ICML 2026

STAR-VAE: Structured Topology-Aware Regularization for Audio Reconstruction and Generation

Huadai Liu[1,2] Wen Wang[2] Kaicheng Luo[2] Qian Chen[2] Xiangang Li[2] **Wei Xue**[1]

[1]Hong Kong University of Science and Technology · [2]Tongyi Fun Team, Alibaba Group

PROJECT PAGE & AUDIO SAMPLES

↗ STAR-VAE.github.io

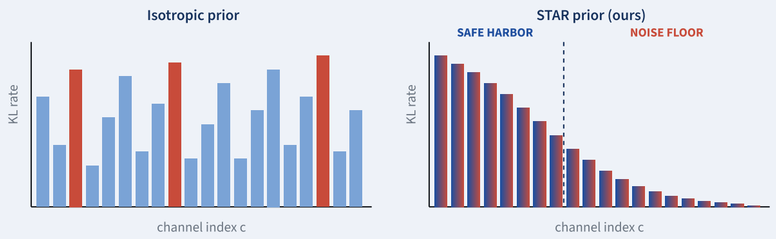
1 The Rate–Distortion–Regularity Trilemma

High-fidelity audio generation is upper-bounded by the **continuous tokenizer**. Standard VAEs juggle three objectives at once — and the **isotropic Gaussian prior** imposes a flat geometry that ignores audio's spectral hierarchy.

RATE ↓ · DISTORTION ↓ · REGULARITY ↑

- **Low-freq** components → structured, low-entropy, compressible.
- **High-freq** components → stochastic, high-entropy, incompressible.
- Uniform KL ⇒ *disordered information packing*: semantic features randomly interleaved with high-entropy noise.

Channel-wise KL · before vs. after STAR



Channel-wise KL divergence on AudioCaps. STAR yields monotonic capacity decay — structure to head channels, texture to tail.

2 STAR Regularization

Standard VAE compound objective (Eq. 1):

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Rec}}(x, \hat{x}) + \lambda_{\text{Adv}} \cdot \mathcal{L}_{\text{Adv}}(x, \hat{x}) + \beta \cdot \mathcal{L}_{\text{KL}}(q_\phi \| p) \quad (1)$$

STAR replaces scalar β with a **channel-indexed** penalty vector $\beta \in \mathbb{R}^C$ via a **Gamma-Growth function** (Eq. 4):

$$\beta_c = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \left((c-1)/(C-1) \right)^{\gamma} \quad (4)$$

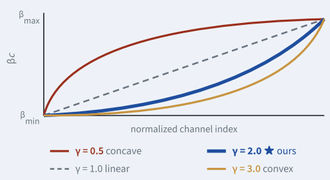
$\beta_{\min}$ 1e-4 · $\beta_{\max}$ 4e-4 · γ 2.0 (convex)

The KL term in Eq. 1 is replaced by the structured objective (Eq. 5):

$$\mathcal{L}_{\text{STAR}} = \sum_{c=1}^{C} \beta_c \cdot D_{\text{KL}}(q_\phi(z_c|x) \| \mathcal{N}(0,1)) \quad (5)$$

Growth function · γ chooses the allocation

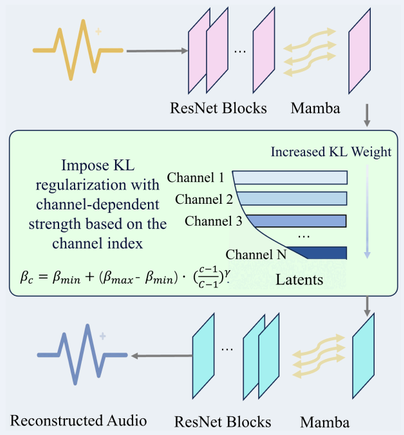
Theory → audio energy follows a heavy-tailed ($\approx 1/f$) power law, so optimal capacity allocation is *convex* ($\gamma > 1$). Concave/linear allocations underfit structure — ablations in Table 4 confirm $\gamma = 2.0$ is optimal.



3 STAR-VAE & STAR-Gen

STAR-VAE · hybrid CNN-Mamba tokenizer

SAO convolutional backbone (5 blocks, ↓2048×) + bidirectional **Mamba** adapters. STAR bottleneck applied in **Phase II** fine-tuning.



STAR-VAE: hybrid CNN-Mamba tokenizer with the STAR bottleneck (Phase II). Figure 2 (left).

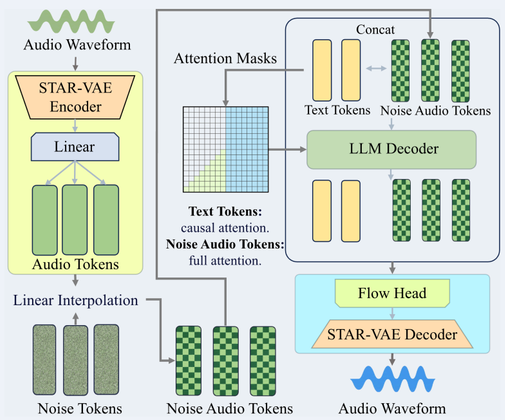
Reconstruction Drift → high-capacity Mamba under *isotropic* KL spontaneously sacrifices high-entropy textures to minimize the uniform penalty, yielding semantically coherent but texturally hollow audio. STAR's capacity gradient eliminates this pathology — see Table 1 (Hybrid CNN-Mamba *w/o* STAR).

STAR-Gen · LLM-based Flow Matching

A causal **Qwen3** decoder, repurposed as a conditional velocity estimator $v_\theta(z_t, t \mid c)$ over continuous STAR latents — no vector quantization (Eq. 7):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,z_0,z_1} \| v_\theta(z_t, t \mid c) - (z_1 - z_0) \|^2 \quad (7)$$

$z_T = (1-T) \cdot z_0 + T \cdot z_1$ · LOGIT(T) ~ $\mathcal{N}(0,1)$ · $z_0 \sim \mathcal{N}(0, I)$



STAR-Gen: Qwen3 LLM decoder as a conditional velocity estimator over continuous STAR latents. The attention mask (causal on text, full on audio) is applied inside the LLM decoder. Figure 2 (right).

4 Empirical Results

Reconstruction · Table 1 (44.1 kHz eval)

MODEL	RATE	AUDIOCAPS · SOUND			SONG DESCRIBER · MUSIC		
		STFT↓	FAD↓	LC↓	STFT↓	FAD↓	LC↓
Mel-VAE	31.2 Hz	2.53	2.86	0.33	3.04	0.84	0.25
AudioGen	100 Hz	2.18	2.36	0.06	2.62	1.16	0.02
ɛar-VAE	43 Hz	1.08	4.44	0.13	0.96	0.29	0.11
Stable Audio Open	21.5 Hz	1.25	3.29	0.11	1.59	0.69	0.09
Hybrid CNN-Mamba (w/o STAR)	21.5 Hz	1.35	2.74	0.10	1.57	0.39	0.10
CNN-STAR (w/o Mamba)	21.5 Hz	1.22	2.65	0.09	1.40	0.38	0.08
CNN-VAE (w/o STAR & Mamba)	21.5 Hz	1.28	3.36	0.11	1.46	0.45	0.12
Ours STAR-VAE	21.5 Hz	**1.17**	**2.31**	**0.08**	**1.32**	**0.25**	**0.08**

Text-to-audio generation · Table 2 (AudioCaps)

MODEL	FD_{OPENL3} ↓	KL ↓	CLAP ↑
AudioLDM 2-Large	108.3	1.81	0.42
Tango 2	108.4	1.11	0.44
TangoFlux	80.2	1.22	0.43
Stable Audio Open (SAO)	89.2	2.58	0.29
SAO w/ STAR-VAE	72.5	2.15	0.35
STAR-Gen w/ ɛar-VAE	76.5	1.53	0.41
STAR-Gen w/ SAO-VAE	67.4	1.21	0.44
Ours STAR-Gen	**55.8**	**1.09**	**0.48**

Backbone ablation · Song Describer (Table 3)

ARCHITECTURE (ALL W/ STAR)	STFT-D ↓	FAD ↓	INFER (S) ↓
CNN-STAR (no Mamba)	1.40	0.38	**0.68**
Transformer-STAR	1.35	0.30	0.92
Mamba-STAR · STAR-VAE	**1.32**	**0.25**	0.85

Growth-function ablation · AudioCaps (Table 4)

GROWTH FUNCTION	STFT-D ↓	FAD ↓	LC ↓
Step	1.58	3.15	0.11
Linear ($\gamma = 1.0$)	1.38	2.65	0.08
$\gamma = 0.5$ · concave	1.42	2.75	**0.07**
$\gamma = 1.5$	1.32	2.45	0.08
$\gamma = 2.0$ · ours	**1.17**	**2.31**	0.08
$\gamma = 3.0$	1.25	2.52	0.09

CONCLUSION

STAR traces the Rate–Distortion–Regularity Trilemma to the isotropic prior and resolves it with a capacity gradient matched to audio's spectral hierarchy — generalizing across architectures and domains to set a new continuous-tokenization paradigm for neural audio generation. Validated by linear probing (70.3% vs. 65.0%) and human MOS (4.32 vs. 4.05).

DIAGNOSIS

Continuous audio VAEs face the Rate–Distortion–Regularity Trilemma: isotropic priors fragment audio's spectral hierarchy into disordered latents.

FIX · STAR

A channel-indexed Gamma-Growth KL ($\gamma = 2.0$, $\beta_{\min} = 1e{-}4$, $\beta_{\max} = 4e{-}4$) routes structure to head channels and texture to tail channels.

RESULT

STAR-VAE + STAR-Gen set SOTA reconstruction on AudioCaps & Song Describer (FAD 2.31 / 0.25) and text-to-audio generation (FD_{openl3} 55.8) — all at 21.5 Hz.