hat{d}| - \lambda_d\, \hat{d}}{\hat{d}}\big)$.

Notes. (i) We compute $\hat{d}$ with the tempo/prosody normalization S_{ratio} (Eq. **??**), then apply λ_d to $\hat{d}$; this avoids double-counting rate differences. (ii) λ_d is asymmetric in (ℓ_x, ℓ_y), which matches the directed substitution setting.

Table 9: Computed λ_d for Swahili as the monolingual language. Durations in seconds.

Infusion ℓ_y	$\tilde{d}_{\ell_y}$	λ_{base}	λ_d (final)
Luo	16.5	0.0855	0.1355
Nandi	17.1	0.1250	0.1750
Kikuyu	15.6	0.0263	0.1000
English	15.0	0.0132	0.1000
Swahili robust duration: $\tilde{d}_{\ell_x} = 15.2$			

H Ablation

H.1 Effect of Removing the Language-Classifier Constraint

We study the role of $c_1(x, y)$ (foreignness gating and global rate). In the *classifier-free* variant we remove c_1 from the guidance and select the edited segment per step uniformly at random (one span per step; cf. §3.2). The inference objective becomes

$$F(\eta) \;=\; \mathbb{E}_{t\sim\text{Unif}\{1:T\},\,\epsilon_0\sim\mathcal{N}(0,I)}\big[\|\hat{\epsilon}_\theta(x_t, t) - \epsilon_0\|_2^2\big] \;+\; w(t)\,\lambda_2\,\mathcal{L}_{c_2}(\eta, y),$$

i.e., identical to the full model but without the c_1 term (and no $g^{(i)}$ gates).

Setup. From the 8,500 synthetic CS utterances, we sample 2,000 per source language (N= 8,000 total). We detect language spans with the frozen LID model f_{cl} and compute: (i) **CS frequency** = proportion of foreign frames (or ASR tokens) in the utterance; (ii) **alternation rate** = number of language switches per utterance; (iii) **human ratings** (fluency, coherence, realism; 5-point Likert, §6.1). We report means with 95% bootstrap CIs (1,000 resamples).

Table 10: Ablation of the language-classifier constraint (c_1). Means [95% CI].

Metric	**Full model**	**No** c_1	Δ
CS frequency (%)	4.33 [4.10, 4.58]	18.3 [17.6, 18.9]	+13.97
Alternation rate (#/utt)	3.88 [3.74, 4.02]	17.9 [17.2, 18.6]	+14.02
Fluency (↑)	4.10 [4.07, 4.13]	2.70 [2.65, 2.75]	-1.40
Coherence (↑)	4.05 [4.01, 4.09]	3.30 [3.24, 3.36]	-0.75
Realism (↑)	4.00 [3.96, 4.04]	2.60 [2.54, 2.66]	-1.40

Findings. Removing c_1 causes substantial *over-switching* (CS frequency +14 pp; alternation rate +14) and markedly degrades perceived quality (fluency -1.40, realism -1.40, coherence -0.75). This confirms that c_1 is essential for regulating *where/how much* to switch and for preserving utterance-level fluency and discourse coherence.

H.2 Effect of Removing Temporal Alignment and Onset Constraints

We quantify the contribution of the timing terms in c_2 that enforce prosodic alignment between the inserted foreign segment and the host utterance. Concretely, we remove the duration and onset penalties by setting $\gamma = 0$ in

$$\mathcal{L}_{c_2} \;=\; \alpha\,\mathcal{L}_{\text{semantic}} + \beta\,\mathcal{L}_{\text{context}} + \gamma\big(\mathcal{L}_{\text{dur}} + \mathcal{L}_{\text{on}}\big) \quad \text{(see §3.2)},$$

yielding the simplified guidance

$$\mathcal{L}_{c_2} \;=\; \alpha\,\mathcal{L}_{\text{semantic}} + \beta\,\mathcal{L}_{\text{context}}.$$

This ablation permits insertions with unconstrained duration and onset (no explicit prosodic guidance). We generate 8,500 code-switched utterances and evaluate both segment- and utterance-level quality using the protocols in §5.3 and §5.4. Table 11 reports averages across synthetic languages with 95% bootstrap CIs.

Table 11: Impact of removing duration/onset constraints on segment- and utterance-level metrics. Mean with 95% CIs (1,000 bootstrap samples).

Level	Metric	Avg. With Duration/Onset	Avg. Without Duration/Onset
Segment	SacreBLEU (↑)	36.4 [35.2, 37.4]	34.6 [33.3, 35.7]
	BERTScore (↑)	0.807 [0.804, 0.810]	0.796 [0.792, 0.800]
	COMET (↑)	0.815 [0.807, 0.823]	0.790 [0.781, 0.799]
	LaBSE (↑)	0.880 [0.873, 0.886]	0.868 [0.860, 0.874]
Utterance	SacreBLEU (↑)	34.5 [33.5, 35.6]	32.8 [31.5, 33.9]
	BERTScore (↑)	0.755 [0.751, 0.759]	0.743 [0.738, 0.748]
	COMET (↑)	0.653 [0.645, 0.661]	0.624 [0.614, 0.634]
	LaBSE (↑)	0.874 [0.870, 0.878]	0.860 [0.854, 0.867]

Table 12: Human ratings (Likert 1–5) comparing the full model vs. the variant without duration/onset constraints.

Metric	Full Model	Without Duration/Onset	Difference (Δ)
Avg. Fluency	4.1	3.1	-1.0
Avg. Coherence	4.05	3.3	-0.75
Avg. Realism	4.0	2.4	-1.6

Removing timing constraints consistently degrades both segment- and utterance-level metrics (Table 11). While segment-level drops are modest (e.g., -1.8 SacreBLEU, -0.019 COMET), utterance-level scores are more sensitive to prosodic disruption (-0.029 COMET, -0.012 BERTScore), suggesting that small duration/onset mismatches propagate across the utterance and impair discourse-level fluency. Human judgments (Table 12) corroborate this: fluency and realism decline sharply, with listeners noting abrupt transitions and unnatural pacing. Overall, explicit timing control is crucial for producing fluent, natural-sounding code-switched speech; semantics and context alone are insufficient without prosodic alignment.

I Related Work

Code-switching is a well-documented linguistic phenomenon in multilingual communities, particularly across Africa, where speakers frequently alternate between local vernaculars and national or international languages such as English or Swahili. Foundational work by Slabbert & Finlayson (1999) and Myers-Scotton (1993) highlighted code-switching as a communicative strategy influenced by identity, context, and pragmatics. Poplack (1980) and Auer (1998) further explored structural patterns and conversational dynamics, establishing typologies of alternation, insertion, and congruent lexicalization. These studies underscore the naturalness and linguistic richness of code-switching in African speech.

Despite its sociolinguistic prominence, *code-switching has been underrepresented in computational speech research*, largely due to the lack of annotated corpora and standardized tools. While progress has been made in *code-switched text generation* using statistical or neural methods (Tarunesh et al., 2021; Gregorius & Okadome, 2022; Chi et al., 2023), the *speech modality* remains significantly underexplored.

The most notable contribution to *code-switched speech synthesis* is by Cao et al. (2020), who proposed a bilingual phonetic posteriorgram-based model that combines monolingual speech corpora to generate mixed-language speech. However, their method lacks explicit semantic or contextual alignment and does not account for speaker consistency or natural prosodic transitions across languages.

In contrast, our work introduces a *diffusion-based framework* that synthesizes code-switched speech by minimally editing monolingual utterances. We incorporate *linguistic constraints*—a pre-trained language classifier for soft switch control and a multilingual encoder for semantic segment matching—to guide the generation process. Additionally, we address *speaker identity harmonization* by introducing a refinement step based on acoustic conditioning.

To the best of our knowledge, this is the first work that enables the infusion of multiple foreign languages within a single utterance, allowing for rich, naturalistic multilingual code-switching patterns. This represents a significant advancement toward realistic speech generation in low-resource multilingual settings.

J LIMITATIONS

Our proposed framework for controlled code-switched speech generation has demonstrated strong quantitative and human evaluation performance. However, several limitations remain:

Mismatch Between Synthesized and Natural Speech The generated utterances, while fluent and semantically faithful, are synthesized from noise and do not inherit the rich socio-pragmatic cues, emotional tone, or discourse-driven switching patterns present in natural conversations. This limits the realism of certain paralinguistic features such as emphasis, hesitation, or spontaneous repairs.

No Parallel Code-Switched Supervision The model is trained entirely on monolingual utterances without access to parallel code-switched references. This weak supervision constrains the model's ability to learn context-specific switching behavior beyond what is imposed by local segment similarity and predefined constraints.

Language and Domain Generalization Our study focuses on five Kenyan languages in a broadcast news context. While this setting ensures clean and aligned data, the model may not generalize to informal, multi-party, or highly emotional speech domains without further tuning or retraining.

Segment-Level Constraints Without Syntax Awareness Although segment replacement is guided by semantic and prosodic alignment, the model does not enforce syntactic compatibility between the inserted segment and surrounding context. This may occasionally result in grammatically awkward utterances, particularly in morphologically rich languages.

Speaker Identity Harmonization Is Post Hoc While a refinement step is used to harmonize speaker identity, it is applied after generation and not jointly optimized with the diffusion process. As a result, subtle speaker inconsistencies may persist across segments in certain cases.

Metrics May Not Capture Cultural or Pragmatic Fit Automated evaluation metrics (e.g., COMET, LaBSE) and even human Likert ratings may overlook deeper cultural or conversational appropriateness of switches. For instance, switching at discourse boundaries or for emphasis may be underrepresented in synthetic data.