IAE: INTERPOLATIVE AUTOENCODER FOR IMPROVING LATENT CONTINUITY IN AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformer (DiT) based audio generation relies on a pretrained autoencoder to map audio into a latent space, where the latent quality directly influences generation performance. Recent work has significantly improved the semantic structuring of representations through distillation or downstream tasks. However, representation continuity has seen little further progress. Continuity matters because the decoder must robustly reconstruct generation-biased latent caused by imperfect training or sampling of DiT. Existing continuity methods still rely on fixed-magnitude Gaussian noise injection or early KL regularization, both of which may reduce reconstruction fidelity when decoding biased latent, thereby degrading final generation quality. In this paper, we propose the Interpolative Autoencoder (IAE), which preserves the semantic structure of the pretrained encoder while improving representation continuity through semantic reconstruction on latent-pair interpolations rather than single-point perturbation. IAE pairs each latent with its nearest neighbor in the current encoding distribution and imposes semantic reconstruction supervision at interpolation points between pairs. This design allows the noise magnitude to adapt according to the semantic structure around each sample's neighborhood, effectively filling low-density regions between neighboring samples and alleviating the problem of decoding generation-biased latent. From an optimization perspective, IAE is essentially optimizing the interpolated Fréchet Audio Distance (iFAD) indirectly. This metric has been shown in the image domain to assess the structuring and continuity of the latent space at the representation learning stage and to correlate significantly with final generation quality. Experimental results confirm this connection. Compared with existing RAE and VAE variant baselines, IAE achieves a lower iFAD while also obtaining better gFAD and WER on audio generation tasks, demonstrating that by improving representation continuity IAE can tangibly enhance generation quality.

1 INTRODUCTION

Latent generative modeling has become the standard paradigm for diffusion-based audio synthesis (Liu et al., 2023; Ghosal et al., 2023; Dinkel et al., 2026). A pretrained autoencoder first compresses waveforms into a continuous latent space, and a diffusion or flow model is then trained within that space. A growing body of research (Zhong et al., 2026; Yue et al., 2026b) shows that a latent space that is both well-structured and sufficiently continuous tends to improve final generation quality. Recent work has made significant progress on semantic structuring. In both the image and audio domains, aligning latent variables to frozen semantic features or introducing semantic downstream tasks allows semantically similar inputs to occupy neighboring latent positions and improves generation performance (Yao et al., 2025; Leng et al., 2025). Semantic structuring has become a mature means of enhancing latent quality.

However, latent space continuity has seen little corresponding progress, despite being equally critical to generation quality (Gu et al., 2026). The autoencoder trains its decoder only on encoded latent variables, yet DiT optimization and sampling are both imperfect and therefore return latent variables that deviate from clean encoding points. The decoder must robustly reconstruct these generation-biased latent variables. When the latent space is not continuous, a deviation of the same magnitude can carry a sample further away from the region where the decoder was trained, and reconstruction fidelity drops accordingly (Guo et al., 2024; Ning et al., 2024). Existing continuity mechanisms

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.

are still the single-point perturbations introduced for earlier systems. One family, represented by Representation Autoencoders (RAE) (Zheng et al., 2026) and spherical representations (Yue et al., 2026a), injects Gaussian noise at a fixed magnitude during decoder training. The other is the KL regularization of Variational Autoencoders (VAE) (Yao et al., 2025; Doersch, 2021). These approaches perturb each encoding point based solely on that latent variable itself, without explicitly referencing the surrounding encoding distribution. This can leave substantial low-density regions in the representation space. As a result, the perturbation range to which the decoder has adapted may not cover the larger deviations produced by imperfect training and sampling during diffusion generation, causing reconstruction fidelity to degrade for generation-biased latent variables and ultimately reducing generation quality.

In this paper, we propose the Interpolative Autoencoder (IAE). IAE preserves the semantic structure provided by the pretrained encoder while improving representation continuity through semantic reconstruction on latent-pair interpolations rather than single-point perturbation. IAE maintains a queue composed of recent-batch latent variables to approximate the current encoding distribution. Each latent variable retrieves its nearest neighbor from this queue, and an interpolation point between the pair is decoded. The interpolated latent variable has no corresponding waveform ground truth, but the semantic content between the paired samples should remain consistent. This branch therefore does not carry an acoustic reconstruction objective. Instead, a frozen semantic encoder constrains the decoded output at the interpolation point to preserve the semantic content of the source audio. Because the interpolation point lies between nearest-neighbor encoding pairs, the perturbation magnitude adapts to the semantic structure around each sample's neighborhood rather than following a data-agnostic preset scale, effectively filling low-density regions between neighboring samples and further improving the continuity of the representation space. Clean latent variables continue to carry the acoustic reconstruction objective as usual and are aligned to the same frozen encoder to prevent the pretrained semantic structure from collapsing.

The design of IAE is motivated by the recently proposed interpolated Fréchet Inception Distance (iFID) (Xu et al., 2026). Interpolated FID retrieves the nearest neighbor for each latent variable in the latent space, decodes the midpoint between the two, and computes the Fréchet distance against real data. We denote its audio instantiation as iFAD. In the image domain, this quantity has been shown to assess the structuring and continuity of the latent space at the representation learning stage and to correlate strongly with diffusion generation quality. IAE approximates the current encoding distribution by maintaining a queue of recent-batch latent variables. The semantic reconstruction supervision that IAE imposes on interpolation points indirectly reduces the Fréchet distance to real data while further preserving the pretrained semantic structure, ultimately improving the representation's fitness for diffusion generation through better iFAD. Our contributions are as follows.

- We propose IAE, a framework for improving the continuity of structured representations. Unlike existing continuity methods based on single-point perturbation, IAE retrieves nearest neighbors from the current encoding distribution to form latent pairs, allowing the perturbation magnitude to adapt to the semantic structure of each sample's neighborhood and thereby filling low-density regions between neighboring samples.
- We design semantic reconstruction supervision at interpolation points as the training objective. Because interpolated latent variables have no waveform ground truth, IAE applies only semantic-level constraints to them, preserving the pretrained encoder's semantic structure while avoiding the introduction of acoustic noise, and constituting an indirect optimization of iFAD from an optimization-objective perspective.
- Experimental results show that IAE achieves a lower iFAD compared with existing RAE and VAE variant baselines. In downstream audio generation, IAE also obtains better gFAD and WER, confirming that the continuity improvement brought by IAE translates into improved generation quality.

2 METHOD

Latent variables sampled by DiT can deviate from clean encodings and fall into low-density regions between neighboring encodings, where the decoder receives no supervision during training. IAE addresses this by extending decoding supervision into these regions, as illustrated in Figure 1. To locate them, IAE pairs each encoding with its nearest neighbor and samples interpolation points

between the pair (Section 2.1). These points have no waveform ground truth, so IAE supervises them with the semantics of the source audio rather than its acoustics (Section 2.2). Because this supervision acts only on semantics, the clean encodings must still retain the pretrained semantic structure and acoustic fidelity (Section 2.3). Finally, since iFAD evaluates exactly these interpolation points, the resulting design serves as an indirect optimization of iFAD (Section 2.4).

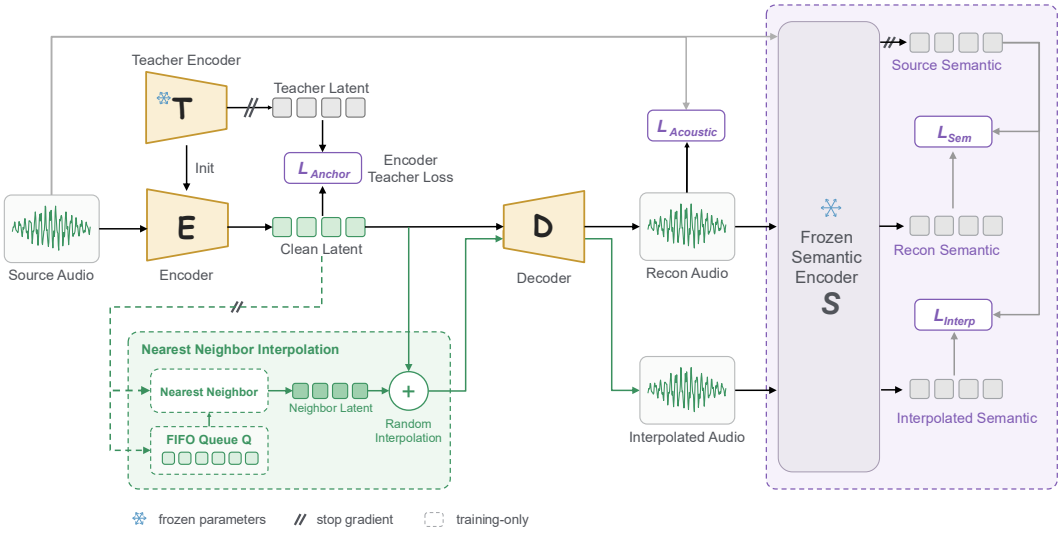


Figure 1: **Overview of IAE.** Each encoding is paired with its nearest neighbor from a queue of recent encodings, and a randomly sampled interpolation point is decoded under semantic supervision from the frozen encoder S. Clean encodings keep acoustic reconstruction and encoder anchoring.

2.1 Neighborhood-Adaptive Latent Interpolation

For an audio waveform x, let

$$z = E_\theta(x) \in \mathbb{R}^{L\times d}, \qquad \hat{x} = D_\phi(z), \tag{1}$$

where E_θ and D_ϕ denote the encoder and decoder, θ and ϕ are their trainable parameters, L is the latent sequence length, and d is the feature dimension.

Pairing within the encoding distribution. We maintain a FIFO queue $\mathcal{Q}$ of encodings from recent batches. The latent variables of each new batch are enqueued to update the queue, so that $\mathcal{Q}$ continually tracks the current state of the representation space. The retrieval key of each latent sequence is obtained by temporal averaging followed by ℓ_2 normalization:

$$k(z) = \frac{\frac{1}{L}\sum_{t=1}^{L} z_t}{\left\|\frac{1}{L}\sum_{t=1}^{L} z_t\right\|_2}. \tag{2}$$

The nearest neighbor is selected from the queue by cosine similarity, excluding entries from the current sample:

$$q^\star = \underset{q\in\mathcal{Q},\, \mathrm{id}(q)\neq\mathrm{id}(x)}{\arg\max} k(z)^\top k(q), \tag{3}$$

where $\mathrm{id}(\cdot)$ denotes the source-sample identifier. When the temporal length of $q^\star$ differs from that of z, $q^\star$ is temporally resampled to match it, yielding the neighbor sequence z^+ used for interpolation. Retrieval uses the pooled keys, whereas interpolation operates on the full latent sequences.

Sampling between paired representations. For each source sample, we independently sample $\alpha \sim \mathcal{U}(0, \frac{1}{2}]$ and construct the interpolation point

$$\tilde{z} = (1-\alpha)z + \alpha\,\mathrm{sg}[z^+], \tag{4}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation, which prevents z^+ from collapsing toward z during training. Restricting α to $(0, \frac{1}{2}]$ keeps the interpolation point on the source-side half of the segment, ensuring that the semantic supervision target derived from x remains reliable. The induced displacement magnitude is

$$\|\Delta z\|_F = \alpha \|z^+ - z\|_F. \tag{5}$$

Unlike fixed-magnitude Gaussian noise injection, the perturbation scale here adapts to the distance between each pair of neighboring encodings. By imposing decoding supervision at these intermediate positions, IAE extends coverage to the low-density regions between neighboring samples, which are precisely where generation-biased latent variables are likely to fall.

2.2 Semantic Reconstruction at Interpolation Points

Interpolating between the latents of neighboring samples alters acoustic details such as phase and timing, so $\tilde{z}$ has no corresponding waveform ground truth. However, since z^+ is selected by semantic similarity, the interpolation point should remain semantically close to the source audio. Therefore, rather than enforcing waveform reconstruction at a location without acoustic ground truth, we impose semantic supervision on the decoded output $\tilde{x} = D_\phi(\tilde{z})$, requiring it to preserve the semantic content of x.

We use a frozen Dasheng encoder S as the semantic encoder that provides the targets; S is initialized from the same pretrained checkpoint as E_θ. For feature sequences $A, B \in \mathbb{R}^{L \times d}$, we define the feature distance

$$\ell_{\mathrm{feat}}(A, B) = \frac{1}{Ld}\|A - B\|_F^2 + \frac{1}{L}\sum_{t=1}^{L}\left(1 - \frac{A_t^\top B_t}{\|A_t\|_2 \|B_t\|_2}\right), \tag{6}$$

which constrains both the magnitude and the direction of the features. The interpolation loss is

$$\mathcal{L}_{\mathrm{interp}} = \mathbb{E}_{x,\alpha}\left[\ell_{\mathrm{feat}}\left(S(\tilde{x}), \mathrm{sg}[S(x)]\right)\right], \tag{7}$$

where $S(x)$ serves as the semantic target. The clean reconstruction $\hat{x}$ receives the same semantic supervision:

$$\mathcal{L}_{\mathrm{sem}} = \mathbb{E}_{x}\left[\ell_{\mathrm{feat}}\left(S(\hat{x}), \mathrm{sg}[S(x)]\right)\right]. \tag{8}$$

The parameters of S remain frozen, and gradients flow through $S(\tilde{x})$ and $S(\hat{x})$ to update the encoder and decoder.

2.3 Preserving Semantic Structure and Acoustic Fidelity

The interpolation branch carries only semantic signals without acoustic supervision, and leaving the encoder unconstrained may destroy the pretrained semantic structure. We prevent this with two objectives on the clean reconstruction path.

Encoder anchoring. The trainable encoder is aligned with the frozen pretrained features:

$$\mathcal{L}_{\mathrm{anchor}} = \mathbb{E}_{x}\left[\ell_{\mathrm{feat}}\left(E_\theta(x), \mathrm{sg}[T(x)]\right)\right], \tag{9}$$

where T is the frozen teacher encoder in Figure 1. This objective preserves the pretrained semantic structure that benefits generation, while allowing the encoder to incorporate acoustic details through joint training with the decoder.

Clean reconstruction. For $\hat{x} = D_\phi(E_\theta(x))$, the acoustic objective is

$$\mathcal{L}_{\mathrm{acoustic}} = \lambda_{\mathrm{mel}}\mathcal{L}_{\mathrm{mel}} + \lambda_{\mathrm{gan}}\left(\mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{fm}}\right), \tag{10}$$

where $\mathcal{L}_{\mathrm{mel}}$ is the Mel-spectrogram reconstruction loss, $\mathcal{L}_{\mathrm{adv}}$ is the generator adversarial loss, and $\mathcal{L}_{\mathrm{fm}}$ is the discriminator feature-matching loss. Encoder anchoring prevents the semantic structure from collapsing, acoustic reconstruction maintains audio fidelity at clean encodings, and interpolation supervision drives the decoder to decode generation-biased latent variables without distortion. The complete objective is

$$\mathcal{L}_{\mathrm{IAE}} = \mathcal{L}_{\mathrm{acoustic}} + \lambda_{\mathrm{anchor}}\mathcal{L}_{\mathrm{anchor}} + \lambda_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{interp}}\mathcal{L}_{\mathrm{interp}}. \tag{11}$$

2.4 Interpolation Supervision as a Surrogate for iFAD

The interpolation supervision of IAE indirectly optimizes iFAD. iFAD retrieves the nearest neighbor for each encoding, decodes the midpoint between the two, and computes the Fréchet distance against real audio, thereby assessing the structuring and continuity of the representation space. A more structured representation yields higher semantic similarity between nearest-neighbor pairs, so the interpolated midpoint undergoes smaller semantic perturbation and the decoded output better preserves the original semantics, lowering iFAD. A more continuous representation makes the decoder more robust to interpolation points that deviate from the encoding distribution, so the decoded outputs exhibit no noticeable distortion, which also lowers iFAD. iFAD is defined as

$$\mathrm{iFAD} = \|\mu_{\mathrm{int}} - \mu_{\mathrm{real}}\|_2^2 + \mathrm{Tr}\left[\Sigma_{\mathrm{int}} + \Sigma_{\mathrm{real}} - 2\left(\Sigma_{\mathrm{real}}^{1/2}\Sigma_{\mathrm{int}}\Sigma_{\mathrm{real}}^{1/2}\right)^{1/2}\right], \tag{12}$$

where $(\mu_{\mathrm{int}}, \Sigma_{\mathrm{int}})$ and $(\mu_{\mathrm{real}}, \Sigma_{\mathrm{real}})$ denote the feature means and covariances of the decoded midpoints and real audio, respectively. In the image domain, iFID has been shown to correlate strongly with generation quality (Xu et al., 2026), making it an effective proxy for assessing how well a representation suits generation.

The design of IAE corresponds directly to the computation of iFAD. Nearest-neighbor pairing combined with semantic supervision trains the decoder to produce semantically faithful outputs at interpolation points, which are exactly the locations that iFAD evaluates. The midpoint used at evaluation,

$$z_{\mathrm{mid}} = \frac{1}{2}(z + z^{+}), \qquad x_{\mathrm{mid}} = D_\phi(z_{\mathrm{mid}}), \tag{13}$$

lies within the source-side interval covered during training (at its endpoint $\alpha = \frac{1}{2}$). Therefore, indirectly optimizing iFAD during training preserves the structuring of the representation space and improves its continuity, ultimately leading to better generation quality.

3 Experimental Setup

3.1 Tokenizer Training

The encoder is initialized from Dasheng (Dinkel et al., 2024) and its outputs serve as latent representations. The decoder follows the Vocos architecture (Siuzdak, 2024) implemented in Dasheng-Tokenizer (Dinkel et al., 2026). The frozen semantic teacher is initialized from the same pretrained checkpoint. Training data cover speech, music, and environmental sound (Table B.2), allowing us to examine the applicability of the same continuity-learning objective to different types of audio. Full architecture, optimization, and loss-weight details are provided in Appendix B.1.

3.2 Baseline Models

Existing continuity mechanisms mainly take two forms: Gaussian noise injection during decoder training, as in the noise-augmented decoding of Representation Autoencoders (RAE) (Zheng et al., 2026), and the KL regularization of Variational Autoencoders (VAE) (Yao et al., 2025). We construct baselines for the first form on the same backbone as IAE, and use VA-VAE for the second.

Anchored AE. RAE pairs a frozen pretrained representation encoder with a trained decoder and uses the pretrained representation directly as the generative latent space. Recent audio work (Dinkel et al., 2026; Yang et al., 2026a) adopts this paradigm but further shows that jointly training the encoder to absorb acoustic details such as phase and timbre improves generation quality without noticeably degrading semantic structure. We therefore train the encoder jointly with the decoder, while $\mathcal{L}_{\mathrm{anchor}}$ aligns its outputs with frozen Dasheng features to preserve the pretrained semantic structure. We refer to this RAE variant as **Anchored AE**. It reconstructs audio only from clean latents and applies no decoding supervision outside clean encodings. All non-VAE baselines and IAE share this backbone, including its encoder initialization, decoder, acoustic reconstruction, and encoder anchoring objectives. They differ only in the additional supervision on decoded outputs.

Anchored AE + Noise (0.25) and Anchored AE + Noise (0.35). Following the noise-augmented decoding of RAE (Zheng et al., 2026), these baselines add a noise reconstruction branch to Anchored AE. Gaussian noise is normalized to a fixed per-sample RMS of 0.25 or 0.35, and the perturbed latent is decoded under supervision from the source Mel spectrogram. The two magnitudes are used to examine whether a larger preset noise can better fill the low-density regions between neighboring encodings and improve the decoder's robustness to generation-biased latents.

VA-VAE (128-d). VA-VAE is a VAE variant that has become a widely adopted approach for semantic structural alignment in latent generative models (Yao et al., 2025). This baseline follows the audio VA-VAE configuration, which combines semantic alignment with KL regularization in a 128-dimensional latent space. Unlike the noise-based baselines above, it encourages continuity by regularizing the posterior toward a prior rather than by decoding perturbed latents.

3.3 Evaluation

We evaluate at both the tokenizer level and the downstream generation level.

At the tokenizer level, we use reconstruction FAD (rFAD) and iFAD to evaluate decoding at clean encodings and nearest-neighbor interpolants, respectively (Kilgour et al., 2019; Xu et al., 2026). Reporting both metrics tests whether improvements in interpolation quality come at the expense of clean reconstruction. We compute FAD using features from VGGish, AudioFlamingo3 (AF3), Qwen3-Omni, Dasheng, and ATST-Frame (Hershey et al., 2017; Goel et al., 2025; Xu et al., 2025a; Dinkel et al., 2024; Li et al., 2024) to reduce dependence on any single representation. Dasheng also provides training supervision, while the other four extractors assess whether improvements extend beyond the training feature space. Details of the evaluation protocol are provided in Appendix B.4.

At the generation level, we freeze each tokenizer and train a latent generative model using the DiT architecture of UniFlow-Audio (Xu et al., 2025b), with the same configuration and training budget for all methods (Appendix B.3). Evaluation covers AudioCaps (Kim et al., 2019), MusicCaps (Agostinelli et al., 2023), LibriTTS (Zen et al., 2019), and MECAT (Niu et al., 2026a). We use generation FAD (gFAD; Kilgour et al., 2019) to measure the distributional quality of generated audio, with FD and KL as auxiliary metrics. We report WER on LibriTTS and, for the speech-containing categories of MECAT, additionally WER and UTMOSv2 (Baba et al., 2024).

4 Results

4.1 Reconstruction and Interpolation Quality

Table 1: **Tokenizer reconstruction and interpolation quality.** We report rFAD and iFAD across five audio feature extractors.

	VGGish		AudioFlamingo3		Qwen3-Omni		Dasheng		ATST-Frame	
Method	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓
Anchored AE	0.1878	0.6588	6.9761	24.1651	0.0189	0.0774	0.0327	0.1311	3.3806	18.6802
Anchored AE + Noise (0.25)	0.1849	0.6472	6.4425	23.5086	0.0177	0.0767	0.0291	0.1314	2.7858	17.1932
Anchored AE + Noise (0.35)	0.1957	0.6986	6.2998	23.8290	0.0167	0.0770	0.0295	0.1306	2.9873	17.2571
VA-VAE (128-d)	0.4909	0.9992	10.2198	28.4798	0.0418	0.1082	0.0515	0.1592	8.9832	25.3319
IAE	**0.0950**	**0.5689**	**4.5273**	**16.3678**	**0.0092**	**0.0615**	**0.0093**	**0.0745**	**1.0580**	**13.8787**

Table 1 reports rFAD and iFAD across five feature spaces. IAE achieves the lowest iFAD under all five extractors, reducing it by 12.1% (VGGish), 30.4% (AudioFlamingo3), 19.8% (Qwen3-Omni), 43.0% (Dasheng), and 19.3% (ATST-Frame) relative to the strongest baseline in each column. The largest reduction appears on Dasheng, as expected, since the semantic reconstruction supervision of IAE operates directly in the Dasheng feature space. More importantly, iFAD also decreases in the four feature spaces not used during training, indicating that IAE improves the continuity of the latent space itself rather than overfitting to the supervising feature space. IAE also attains the lowest rFAD in every column, 28.1–68.0% below the strongest baseline, showing that interpolation supervision does not conflict with acoustic reconstruction at the encoding points.

Table 2: **Generation results on AudioCaps, MusicCaps, and LibriTTS.** Subscripts denote the feature extractor used for gFAD. LibriTTS columns report WER (%). Lower is better; best in bold.

	AudioCaps						MusicCaps						LibriTTS	
Method	FAD_{VGG}	FAD_{AF3}	FAD_{Qwen}	$FAD_{Dasheng}$	FD	KL	FAD_{VGG}	FAD_{AF3}	FAD_{Qwen}	$FAD_{Dasheng}$	FD	KL	clean	other
Anchored AE	9.7868	35.2784	0.1083	0.2816	71.00	3.29	4.4827	33.0231	0.1009	0.2272	38.83	1.58	17.05	18.19
+ Noise (0.25)	5.3008	36.0817	0.0847	0.2804	52.82	2.69	5.2162	34.5469	0.0865	0.2138	34.07	1.68	17.26	18.49
+ Noise (0.35)	5.8703	34.0570	0.0737	0.2563	47.97	2.59	4.1943	25.6922	0.0685	0.1615	28.92	1.61	20.82	22.26
VA-VAE (128-d)	10.0434	42.2147	0.1162	0.2800	64.33	3.19	4.8076	27.1748	0.0972	0.1454	37.78	1.56	15.50	20.57
IAE	**3.8399**	**30.6404**	**0.0673**	**0.2556**	**41.41**	**2.32**	**2.8734**	**20.1472**	**0.0379**	**0.1357**	**28.91**	**1.52**	**14.62**	**15.18**

The noise-based baselines also extend decoding supervision beyond clean encodings, yet the gain is limited. Increasing the noise RMS from 0.25 to 0.35 worsens iFAD under four of the five extractors. A fixed-magnitude perturbation applies the same scale of displacement to every sample, and a larger magnitude is more likely to push the perturbed point away from the true low-density regions between neighboring encodings. IAE instead lets the retrieved nearest neighbor determine the direction and magnitude of the perturbation, covering these regions more precisely. VA-VAE yields the highest rFAD and iFAD in every column. Its 128-dimensional latent space discards substantial information, and the continuity provided by KL regularization cannot compensate for this loss.

4.2 Downstream Audio Generation

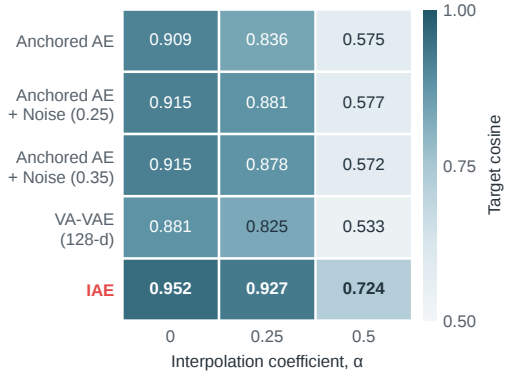
IAE achieves a gFAD of 3.84 on AudioCaps, reducing it by 60.8% from 9.79 for Anchored AE and by 27.5% from 5.30 for the better of the two noise baselines. On MusicCaps, IAE obtains a gFAD of 2.87, again the lowest among all methods. IAE attains the lowest gFAD in all four feature spaces on both datasets. FD and KL follow the same trend, with IAE achieving the best scores on both datasets. Across the seven MECAT categories, IAE attains the lowest gFAD and FD in five (Appendix C.2). Anchored AE and Anchored AE + Noise (0.25) have nearly identical reconstruction quality (VGGish rFAD of 0.1878 and 0.1849, respectively), yet the latter reduces gFAD on AudioCaps from 9.79 to 5.30 and FD from 71.00 to 52.82, indicating that generation quality depends not only on reconstruction fidelity at encoding points but also on the robustness of the decoder to latent variables that deviate from them. IAE achieves the lowest iFAD at the tokenizer level (Table 1) and the best gFAD at the generation level, confirming that improved representation continuity directly leads to better generation quality.

IAE also shows a clear advantage in speech intelligibility. It achieves the lowest WER on all evaluation sets for which WER is reported, reducing it from 17.05 to 14.62 (clean) and from 18.19 to 15.18 (other) on LibriTTS compared with Anchored AE. The noise baselines see WER degrade as the noise magnitude increases, for example rising from 17.26 to 20.82 (clean) and from 18.49 to 22.26 (other) on LibriTTS. This suggests that acoustic reconstruction supervision at larger perturbation points introduces distortion. IAE applies only semantic reconstruction supervision at interpolation points, enabling the decoder to preserve speech content when latent variables deviate from clean encodings. UTMOSv2 corroborates this finding. IAE obtains the highest score in every speech-containing MECAT category (Table C.2), indicating that the generated speech is not only more accurate but also more natural.

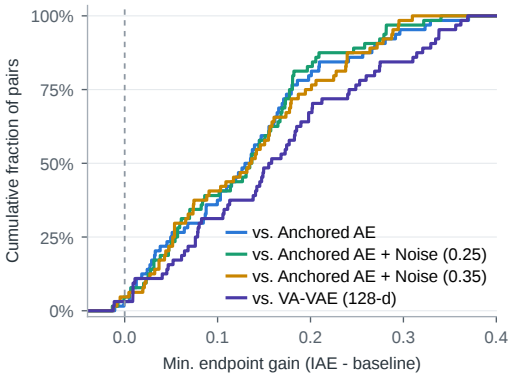
4.3 Decoding Along Interpolation Paths

To verify that IAE fills the low-density regions between neighboring encodings, we decode 64 nearest-neighbor audio pairs at $\alpha \in \{0, 0.25, 0.5\}$ and measure the cosine similarity between the output and a semantic reference that mixes the frozen Dasheng features of the two endpoints according to α. Clean reconstruction is evaluated on 240 clips. Results are shown in Figure 2.

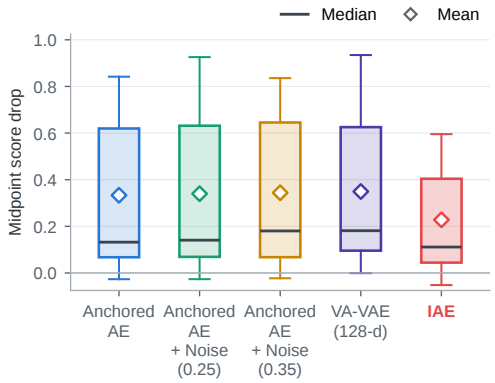
In Figure 2a, the similarity of the baselines falls from 0.88–0.92 at the source to 0.53–0.58 at the midpoint, while IAE retains 0.724. The noise baselines exceed Anchored AE at $\alpha = 0.25$ (0.878–0.881 versus 0.836) but remain nearly unchanged at the midpoint (0.572–0.577). Fixed-magnitude perturbation therefore covers only the vicinity of clean encodings, not the low-density regions between neighboring encodings. IAE achieves the highest similarity at both positions.



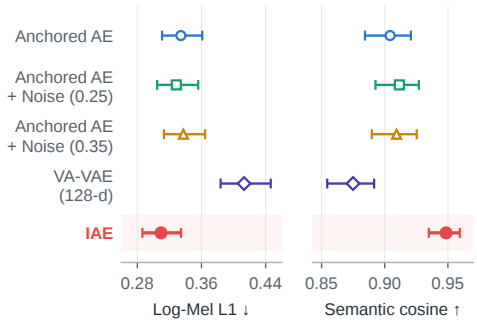
(a) Similarity along the path



(b) Midpoint min-endpoint gain



(c) Midpoint score drop



(d) Clean reconstruction

Figure 2: **Decoding along interpolation paths and at clean encodings.** (a) Semantic similarity between decoded audio and the reference at interpolation coefficients $\alpha \in \{0, 0.25, 0.5\}$. (b) Per-pair gain in min-endpoint similarity at the midpoint. (c) Endpoint score minus midpoint score. (d) Clean reconstruction quality at encoding points.

This gain is not driven by a few pairs. Scoring each pair by the lower of its midpoint similarities to the two endpoints, IAE outperforms every baseline on 61–62 of the 64 pairs (Figure 2b). The drop from the mean endpoint score to the midpoint score has an upper quartile of 0.40 for IAE, compared with 0.62–0.65 for the baselines (Figure 2c), indicating that severe semantic degradation is largely suppressed. IAE also attains the best clean reconstruction, with the lowest log-Mel L1 error (0.310 versus 0.329–0.413) and the highest semantic similarity (0.949 versus 0.875–0.912) (Figure 2d). Together, these results show that IAE fills the low-density regions between neighboring encodings without sacrificing reconstruction at clean encodings, which explains its lower iFAD in Table 1 and its better generation quality in Table 2.

4.4 Ablation Study

Default setting. All ablations start from the full IAE configuration, which uses Dasheng semantic supervision, a queue of online encodings, and $\alpha \sim \mathcal{U}(0, \frac{1}{2}]$. Several variants replace the online queue with an EMA queue, whose encodings come from an exponential moving average of the encoder (He et al., 2020). Appendix D compares the two queue settings, the semantic supervision model, and the queue capacity, and reports the rFAD of all configurations. Each comparison changes a single component while keeping the remaining settings and the training budget fixed.

Supervision target for interpolation points. This comparison uses fixed α and an EMA queue. We replace the semantic reconstruction loss of the interpolation branch with a Mel-spectrogram re-

construction loss against the source audio. iFAD increases in all five feature spaces, from 0.84 to 1.64 under VGGish and from 17.05 to 40.67 under ATST-Frame. rFAD also degrades. Interpolation points have no waveform ground truth, and an acoustic target there harms both interpolation decoding and clean reconstruction, so interpolation points should carry only semantic constraints.

Sampling of the interpolation coefficient. Replacing the fixed $\alpha = \frac{1}{2}$ with $\alpha \sim \mathcal{U}(0, \frac{1}{2}]$ reduces iFAD in all five feature spaces under both semantic and Mel supervision. Under semantic supervision, iFAD decreases from 1.28 to 0.57 under VGGish and from 19.10 to 13.88 under ATST-Frame. The fixed coefficient supervises exactly the midpoint at which iFAD is evaluated, yet its iFAD at that midpoint remains higher than that of random sampling. The improvement thus comes from continuous coverage of the region between neighboring encodings rather than from fitting a single location.

Fixed-magnitude noise. We add the fixed-magnitude noise reconstruction branch of Anchored AE + Noise to IAE. Under random α, adding noise slightly lowers rFAD in four feature spaces but raises iFAD in all five, from 0.57 to 0.62 under VGGish and from 13.88 to 14.99 under ATST-Frame. The fixed-α setting shows no consistent iFAD improvement from noise either. Consistent with Section 4.3, data-agnostic perturbation improves reconstruction only around clean encodings and does not fill the low-density regions between neighboring encodings.

Table 3: **Ablation of IAE components.** iFAD↓ in five feature spaces. Each block changes one component, and † marks the full IAE configuration. (a) Supervision target of the interpolation branch. (b) Fixed versus random interpolation coefficient. (c) Adding the fixed-magnitude noise reconstruction branch.

Setting	Variant	VGGish↓	AF3↓	Qwen3-Omni↓	Dasheng↓	ATST-Frame↓
(a) Interpolation supervision target						
Fixed α, EMA	Mel	1.6394	30.2505	0.1337	0.1559	40.6659
Fixed α, EMA	Semantic	**0.8419**	**18.4516**	**0.0713**	**0.0785**	**17.0542**
(b) Interpolation coefficient						
Semantic, online	$\alpha = \frac{1}{2}$	1.2765	17.5466	0.0842	0.0759	19.1043
Semantic, online	$\alpha \sim \mathcal{U}(0, \frac{1}{2}]^\dagger$	**0.5689**	**16.3678**	**0.0615**	**0.0745**	**13.8787**
Mel, EMA	$\alpha = \frac{1}{2}$	1.6394	30.2505	0.1337	0.1559	40.6659
Mel, EMA	$\alpha \sim \mathcal{U}(0, \frac{1}{2}]$	**1.4619**	**28.3731**	**0.1120**	**0.1496**	**30.5510**
(c) Fixed-magnitude noise						
Random α	w/ noise	0.6170	17.7053	0.0716	0.0830	14.9921
Random α	w/o noise†	**0.5689**	**16.3678**	**0.0615**	**0.0745**	**13.8787**
Fixed α	w/ noise	1.3535	18.2425	**0.0841**	0.0804	**17.8621**
Fixed α	w/o noise	**1.2765**	**17.5466**	0.0842	**0.0759**	19.1043

5 Conclusion

In this paper, we revisit latent continuity in audio representation learning. For diffusion based audio generation, good reconstruction at encoded samples alone is not enough. The latents produced during sampling often drift away from the training distribution, so the decoder must also remain stable in nearby regions. A latent space that is truly suitable for generation needs not only clear semantic structure but also sufficient continuity. To address this problem, we propose IAE. Instead of relying on fixed magnitude single point perturbations, IAE uses nearest neighbor relations in the current encoding distribution, interpolates between neighboring representations, and extends reliable decoding into the low density regions between samples through semantic reconstruction. In this way, continuity learning is no longer a local smoothing strategy detached from the data distribution, but is tied directly to semantic structure, neighborhood geometry, and the goal of generation. Experiments show that IAE improves decoding quality in interpolation regions and also leads to better downstream generation. This suggests that latent continuity is not a secondary property, but an important link between representation learning and generative quality. We hope this work encourages more explicit attention to this issue and provides a clearer direction for audio representation learning aimed at generation.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods (code reproduction, development, and debugging), to assist with translation, and to help interpret results. We have not used generative AI tools for core research conceptualization or methodology design, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature search and survey, research summarization, as well as manuscript outlining and language polishing.

We have reviewed all AI-assisted work. Specifically, literature summaries were manually cross-referenced against original publications for accuracy. AI-assisted code was thoroughly tested and verified for correctness by the authors. AI-assisted interpretations of results were checked by the authors against the experimental results. In addition, all manuscript outlines and translated and refined text were carefully reviewed to ensure technical precision while preserving the original intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023. URL https://arxiv.org/abs/2301.11325.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Orhun Buğra Baran, Melih Kandemir, and Ramazan Gokberk Cinbis. Policy-based tuning of autoregressive image models with instance- and distribution-level rewards. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 320–339, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37013-6.

Changhao Cheng, Wei Wang, Wangyou Zhang, Dongya Jia, Jian Wu, Zhuo Chen, and Yanmin Qian. On the distillation loss functions of speech vae for unified reconstruction, understanding, and generation, 2026. URL https://arxiv.org/abs/2604.12383.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Interspeech 2024*, pp. 547–551, 2024. doi: 10.21437/Interspeech.2024-246.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, and Jian Luan. Dashengtokenizer: One layer is enough for unified audio understanding and generation, 2026. URL https://arxiv.org/abs/2602.23765.

Carl Doersch. Tutorial on variational autoencoders, 2021. URL https://arxiv.org/abs/1606.05908.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning, 2023. URL https://arxiv.org/abs/2307.16372.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2022. doi: 10.1109/TASLP.2021.3133208.

Mingju Gao, Jingkai Zhou, Kun Gai, Changqian Yu, and Hao Tang. Advfd: Boosting visual generation via adversarial fr'echet distance loss, 2026. URL https://arxiv.org/abs/2608.11205.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3590–3598, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612348. URL https://doi.org/10.1145/3581783.3612348.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025. URL https://arxiv.org/abs/2507.08128.

Jing Gu, Morteza Mardani, Wonjun Lee, Dongmian Zou, and Gilad Lerman. Understanding latent diffusability via fisher geometry, 2026. URL https://arxiv.org/abs/2604.02751.

Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7548–7558, 2024. doi: 10.1109/CVPR52733.2024.00721.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024. doi: 10.1109/SLT61566.2024.10832365.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017. doi: 10.1109/ICASSP.2017.7952132.

Kuan-Po Huang, Bo-Ru Lu, Ho-Lam Chung, Shih-Hsin Wang, and Hung yi Lee. Fdaudio: Meanflow-anchored fréchet-distance post-training for one-step text-to-audio generation, 2026. URL https://arxiv.org/abs/2607.10421.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A Reference-Free Metric for Evaluating Music Enhancement Algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Qifan Li, Xingyu Zhou, Jinhua Zhang, Weiyi You, and Shuhang Gu. Taming sampling perturbations with variance expansion loss for latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43643–43652, June 2026a.

Ruihang Li, Mengde Xu, Shuyang Gu, Leigang Qu, Fuli Feng, Han Hu, and Wenjie Wang. Optimizing visual generative models via distribution-wise rewards, 2026b. URL https://arxiv.org/abs/2607.02291.

Xian Li, Nian Shao, and Xiaofei Li. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1336–1351, 2024. doi: 10.1109/TASLP.2024.3352248.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

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. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xEJMoj1SpX.

Yadong Niu, TIANZI WANG, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, and Jian Luan. MECAT: A multi-experts constructed benchmark for fine-grained audio understanding tasks. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=MgjXTLpmgf.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Junbo Zhang, and Jian Luan. Acavcaps: Enabling large-scale training for fine-grained and diverse audio understanding. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15347–15351, 2026b. doi: 10.1109/ICASSP55912.2026.11463626.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Julian D. Parker, Zach Evans, CJ Carr, Zachary Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. Same: A semantically-aligned music autoencoder, 2026. URL https://arxiv.org/abs/2605.18613.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Interspeech 2021*, pp. 2756–2760, 2021. doi: 10.21437/Interspeech.2021-755.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vY9nzQmQBw.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025a. URL https://arxiv.org/abs/2509.17765.

Tongda Xu, Mingwei He, Shady Abu-Hussein, Jose Miguel Hernandez-Lobato, Chunhang Zheng, Kai Zhao, Chao Zhou, Ya-Qin Zhang, and Yan Wang. Making reconstruction fid predictive of diffusion generation fid, 2026. URL https://arxiv.org/abs/2603.05630.

Xuenan Xu, Jiahao Mei, Zihao Zheng, Ye Tao, Zeyu Xie, Yaoyun Zhang, Haohe Liu, Yuning Wu, Ming Yan, Wen Wu, Chao Zhang, and Mengyue Wu. Uniflow-audio: Unified flow matching for audio generation from omni-modalities, 2025b. URL https://arxiv.org/abs/2509.24391.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Guanrou Yang, Tian Tan, Qian Chen, Zhikang Niu, Yakun Song, Ziyang Ma, Yushen Chen, Zeyu Xie, Tianrui Wang, Yifan Yang, Wenxi Chen, Qi Chen, Wenrui Liu, Shan Yang, and Xie Chen. Wavcube: Unifying speech representation for understanding and generation via semantic-acoustic joint modeling, 2026a. URL https://arxiv.org/abs/2605.06407.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation, 2026b. URL https://arxiv.org/abs/2604.28190.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025. doi: 10.1109/CVPR52734.2025.01464.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, Yike Guo, and Wei Xue. Codec does matter: exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i24.34761. URL https://doi.org/10.1609/aaai.v39i24.34761.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87400–87442, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9e42b4d7163931f3689d6d6fbaa11d0-Paper-Conference.pdf.

Kaiyu Yue, Menglin Jia, Ji Hou, and Tom Goldstein. Image generation with a sphere encoder, 2026a. URL https://arxiv.org/abs/2602.15030.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, and Yali Wang. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion, 2026b. URL https://arxiv.org/abs/2605.07915.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186, 2022. doi: 10.1109/ICASSP43922.2022.9746682.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AF9Q8Vip84.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability, 2026. URL https://arxiv.org/abs/2606.03578.

A Related Work

Semantic structure and continuity. Latent diffusion models benefit from latent spaces that are both semantically structured and continuous (Zhong et al., 2026; Yue et al., 2026b). Semantic structure is typically obtained by aligning latents with pretrained representations, building on pretrained encoders, or introducing supervision from downstream understanding tasks. This paradigm has been widely adopted in vision (Zheng et al., 2026; Yu et al., 2025; Leng et al., 2025) and has also advanced in audio (Zhang et al., 2024; Ye et al., 2025; Parker et al., 2026). Beyond semantic structuring, continuity is equally essential for high-quality generation. At generation time, the decoder must reconstruct not only training encodings but also nearby shifted latents, and whether such latents can be decoded reliably depends on how continuous the latent space is (Ning et al., 2024; Guo et al., 2024). Existing methods typically improve continuity by applying KL regularization or Gaussian perturbations to individual encodings (Doersch, 2021; Zheng et al., 2026; Yue et al., 2026a). While these methods can locally smooth the latent space, they mainly operate on isolated encoding points rather than explicitly constraining the regions between neighboring encodings. Consequently, the decoder receives limited supervision over these intermediate regions during training, when generated latents drift into such low-density regions, it leads to out-of-distribution decoding and degrades generation quality (Yue et al., 2026b; Parker et al., 2026; Li et al., 2026a).

Interpolated FID. Recent studies find that the reconstruction quality of an autoencoder does not predict the generation quality of a diffusion model trained in its latent space (Zhong et al., 2026; Xu et al., 2026). Xu et al. (2026) therefore propose interpolated FID (iFID), which measures the FID of decoded interpolations between nearest-neighbor encodings and correlates strongly with generation FID. Since semantic structure determines which encodings are neighbors and continuity determines how their interpolations decode, iFID reflects both properties; however, it is proposed only as an evaluation metric, and its optimization is left to future work.

Fréchet distance optimization. Other work optimizes Fréchet distances directly, either as differentiable training losses for generators (Yang et al., 2026b; Gao et al., 2026; Huang et al., 2026) or as distribution-level rewards in reinforcement learning (Li et al., 2026b; Baran et al., 2026). These methods, however, optimize the output distribution of a generator rather than the latent space of an autoencoder. IAE instead targets the property measured by iFID during autoencoder training, using semantic reconstruction at nearest-neighbor interpolations as a surrogate that indirectly reduces iFAD, the audio counterpart of iFID.

B Experimental Details

B.1 Tokenizer Architecture and Training

Architecture. The encoder E_θ is the Dasheng base audio transformer (Dinkel et al., 2024), and the decoder D_ϕ is the Vocos-style decoder of DashengTokenizer (Dinkel et al., 2026). The teacher encoder T and the semantic encoder S are frozen copies of the same pretrained Dasheng checkpoint as E_θ. The final-layer features of T serve as the anchoring target in equation 9, and those of S serve as the semantic target in equation 7 and equation 8.

Objective. The weights in equation 11 are $\lambda_{\text{mel}} = 4.5$, $\lambda_{\text{gan}} = 0.1$, $\lambda_{\text{anchor}} = 1.0$, $\lambda_{\text{sem}} = 0.4$, and $\lambda_{\text{interp}} = 0.4$. $\mathcal{L}_{\text{mel}}$ is the ℓ_1 distance between log-Mel spectrograms. $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{fm}}$ use a multi-resolution discriminator (Siuzdak, 2024), which is updated once every three generator steps.

Latent queue and interpolation. The queue stores 32768 latent sequences and is shared across all GPUs by gathering each batch before enqueuing. At each training step, neighbors are retrieved before the current batch is enqueued, so all candidates come from previous steps. Retrieval uses the pooled key of equation 2 and excludes entries with the same sample identifier. When a retrieved neighbor has a different number of frames from the query, it is resampled by linear interpolation along the time axis. α is sampled independently per sample from $\mathcal{U}(0, \frac{1}{2}]$.

Optimization. Table B.1 lists the optimization settings. All tokenizers, including the baselines, share these settings. For each tokenizer, we use the checkpoint with the lowest validation loss for all downstream experiments.

Table B.1: Tokenizer training hyperparameters. Queue size and interpolation coefficient apply to IAE only. VA-VAE uses a 128-dimensional latent.

Setting	Value
Sample rate	16 kHz
Latent dimension / frame rate	768 / 25 Hz
Batch size	32 per GPU $\times$ 8 GPUs
Optimizer	AdamW, $\beta = (0.8, 0.9)$, for both generator and discriminator
Learning rate	1×10^{-4}, 5k warmup steps, cosine decay
Queue size	32768
Interpolation coefficient	$\alpha \sim \mathcal{U}(0, \frac{1}{2}]$

Training data. Domains are sampled with the weights in Table B.2.

Table B.2: Tokenizer training data and domain sampling weights.

Domain	Dataset	Weight
English speech	LibriSpeech (Panayotov et al., 2015)	20
Mandarin speech	WenetSpeech (Zhang et al., 2022)	35
Mandarin speech	AISHELL-3 (Shi et al., 2021)	5
Music	MSD subset with LP-MusicCaps captions (Doh et al., 2023)	18
Music	MagnaTagATune subset with LP-MusicCaps captions (Doh et al., 2023)	2
General audio	WavCaps (Mei et al., 2024)	16
General audio	FSD50K (Fonseca et al., 2022)	4

B.2 Baseline Configurations

All Anchored AE variants use the same encoder, decoder, initialization, data, and optimization settings as IAE (Tables B.1 and B.2).

- **Anchored AE** is trained with $\mathcal{L}_{\text{acoustic}} + \lambda_{\text{anchor}}\mathcal{L}_{\text{anchor}}$ only. It has no perturbation branch and no decoded-output semantic loss.
- **Anchored AE + Noise (σ)** adds a noise branch to Anchored AE. Gaussian noise is rescaled to a per-sample RMS of $\sigma \in \{0.25, 0.35\}$ and added to the clean latent. The perturbed latent is decoded and supervised by the ℓ_1 log-Mel loss against the source audio, with the same weight as $\mathcal{L}_{\text{mel}}$ (4.5).
- **VA-VAE (128-d)** replaces the identity projection with a 128-dimensional variational bottleneck and replaces $\mathcal{L}_{\text{anchor}}$ with a semantic alignment loss. Our implementation follows VA-VAE (Yao et al., 2025) and JMAS-VAE (Cheng et al., 2026).

VA-VAE objective. The encoder output is mapped to posterior parameters $(\mu_t, \sigma_t) \in \mathbb{R}^{128} \times \mathbb{R}^{128}$, and $z_t = \mu_t + \sigma_t \odot \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, I)$. A linear projection maps z_t to $z'_t \in \mathbb{R}^d$, which is aligned with the frozen teacher features $f_t = T(x)_t$ by

$$\mathcal{L}_{\text{mcos}} = \frac{1}{L}\sum_{t=1}^{L} \text{ReLU}(1 - m_1 - \cos(z'_t, f_t)) \,, \tag{B.1}$$

$$\mathcal{L}_{\text{mdss}} = \frac{1}{L^2}\sum_{i=1}^{L}\sum_{j=1}^{L} \text{ReLU}\big(\big|\cos(z'_i, z'_j) - \cos(f_i, f_j)\big| - m_2\big) \,, \tag{B.2}$$

with $m_1 = 0.5$ and $m_2 = 0.25$. The complete objective is

$$\mathcal{L}_{\text{VA-VAE}} = \mathcal{L}_{\text{acoustic}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{align}}\left(w_{\text{mcos}}\mathcal{L}_{\text{mcos}} + w_{\text{mdss}}\mathcal{L}_{\text{mdss}}\right) , \tag{B.3}$$

where $\lambda_{\text{KL}} = 10^{-3}$, $\lambda_{\text{align}} = 2.5$, and w_{mcos} and w_{mdss} are adaptive weights given by the ratio of the gradient norms of $\mathcal{L}_{\text{mel}}$ and the corresponding alignment loss with respect to the posterior projection layer.

B.3 Downstream Generation Model

Following UniFlow-Audio (Xu et al., 2025b), we train a flow-matching DiT (Peebles & Xie, 2023; Lipman et al., 2023) on the frozen latents of each tokenizer. The DiT has 32 layers, hidden size 1536, 24 attention heads, GEGLU feed-forward layers with ratio 4, QK LayerNorm, and adaptive-norm timestep conditioning. Text is encoded by mT5 (Xue et al., 2021) into 1024-dimensional embeddings and injected through cross-attention. Table B.3 lists the training and sampling settings. All generation models share these settings, and only the tokenizer differs.

Table B.3: Generation model training and sampling settings.

Setting	Value
Batch size	24 per GPU $\times$ 8 GPUs
Optimizer	AdamW, lr 5×10^{-5}, weight decay 0.01
LR schedule	2k warmup steps, then constant
Gradient clipping	1.0
Condition dropout (for CFG)	0.2
Sampling steps / CFG scale	25 / 5.0 (guidance rescale 0.5)

Training data. Table B.4 lists the training datasets of the generation model and their sampling weights.

Table B.4: Generation model training data.

Dataset	Content	Weight
ACAVCaps (Niu et al., 2026b)	Speech, music, sound events, and mixtures	59.9%
Emilia (He et al., 2024)	Speech	38.2%
LP-MusicCaps (MSD subset) (Doh et al., 2023)	Music	1.9%

B.4 Evaluation Protocols

Tokenizer evaluation. We sample 500 query clips and a disjoint candidate pool of 5,000 clips from the held-out splits of the tokenizer training domains. For each query, the nearest neighbor is retrieved from the candidate pool, and the midpoint ($\alpha = 0.5$) is decoded. rFAD and iFAD compare the reconstructed and interpolated audio, respectively, against the 500 original query clips, using the same feature extractor.

Generation evaluation. Table B.5 lists the evaluation sets. FAD_{VGG} and FD are computed with VGGish and PANNs CNN14 (Kong et al., 2020) features using the AudioLDM evaluation toolkit (Liu et al., 2023), against the ground-truth audio of the same evaluation set; for MECAT, each category uses its own ground-truth audio. On AudioCaps and MusicCaps, gFAD is additionally computed with AudioFlamingo3, Qwen3-Omni, and Dasheng features, using the same extractors as the tokenizer evaluation. KL is computed with the AudioLDM evaluation toolkit from the PANNs class posteriors of generated and reference audio. WER is computed with Whisper large-v3 (Radford et al., 2023) in English transcription mode, and UTMOSv2 (Baba et al., 2024) is predicted directly from the generated waveform without a reference. For LibriTTS, the model is conditioned only on the normalized target transcript, without a speaker prompt or reference audio.

Table B.5: Generation evaluation sets.

Benchmark	Split	# Prompts
AudioCaps (Kim et al., 2019)	test	957
MusicCaps (Agostinelli et al., 2023)	test	4,950
MECAT (Niu et al., 2026a)	00A / 0M0 / 0MA	848 / 2,593 / 199
	S00 / S0A / SM0 / SMA	7,839 / 2,439 / 5,312 / 643
LibriTTS (Zen et al., 2019)	test-clean / test-other	3,824 / 4,591

C Additional Results

C.1 Tokenizer Results on a Second Evaluation Set

Table C.1 repeats the evaluation of Table 1 on a second evaluation set, with the query clips and the candidate pool resampled using a different seed. IAE again attains the lowest rFAD and iFAD under every extractor, and VA-VAE (128-d) the highest.

Table C.1: **Tokenizer reconstruction and interpolation quality on a second evaluation set.** Same protocol as Table 1, with query clips and candidate pool resampled using a different seed.

	VGGish		AudioFlamingo3		Qwen3-Omni		Dasheng		ATST-Frame	
Method	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓	rFAD↓	iFAD↓
Anchored AE	0.2106	0.7179	7.9073	25.9503	0.0203	0.0784	0.0364	0.1361	3.7142	19.4918
Anchored AE + Noise (0.25)	0.1997	0.6814	7.1720	24.7374	0.0194	0.0774	0.0316	0.1364	3.1561	17.4952
Anchored AE + Noise (0.35)	0.2305	0.7486	7.3938	25.2014	0.0177	0.0774	0.0334	0.1374	3.5036	17.7965
VA-VAE (128-d)	0.5705	1.0652	10.9423	29.7798	0.0439	0.1104	0.0565	0.1674	9.4271	25.8942
IAE	**0.1016**	**0.5873**	**5.0498**	**18.1954**	**0.0105**	**0.0616**	**0.0101**	**0.0808**	**1.5307**	**14.5656**

C.2 Generation Results on MECAT

Table C.2 reports generation results on the seven MECAT categories. IAE attains the lowest gFAD and FD in five categories, and the lowest WER and highest UTMOSv2 in every speech-containing category.

D Additional Ablations

Semantic supervision model. Table D.1 compares Dasheng and Qwen3-Omni as the frozen semantic supervision model under fixed α. Each model performs best in its own feature space. In the three feature spaces used by neither model, Dasheng yields lower iFAD under AudioFlamingo3 and ATST-Frame for both queue settings, while Qwen3-Omni performs better only under VGGish. Dasheng shares its pretrained checkpoint with the trainable encoder, so the semantic reconstruction target and the encoder anchoring target lie in the same representation space.

Queue encoder. Table D.1 also compares the online and EMA queues defined in Section 4.4. Under fixed α, the EMA queue yields lower iFAD in three of the five feature spaces for both supervision models, but this gain is smaller than that of random α sampling (Table 3(b)). IAE therefore uses the online queue, which directly reflects the current encoding distribution and requires no additional copy of the encoder.

Queue capacity. Table D.2 varies the queue capacity under the full IAE configuration. Except under VGGish, iFAD varies little across capacities, and IAE is insensitive to this choice. The smallest capacity, 16384, performs slightly worse in four of the five feature spaces, likely because a smaller candidate set provides neighbors that are less semantically similar.

Reconstruction quality of all configurations. Table D.3 reports the rFAD of every configuration in Tables 3, D.1, and D.2.

Table C.2: **Generation results on MECAT.** Categories are named by the presence of speech (S), music (M), and environmental sound (A); for example, S0A contains speech and environmental sound. WER and UTMOSv2 are reported only for speech-containing categories. Best in bold.

Category	Method	$\text{FAD}_{\text{VGG}}\downarrow$	FD↓	KL↓	WER↓	UTMOSv2↑
00A	Anchored AE	12.91	59.50	2.609	–	–
	Anchored AE + Noise (0.25)	8.16	40.19	2.008	–	–
	Anchored AE + Noise (0.35)	8.77	37.79	1.949	–	–
	VA-VAE (128-d)	13.48	59.64	2.540	–	–
	IAE	**5.67**	**30.02**	**1.737**	–	–
0M0	Anchored AE	2.82	23.20	0.673	–	–
	Anchored AE + Noise (0.25)	2.19	22.69	0.669	–	–
	Anchored AE + Noise (0.35)	**1.67**	**15.66**	**0.646**	–	–
	VA-VAE (128-d)	3.19	23.37	0.776	–	–
	IAE	1.74	21.85	0.659	–	–
0MA	Anchored AE	7.69	56.25	2.030	–	–
	Anchored AE + Noise (0.25)	4.97	46.63	1.885	–	–
	Anchored AE + Noise (0.35)	**4.74**	**35.88**	1.585	–	–
	VA-VAE (128-d)	7.60	52.30	1.734	–	–
	IAE	5.12	41.20	**1.575**	–	–
S00	Anchored AE	4.14	18.08	0.512	32.45	2.067
	Anchored AE + Noise (0.25)	2.58	12.52	**0.432**	32.93	2.211
	Anchored AE + Noise (0.35)	3.09	11.71	0.450	32.73	2.206
	VA-VAE (128-d)	4.34	18.44	0.512	30.63	2.094
	IAE	**2.29**	**10.87**	0.438	**29.29**	**2.392**
S0A	Anchored AE	7.10	32.79	1.150	33.14	2.009
	Anchored AE + Noise (0.25)	4.02	17.86	0.829	33.23	1.923
	Anchored AE + Noise (0.35)	4.41	17.69	0.830	35.32	2.070
	VA-VAE (128-d)	7.30	32.55	1.058	31.34	2.009
	IAE	**2.48**	**13.22**	**0.775**	**28.14**	**2.308**
SM0	Anchored AE	3.68	20.10	0.421	31.66	2.240
	Anchored AE + Noise (0.25)	1.87	13.98	0.400	31.69	2.163
	Anchored AE + Noise (0.35)	1.73	13.26	**0.372**	32.04	2.243
	VA-VAE (128-d)	3.46	20.94	0.457	27.19	2.131
	IAE	**1.34**	**12.74**	0.404	**26.94**	**2.361**
SMA	Anchored AE	4.98	36.59	1.148	33.33	2.268
	Anchored AE + Noise (0.25)	2.34	25.26	0.980	37.42	2.017
	Anchored AE + Noise (0.35)	2.40	23.69	**0.871**	36.17	2.194
	VA-VAE (128-d)	4.97	35.01	1.062	36.17	2.088
	IAE	**1.93**	**22.99**	0.900	**30.06**	**2.335**

Table D.1: **Semantic supervision model and queue encoder.** iFAD↓ in five feature spaces, with fixed $\alpha = \frac{1}{2}$.

Supervision	Queue	VGGish	AF3	Qwen3-Omni	Dasheng	ATST-Frame
Qwen3-Omni	EMA	**0.6925**	20.5873	0.0364	0.1048	25.5587
Qwen3-Omni	Online	0.8057	19.6011	**0.0362**	0.1086	27.6657
Dasheng	EMA	0.8419	18.4516	0.0713	0.0785	**17.0542**
Dasheng	Online	1.2765	**17.5466**	0.0842	**0.0759**	19.1043

Table D.2: **Queue capacity.** iFAD↓ under the full IAE configuration. † marks the default capacity.

Capacity	VGGish	AF3	Qwen3-Omni	Dasheng	ATST-Frame
16384	0.7254	16.7165	0.0646	0.0755	13.8632
65536	0.6349	16.4988	**0.0611**	0.0754	**13.5674**
32768†	**0.5689**	**16.3678**	0.0615	**0.0745**	13.8787

Table D.3: **rFAD of all ablation configurations.** rFAD↓ in five feature spaces. † marks the full IAE configuration.

Supervision	Target	α	Queue	Noise	VGGish	AF3	Qwen3-Omni	Dasheng	ATST-Frame
Dasheng	Mel	fixed	EMA	–	0.2212	6.9158	0.0196	0.0304	3.2748
Dasheng	Mel	random	EMA	–	0.1679	6.8281	0.0165	0.0310	3.1387
Dasheng	Semantic	fixed	EMA	–	0.1216	4.9538	0.0118	0.0106	1.2967
Dasheng	Semantic	fixed	Online	–	0.1206	4.4578	0.0132	0.0110	1.3077
Dasheng	Semantic	fixed	Online	✓	0.1083	4.9882	0.0102	0.0119	1.1847
Dasheng	Semantic	random	Online	✓	0.0931	4.4371	0.0092	0.0092	**0.9683**
Dasheng	Semantic	random	Online	–†	0.0950	4.5273	0.0092	0.0093	1.0580
Qwen3-Omni	Semantic	fixed	EMA	–	0.1083	4.8940	**0.0055**	0.0234	1.8491
Qwen3-Omni	Semantic	fixed	Online	–	0.1208	5.3635	0.0070	0.0258	2.3104
IAE, queue capacity 16384					0.1039	4.3795	0.0101	0.0094	1.0960
IAE, queue capacity 65536					**0.0887**	**4.0407**	0.0096	**0.0091**	1.0772