DASHENGTOKENIZER: ACOUSTICALLY AUGMENTED SEMANTIC REPRESENTATIONS FOR AUDIO UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio understanding and generation place different demands on representations: accurate reconstruction does not ensure useful understanding features, while semantic features may not directly support sound synthesis. DashengTokenizer connects these capabilities by reusing continuous, high-dimensional features from a frozen audio encoder and injecting acoustics through a parallel lightweight Mel projection. The unquantized fusion preserves the original feature dimension. The projection and waveform decoder are jointly trained in one adaptation stage; understanding tasks read the fused features, while conditional generators predict the same type of representation before decoding. Experiments demonstrate competitive reconstruction across speech, music, and environmental sound and advantages over compared encoders and tokenizers on several sound and music understanding tasks. Full-sequence flow matching and continuous autoregression demonstrate use as a generation target. In multi-domain comparisons with the same DiT backbone configuration, the representation and decoder outperform an acoustic VAE scheme on text-to-audio, music, and speech metrics. Acoustic adaptation can thus extend pretrained understanding features into a continuous representation connecting reconstruction, understanding, and conditional generation.

1 INTRODUCTION

Human speech perception and production form a feedback loop: speakers turn linguistic content into sound and monitor their own speech to regulate production (Tourville et al., 2008). The motor theory of speech perception offers a related perspective, identifying intended articulatory gestures as the objects of perception and proposing common articulatory representations to connect perception and production (Liberman & Mattingly, 1985). In audio modeling, understanding identifies linguistic content, sound events, and musical attributes, while generation renders them as concrete sounds. The information extracted by understanding also helps determine what generated audio should convey. This overlap motivates representations connecting sound recognition and description with synthesis.

Shared knowledge does not imply identical information requirements. Understanding features emphasize task-relevant information and robustness to irrelevant variation (Dinkel et al., 2024; 2025). Speech recognition benefits from similar content representations across speakers and recording conditions, whereas reconstruction and controlled generation must express specific timbre, prosody, and environmental acoustics (Défossez et al., 2022; Liu et al., 2023). Useful invariance can therefore conflict with rendering a particular sound. A common representation must reconcile these demands and specify how content and acoustics enter and are used downstream.

Existing approaches semantically supervise reconstruction-based acoustic representations, supplement quantized understanding features with acoustic codes, or map acoustic latents into high-dimensional semantic spaces (Zhang et al., 2024; Défossez et al., 2024; Liu et al., 2024a; Yan et al., 2025). This paper investigates directly reusing pretrained understanding features as the basis of the final representation, supplemented by acoustic adaptation. Visual representation autoencoders (RAEs) likewise demonstrate reconstruction and generation with frozen pretrained encoders and learned decoders (Zheng et al., 2025). The question here is whether audio understanding features can retain

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.

their continuous form and dimensionality while a simple acoustic supplement supports reconstruction and conditional generation alongside understanding.

DashengTokenizer adds a parallel lightweight Mel projection to frozen pretrained audio features at matching temporal resolution and dimension. It requires neither vector quantization nor a separate low-dimensional acoustic space. After semantic pretraining, the projection and waveform decoder are jointly trained in one adaptation stage, with the understanding backbone frozen. Understanding tasks read the fusion, reconstruction decodes it, and conditional generators predict the same type of features before decoding. Existing understanding features thus directly constitute the representation for all three uses, beyond serving as supervision or conditioning.

The main contributions are:

- **A continuous audio tokenizer built on understanding representations.** Frozen understanding features, parallel acoustic supplementation, and jointly learned decoding support reconstruction, understanding, and generation in the original high-dimensional continuous space.
- **Cross-domain evidence for understanding and generation.** Experiments cover speech, music, and environmental sound, with advantages on selected understanding tasks and generation comparisons using the same DiT configuration. Full-sequence flow matching and continuous autoregression demonstrate both generation uses.

Table 1: Representative audio representations and their construction.

Method	Form	Semantic information	Audio decoding
EnCodec	Discrete	No semantic teacher	Waveform decoder
Mimi	Discrete	Teacher distillation	Waveform decoder
Whisper	Continuous	Transcription supervision	–
AudioLDM VAE	Continuous	No semantic teacher	Decoder + vocoder
DashengTokenizer	Continuous	Frozen features + acoustic injection	Waveform decoder

2 Related work

Audio representations must organize content, attributes, and acoustics while supporting sound recovery and conditional generation. The following separates *construction*—how semantic and acoustic information enter the features—from *decoding and generation*—what generators predict and how predictions become sound.

2.1 Constructing audio representations

Two starting points are relevant: enriching reconstruction-based acoustic representations with semantic supervision or mappings, and directly reusing pretrained understanding features with acoustic supplementation. These routes emphasize semantic enrichment of acoustic representations and acoustic adaptation of understanding representations, respectively.

The first route builds on reconstruction-based representations exemplified by EnCodec and DAC (Défossez et al., 2022; Kumar et al., 2023). SpeechTokenizer supervises the first residual vector quantization (RVQ) layer with HuBERT features and uses later layers for residual acoustics. Mimi instead adds a semantically distilled quantization branch to parallel acoustic RVQ for decoding (Zhang et al., 2024; Défossez et al., 2024). Semantic-VAE aligns continuous latents with pretrained speech features; JMAS-VAE combines frame-level and inter-frame relational constraints to examine distillation's effects on reconstruction, understanding, and generation (Niu et al., 2026; Cheng et al., 2026). MingTok-Audio freezes a pretrained acoustic VAE and adapts a Whisper-initialized semantic module through distillation and joint semantic–reconstruction objectives, mapping low-dimensional acoustic latents to high-dimensional unified representations (Yan et al., 2025).

The second route reuses content and attribute information from understanding encoders such as WavLM, Whisper, and Dasheng (Chen et al., 2022; Radford et al., 2023; Dinkel et al., 2024). For

low-bitrate compression, SemantiCodec quantizes frozen AudioMAE features and feeds pre- and post-quantization features into an acoustic encoder. Additional discrete codes compensate for quantization loss, and the streams are concatenated for recovery (Liu et al., 2024a). DashengTokenizer preserves unquantized understanding features, learns acoustic supplements in parallel from the same audio's Mel spectrogram, and adds them in the original dimension. Its parallel fusion resembles Mimi, but understanding features themselves enter the final representation instead of serving only as distillation targets (Défossez et al., 2024). Jointly learning the supplement and decoder extends the frozen continuous features to sound recovery.

2.2 Decoding and generating audio representations

Decoding recovers a waveform from an existing encoding; conditional generation must additionally produce new decodable features from text or other conditions. Both the decoding pathway and the generation target's representation space and prediction process matter.

Sound recovery from pretrained features has been explored in enhancement. Sun et al. (2025) freeze the encoder, train a decoder on clean-speech features, and separately learn feature denoising. Song et al. (2024) fuse frozen self-supervised features with the input spectrum for vocoding. SemantiCodec instead conditions diffusion in an acoustic VAE space on concatenated semantic/acoustic codes, then uses a VAE decoder and vocoder (Liu et al., 2024a). DashengTokenizer jointly learns acoustic supplementation and decoding so that fused features directly enter the waveform decoder.

A common generation route first learns compact acoustic VAE latents, then models their conditional distribution. AudioLDM compresses Mel spectrograms for latent diffusion and waveform recovery through a VAE decoder and vocoder; UniFlow-Audio generates waveform-VAE latents with flow matching (Liu et al., 2023; Xu et al., 2025). The VAE defines the reconstruction space, while diffusion or flow matching produces new samples.

Understanding features can also mediate between conditions and acoustic latents. AudioLDM 2 predicts AudioMAE representations, or language of audio (LOA), to condition acoustic VAE diffusion (Liu et al., 2024b). DashengTokenizer instead makes acoustically adapted fusion features the directly predicted and decodable target, without generating separate VAE latents. Understanding tasks read the same type of feature. This changes the basis of the generation space, beyond its dimension: existing understanding features and acoustic supplements constitute the target.

Continuous generation can jointly model an entire sequence through diffusion or flow matching, or proceed autoregressively (Liu et al., 2023; Xu et al., 2025). MELLE predicts continuous Mel frames without vector quantization (Meng et al., 2025); MiDashengLM-Gen combines autoregression with context-conditioned flow matching for each next feature block (Sun et al., 2026). Both full-sequence and blockwise settings are evaluated here as ways to predict acoustically augmented understanding features and decode them into audio.

3 Method

DashengTokenizer supplements pretrained understanding features with acoustics and learns to decode their fusion. The following describes its branches, adaptation objectives, and reconstruction, understanding, and generation uses (Figure 1c).

3.1 Constructing acoustically augmented semantic representations

Given audio x, a frozen encoder S extracts understanding features s. A parallel acoustic branch A projects the same audio's Mel spectrogram into supplementary features a with matching length and dimension. Addition yields fused features for waveform decoder G:

$$s = S(x), \qquad a = A(x) = \mathrm{LayerNorm}(\phi(\mathrm{Mel}(x))), \tag{1}$$

$$z = s + a, \qquad \hat{x} = G(z). \tag{2}$$

The trainable patch projection ϕ gives $s, a, z \in \mathbb{R}^{T \times D}$, with length T and dimension D. Fusion preserves both dimensions and remains continuous without vector quantization. S and the Mel transform are fixed; A and G are jointly adapted.

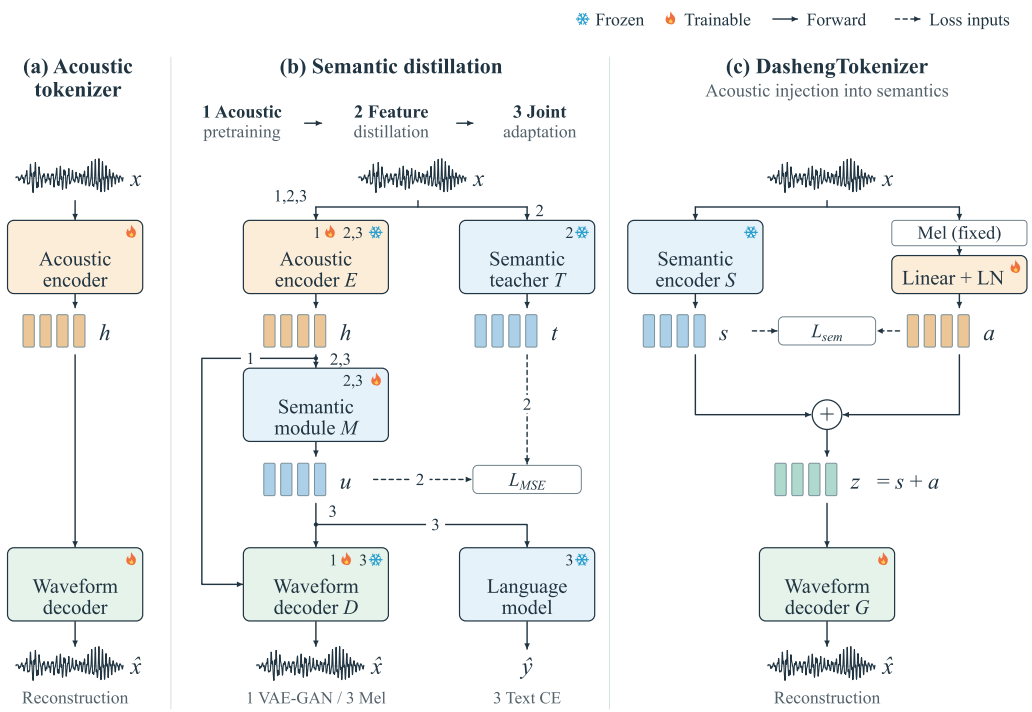


Figure 1: Construction routes: (a) acoustic reconstruction; (b) representative staged semantic distillation (Yan et al., 2025); (c) joint acoustic-projection and decoder adaptation with frozen semantic features. Stage labels identify active paths and module states. Decoder interface projections are omitted.

The main encoder, from MiDashengLM-7B, produces 1280-dimensional features at 25 Hz (Dinkel et al., 2025). The acoustic branch projects non-overlapping 128×4 patches of a 128-bin Mel spectrogram with a 10 ms hop, using a two-dimensional convolution followed by LayerNorm. The decoder upsamples from 25 to 50 Hz and uses a 12-layer, width-1280 Vocos to produce 16 kHz waveforms (Siuzdak, 2023). Table 2 summarizes the modules.

Table 2: DashengTokenizer architecture.

Attribute	Semantic encoder	Acoustic branch	Decoder
Parameters	630M	0.66M	173M
Architecture	32-layer Transformer	2D Conv + LN	Vocos
Hidden dim.	1280	1280	1280
Input	64-bin Mel	128-bin Mel	Fused features
Frame rate	25 Hz	25 Hz	$25 \rightarrow 50$ Hz

3.2 Reconstruction and Branch-Alignment Objectives

Adaptation jointly optimizes A and G. Reconstruction teaches them to recover sound from fusion, while alignment constrains a using frozen features s. Understanding features thus both enter fusion and guide the supplement:

$$\mathcal{L}_G = \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} + \mathcal{L}_{\text{fm}} + \mathcal{L}_{\text{adv}}, \qquad \mathcal{L}_{\text{sem}} = \|s - a\|_2^2. \tag{3}$$

Here $\mathcal{L}_{\text{mel}}$, $\mathcal{L}_{\text{fm}}$, and $\mathcal{L}_{\text{adv}}$ are Mel-domain L1, feature-matching, and adversarial losses. Adversarial training uses multi-frequency discriminators and a hinge objective; the main configuration sets $\lambda_{\text{sem}} = \lambda_{\text{mel}} = 45$. Reconstruction requires information for sound recovery, while alignment constrains the new branch using understanding features. The supplement can consequently contain

both understanding-relevant information and fine acoustics; it is not restricted to a residual missing from semantic features. Section 4.5 examines the resulting trade-offs through objective weights.

3.3 Using the representation for reconstruction, understanding, and generation

Reconstruction decodes real-audio features z through G; understanding tasks read z. Conditional generation learns to predict the same type of features from text using real-audio encodings as targets, then decodes its predictions.

Full-sequence flow matching learns a text-conditioned transformation from noise to the entire feature sequence before decoding. Continuous autoregression instead organizes features into temporal blocks: text and feature history condition flow matching for each next block, and the resulting sequence is decoded (Sun et al., 2026). Section 4.4 specifies both configurations.

4 Experiments

Evaluation proceeds from reconstruction and understanding of real-audio encodings to prediction and decoding of new features across speech, music, and general sound. Objective and decoder-input analyses, speech enhancement, and discussion then clarify the representation's behavior and scope. Figure 2 previews selected results.

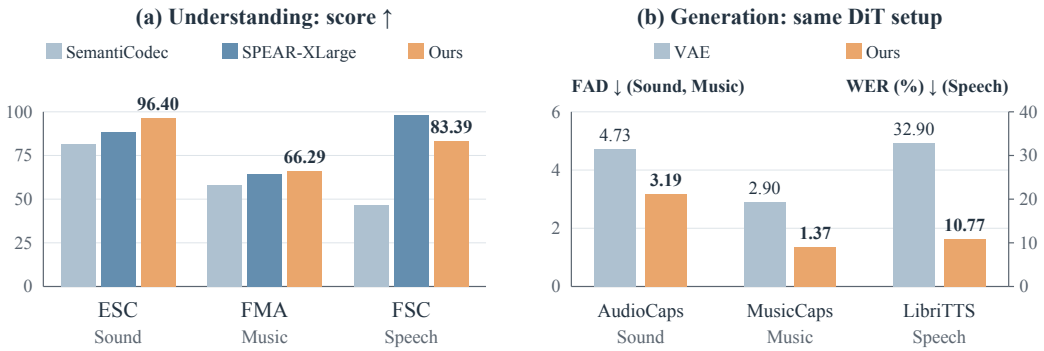


Figure 2: Selected results, ordered sound, music, and speech. (a) Individual understanding tasks with different pretraining. (b) Generation with the same DiT configuration: FAD uses the left axis; speech WER uses the right. Full results: Tables 5 and 8.

4.1 Experimental setup

Training. Adaptation uses approximately 282k hours of speech, music, and general audio resampled to 16 kHz. AdamW runs for 1M updates with batch size 256 and learning rate 5×10^{-4}, reduced to 10% by cosine decay (Loshchilov & Hutter, 2019).

Reconstruction. Original and decoded waveforms are compared at 16 kHz. PESQ and STOI evaluate Chinese/English Seed-TTS speech (Anastassiou et al., 2024; Rix et al., 2001; Taal et al., 2010); Mel and STFT distances evaluate AudioSet general sound and MUSDB18 music (Gemmeke et al., 2017; Rafii et al., 2019).

Understanding. The 22 X-ARES tasks cover classification, detection, retrieval, and speech recognition (Zhang et al., 2025). Evaluation compares fusion with external encoders and tokenizers and with the model's semantic and acoustic branches. Discrete codecs use continuous pre-quantization features.

Conditional generation. Full-sequence flow matching separately models both representations using the same DiT configuration and their respective decoders (Anonymous, 2026). AudioCaps and

MusicCaps use FAD, FD, KL, and CLAP for distributional differences and text relevance; LibriTTS uses WER and UTMOS-v2 for intelligibility and predicted quality. MiDashengLM-Gen provides a continuous autoregressive application across the same domains, reporting UTMOS-v2 for speech (Sun et al., 2026). Section 4.4 and Appendix A give configurations and evaluation details.

4.2 AUDIO RECONSTRUCTION

Reconstruction tests whether understanding-based representations recover concrete sounds, first across Chinese and English speech and then general sound and music (Tables 3–4).

Speech. Chinese/English PESQ reaches 4.163/4.125: English exceeds the listed references, while Chinese approaches MingTok-Audio's 4.210. STOI of 0.988/0.987 exceeds MingTok-Audio's 0.980 but is slightly below UniFlow-Audio. Both metrics exceed DAC and Mimi in both languages, demonstrating competitive recovery with metric-dependent strengths.

Table 3: Speech reconstruction at 16 kHz. Hz denotes feature frame rate; a dash indicates an unreported value.

Model	Hz	SeedTTS-ZH PESQ ↑	SeedTTS-ZH STOI ↑	SeedTTS-EN PESQ ↑	SeedTTS-EN STOI ↑
SNAC	–	1.841	0.862	1.804	0.870
Mimi	12.5	2.050	0.890	2.010	0.890
XCodec 2.0	50	2.190	0.920	2.370	0.930
XY-Tokenizer	12.5	2.270	0.900	2.140	0.900
DAC	50	3.860	0.967	3.763	0.969
EzAudio	50	3.857	0.987	3.668	0.989
UniFlow-Audio	50	4.048	0.990	3.858	0.992
MingTok-Audio	50	4.210	0.980	4.040	0.980
DashengTokenizer	25	4.163	0.988	4.125	0.987

General sound and music. On AudioSet, Mel distance of 0.320 is close to UniFlow-Audio's 0.321. STFT distance of 1.585 exceeds UniFlow-Audio's 1.385 but is below MingTok-Audio, DAC, and SNAC. On MUSDB18, Mel/STFT distances of 0.293/1.458 improve on MingTok-Audio's 0.432/1.616 and on DAC and SNAC, whereas EzAudio and UniFlow-Audio score better. Advantages over MingTok-Audio, DAC, and SNAC thus span general sound and music; comparisons with EzAudio and UniFlow-Audio vary by dataset and metric. These results establish cross-domain recovery before examining understanding and generation.

Table 4: General-sound and music reconstruction at 16 kHz. Lower spectral distances are better.

Model	Hz	AudioSet Mel ↓	AudioSet STFT ↓	MUSDB18 Mel ↓	MUSDB18 STFT ↓
XY-Tokenizer	12.5	0.901	2.361	0.607	1.817
DAC	50	0.664	1.986	0.965	2.352
SNAC	–	1.195	2.690	1.242	2.655
XCodec 2.0	50	1.253	2.925	1.273	2.855
EzAudio	50	0.349	1.434	0.258	1.168
UniFlow-Audio	50	0.321	1.385	0.232	1.129
MingTok-Audio	50	0.462	1.768	0.432	1.616
DashengTokenizer	25	0.320	1.585	0.293	1.458

4.3 AUDIO UNDERSTANDING

External comparisons establish downstream performance; comparisons with the actual semantic branch reveal changes from acoustic adaptation.

External comparisons. In Table 5, ESC and MAESTRO scores of 96.40/57.65 exceed Dasheng 0.6B's 87.95/46.81; FMA and GTZAN also exceed all listed references. FSC exceeds SemantiCodec and UniFlow-Audio but falls below MingTok-Audio, SPEAR-XLarge, and Dasheng 0.6B; NSynth's 77.91 approaches Dasheng 0.6B's 78.39. Together with reconstruction, this establishes recovery and downstream readout, with particular strengths on sound and music tasks.

Table 5: Selected external understanding comparisons; higher is better. Architectures and pretraining differ; discrete codecs use pre-quantization features. Complete results are in Appendix B.1.

Model	FSC	ESC	FMA	GTZAN	MAESTRO	NSynth
SemantiCodec	46.40	81.75	57.94	66.67	8.62	68.14
UniFlow-Audio	1.32	19.85	30.17	40.74	16.44	31.73
MingTok-Audio (Unified)	98.58	62.25	54.45	71.17	25.00	57.00
SPEAR-XLarge	98.20	88.45	64.16	84.59	13.63	67.85
Dasheng 0.6B	97.79	87.95	64.34	89.49	46.81	78.39
DashengTokenizer	83.39	96.40	66.29	89.99	57.65	77.91

Changes across branches. Table 6 compares actual semantic features s, acoustic features a, and fusion z. MAESTRO rises from 38.65 to 57.65, FMA and NSynth improve, and FSC drops from 98.26 to 83.39. Acoustic features alone reach 58.84 on MAESTRO, showing task-relevant information in the new branch.

Table 6: Actual semantic features, the complete model's acoustic branch, and fusion. $\Delta = \mathrm{Unified} - \mathrm{Semantic}$ is the displayed score difference. Metrics are defined in Appendix A.3.

Task	Semantic	Acoustic	Unified	Δ
FSC	98.26	2.71	83.39	-14.87
LS100h ↑	80.32	0.00	68.70	-11.62
VoxL33 ↑	75.40	34.34	74.06	-1.34
CREMA-D ↑	80.10	44.44	80.55	+0.45
DES ↑	52.02	21.49	55.40	+3.38
FMA ↑	65.14	41.71	66.29	+1.15
MAESTRO ↑	38.65	58.84	57.65	+19.00
NSynth ↑	77.59	43.16	77.91	+0.32

Across 22 tasks, fusion improves 5 scores, reduces 16, and ties 1 relative to semantic features (Appendix B.2). External performance and adaptation effects differ: fused ESC exceeds the external references but drops from 96.95 to 96.40 relative to the actual semantic branch. The decodable representation remains competitive on several sound/music tasks while changing the original features' task profile.

4.4 Audio generation

Conditional generation must predict new fused features from text, beyond decoding observed-audio encodings. Full-sequence flow matching compares DashengTokenizer with an acoustic VAE across audio, music, and speech; continuous autoregression then tests blockwise prediction.

4.4.1 Generation settings

Both full-sequence schemes use the same 32-layer, width-1536 DiT configuration, train on an approximately 77k-hour private ACAVCaps extension spanning speech, music, and environmental sound, and use their respective decoders (Anonymous, 2026). Main DashengTokenizer features run at 25 Hz; dimensions are in Table 7. Width selection considers prior analysis of bottlenecks with linear input/output projections (Zheng et al., 2025).

Table 7: Generation configurations.

Generation	Generator	Latent dimensions
Flow matching	DiT (32 layers, width 1536)	DashengTokenizer: 1280; VAE: 128
Autoregressive	5 frames/step	DashengTokenizer: 768

4.4.2 FULL-SEQUENCE FLOW-MATCHING GENERATION

Text-to-audio and music. DashengTokenizer improves all four metrics over the acoustic VAE in the multi-domain comparison (Table 8). AudioCaps FAD drops from 4.73 to 3.19 and CLAP rises from 32.1 to 43.8; MusicCaps FAD drops from 2.90 to 1.37 and CLAP rises from 28.9 to 33.4. FD and KL also decrease on both datasets. Decoding predicted features establishes fusion as a full-sequence target beyond real-audio reconstruction.

Text-to-speech. LibriTTS WER drops from 32.90% to 10.77% and UTMOS-v2 rises from 1.52 to 3.12, extending the improvements to speech content and predicted quality (Anonymous, 2026). Appendices C.1–C.2 provide single-domain comparisons.

4.4.3 CONTINUOUS AUTOREGRESSIVE GENERATION

MiDashengLM-Gen predicts five continuous feature frames per step, conditioned on text and feature history, then decodes them into sound (Sun et al., 2026). This tests temporally unfolding fusion targets rather than joint full-sequence prediction; Table 7 specifies the variant.

Table 8 reports distributional and text-relevance metrics on AudioCaps/MusicCaps and 7.75% WER and 2.78 UTMOS-v2 on LibriTTS. These application results establish cross-domain autoregressive use, complementing the representation comparison under full-sequence generation.

Table 8: Cross-domain generation. CLAP and WER are percentages. Configurations are in Table 7.

Generation	Representation	AudioCaps				MusicCaps				LibriTTS	
		FAD ↓	FD ↓	KL ↓	CLAP ↑	FAD ↓	FD ↓	KL ↓	CLAP ↑	WER ↓	UTMOS-v2 ↑
Flow matching	Acoustic VAE	4.73	41.06	2.56	32.1	2.90	24.34	1.62	28.9	32.90	1.52
Flow matching	DashengTokenizer	3.19	26.06	1.86	43.8	1.37	18.45	1.37	33.4	10.77	3.12
Autoregressive	DashengTokenizer	5.01	26.13	1.91	36.0	2.28	14.58	1.24	33.0	7.75	2.78

4.5 FURTHER ANALYSIS

Further analysis examines objective effects on recovery and readout, decoder responses to different branches, and feature prediction for audio restoration.

Training objectives. An 86M semantic backbone is trained on LibriSpeech for 200K updates with both branches retained. At $\lambda_{\text{mel}} = 45$, raising λ_{sem} from 0 to 10 improves NSynth from 38.67 to 76.56 and ESC from 35.20 to 85.25, but increases all three Mel distances. Raising alignment from 10 to 60 only increases NSynth to 77.17 while AudioSet/MusicCaps Mel distances grow from 0.44/0.52 to 0.61/0.68. From 45 to 60, FSD50 rises from 55.94 to 56.09 while ESC drops from 91.85 to 91.60.

Removing the Mel term at $\lambda_{\text{sem}} = 45$ substantially increases all Mel distances, with feature-matching and adversarial objectives retained. Weight selection must account for both recovery and readout. Figure 3 shows reconstruction and four understanding tasks; Appendix D.1 gives all reconstruction and ten-task understanding results.

Decoder-input diagnostic. The fixed trained decoder G receives s, a, or $s + a$. Acoustic-only input improves all six reconstruction metrics over fusion, with Chinese/English PESQ of 4.178/4.131 close to fusion's 4.163/4.125; semantic-only input reconstructs poorly. The acoustic branch already recovers sound well under this decoder, whereas fusion substantially improves several understanding tasks over that branch (Section 4.3). Fusion serves readout, recovery, and conditional prediction

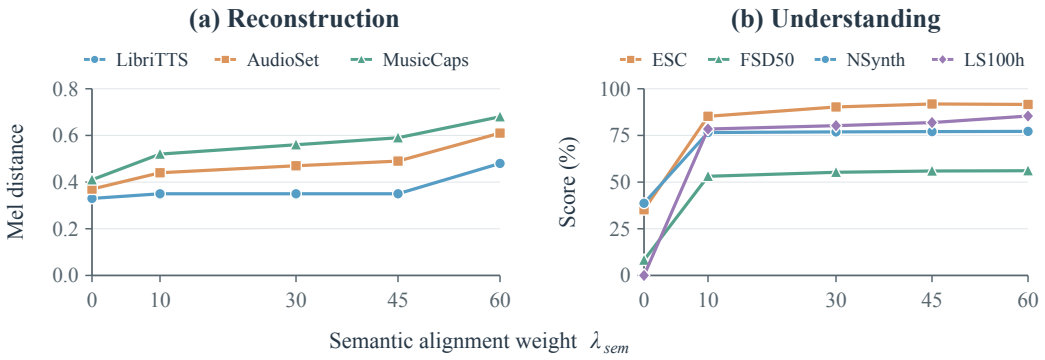


Figure 3: Alignment-weight effects (86M backbone, LibriSpeech, 200K updates; $\lambda_{\text{mel}} = 45$). Understanding scores are accuracy (ESC, NSynth), mAP (FSD50), and $1 - \text{WER}$ (LS100h), all in percent. Full results are in Appendix D.1.

together, with task trade-offs relative to original semantic features. Inputs are not evaluated with separately retrained decoders (Appendix D.2).

Extension: latent-space speech enhancement. A denoiser predicts clean features from noisy audio for waveform decoding, extending text-conditioned prediction to audio-conditioned restoration. PESQ, DNSMOS OVL, and NISQAv2 improve over noisy inputs on Valentini and DNS1 (Appendix E).

4.6 Discussion

Real-audio features support understanding and reconstruction; conditional models predict the same type of features to synthesize sound. Both full-sequence and blockwise results support incorporating pretrained knowledge directly into a decodable generation target.

Freezing preserves the original extractor, but acoustic supplementation changes fusion's task profile: some sound/music tasks benefit while most understanding scores decline relative to semantic features. This reflects the introduction's differing information demands. The alignment-weight analysis connects this tension to the training objective: stronger alignment can improve selected understanding scores while increasing reconstruction error. Evaluating recovery and readout together therefore helps characterize the consequences of adapting a pretrained representation.

Generation comparisons assess the representation-and-decoder scheme jointly; isolating designs requires matched training, and performance depends on task and generator. Reconstruction uses 16 kHz audio and objective metrics. Higher-rate reconstruction and tokenizer-specific listening tests remain needed; listening tests of complete generation systems cannot replace them.

5 Conclusion

DashengTokenizer combines frozen understanding features with jointly learned acoustic supplementation and waveform decoding to support reconstruction, understanding, and conditional generation across speech, music, and environmental sound. The construction preserves the continuous form and dimensionality of the pretrained features, allowing existing audio knowledge to enter the final representation directly. It improves selected sound and music understanding scores over the listed baselines and multi-domain generation over an acoustic VAE with the same DiT configuration. Flow matching and continuous autoregression demonstrate its use with different feature prediction mechanisms. The resulting representation can be decoded for reconstruction, read for understanding, or predicted for synthesis, connecting these uses through a common feature interface. Acoustic adaptation of existing understanding features thus provides a practical route to connecting audio understanding and synthesis.

REPRODUCIBILITY STATEMENT

Section 3 describes the architecture and objectives, and Section 4.1 and Appendix A specify training and evaluation. Appendices B–E provide complete understanding results, supplementary generation comparisons, objective and decoder-input analyses, and speech enhancement. The generation experiments use a private extension of ACAVCaps, which limits independent reproduction of that training setting.

AI USE STATEMENT

Generative AI tools assisted with manuscript review, translation and preparation, interpretation of existing results, and discussion of possible follow-up experiments. They were not used to conduct new experiments. The authors are responsible for the final content.

REFERENCES

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

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Anonymous. Dasheng AudioGen: A unified model for generating coherent audio scenes from text. Unpublished companion manuscript, 2026.

Shikhar Bharadwaj, Samuele Cornell, Kwanghee Choi, Satoru Fukayama, Hye jin Shim, Soham Deshmukh, and Shinji Watanabe. Openbeats: A fully open-source general-purpose audio encoder. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2025.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

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. *arXiv preprint arXiv:2604.12383v2*, 2026. URL `https://arxiv.org/abs/2604.12383v2`. Version 2, May 25.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

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.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. In *Ismir 2023 Hybrid Conference*, 2023.

Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer. *arXiv preprint arXiv:2409.10819*, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *The Thirteenth International Conference on Learning Representations*.

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.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jiaqi Li, Yao Qian, Yuxuan Hu, Leying Zhang, Xiaofei Wang, Heng Lu, Manthan Thakker, Jinyu Li, Sheng Zhao, and Zhizheng Wu. Flexicodec: A dynamic neural audio codec for low frame rates. *arXiv preprint arXiv:2510.00981*, 2025.

Alvin M. Liberman and Ignatius G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985. doi: 10.1016/0010-0277(85)90021-6. URL https://pubmed.ncbi.nlm.nih.gov/4075760/.

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.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*, 2024a.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734v3*, 2024b. URL https://arxiv.org/abs/2308.05734v3. Version 3, May 11; first submitted in 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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.

Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, Helen Meng, and Furu Wei. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551v2*, 2025. URL https://arxiv.org/abs/2407.08551v2.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets. In *Interspeech 2021*, pp. 2127–2131, 2021.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. In *HEAR: Holistic Evaluation of Audio Representations*, pp. 1–24. PMLR, 2022.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *arXiv preprint arXiv:2507.23511*, 2025.

Zhikang Niu, Shujie Hu, Jeongsoo Choi, Yushen Chen, Peining Chen, Pengcheng Zhu, Yunting Yang, Bowen Zhang, Jian Zhao, Chunhui Wang, and Xie Chen. Semantic-VAE: Semantic-alignment latent representation for better speech synthesis. *arXiv preprint arXiv:2509.22167v3*, 2026. URL https://arxiv.org/abs/2509.22167v3. Version 3, June 14; first submitted in 2025.

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.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. Musdb18-hq - an uncompressed version of musdb18, August 2019.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 886–890. IEEE, 2022.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2:749–752, 2001.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023.

Yanjue Song, Doyeon Kim, Hong-Goo Kang, and Nilesh Madhu. Spectrum-aware neural vocoder based on self-supervised learning for speech enhancement. In *European Signal Processing Conference (EUSIPCO)*, pp. 16–20, 2024. URL https://eurasip.org/Proceedings/Eusipco/Eusipco2024/pdfs/0000016.pdf.

Maximilian Strake, Bruno Defraene, Kristoff Fluyt, Wouter Tirry, and Tim Fingscheidt. INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising. In *Interspeech 2020*, pp. 2467–2471, 2020.

Xingwei Sun, Heinrich Dinkel, Yadong Niu, Linzhang Wang, Junbo Zhang, and Jian Luan. Efficient Speech Enhancement via Embeddings from Pre-trained Generative Audioencoders. In *Interspeech 2025*, pp. 4848–4852, 2025.

Xingwei Sun, Heinrich Dinkel, Gang Li, Jiahao Mei, Yadong Niu, Zerui Han, Yuepeng Jiang, Jiahao Zhou, Lichun Fan, and Jian Luan. MiDashengLM-Gen: Unified audio scene generation via LLM-driven autoregressive flow matching. *arXiv preprint arXiv:2608.11804v1*, 2026. URL https://arxiv.org/abs/2608.11804v1. Technical report, version 1, August 12, 2026.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 4214–4217, 2010.

Jason A. Tourville, Kevin J. Reilly, and Frank H. Guenther. Neural mechanisms underlying auditory feedback control of speech. *NeuroImage*, 39(3):1429–1443, 2008. doi: 10.1016/j.neuroimage.2007.09.054. URL https://pubmed.ncbi.nlm.nih.gov/18035557/.

Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and tts models, 2017.

Yusong Wu, Ke Chen, Tianyu Zhang, Yutong Hui, Taylor Berg-Kirkpatrick, and Shlomo Shier. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023*, 2023. Standard citation for CLAP-based metrics.

Xuenan Xu, Jiahao Mei, Zihao Zheng, Ye Tao, Zeyu Xie, Yaoyun Zhang, Haohe Liu, Yuning Wu, Ming Yan, Wen Wu, et al. Uniflow-audio: Unified flow matching for audio generation from omni-modalities. *arXiv preprint arXiv:2509.24391*, 2025.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, et al. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:2511.05516*, 2025.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland. Spear: A unified ssl framework for learning speech and audio representations. *arXiv preprint arXiv:2510.25955*, 2025.

Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, and Jian Luan. X-ares: A comprehensive framework for assessing audio encoder performance. In *Interspeech 2025*, 2025.

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.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A IMPLEMENTATION AND EVALUATION DETAILS

This appendix supplements the method and evaluation settings, following reconstruction, understanding, and conditional generation in turn.

A.1 FEATURE EXTRACTION AND WAVEFORM DECODING

The main configuration encodes 16 kHz audio. A frozen semantic encoder takes 64-bin Mel spectrograms and produces 1280-dimensional features at 25 Hz. The acoustic branch computes a 128-bin Mel spectrogram with a 10 ms hop, projects non-overlapping 128×4 patches, and applies LayerNorm. Both branches therefore match in temporal length and dimensionality. A one-dimensional transposed convolution upsamples fused features from 25 to 50 Hz, followed by a 12-layer, width-1280 Vocos waveform decoder (Siuzdak, 2023). Approximate parameter counts in Table 2 exclude training-time GAN discriminators.

A.2 TOKENIZER ADAPTATION TRAINING

Adaptation uses approximately 282k hours of speech, music, and general audio resampled to 16 kHz. The semantic encoder is frozen, while the acoustic branch and waveform decoder are jointly updated with multi-frequency adversarial discriminators. AdamW uses global batch size 256 for 1M updates. The initial learning rate of 5×10^{-4} decays to 10% with a cosine schedule. Reconstruction and branch alignment use $\lambda_{\text{sem}} = \lambda_{\text{mel}} = 45$, alongside feature-matching and adversarial terms (Section 3.2). The objective-weight scan in Appendix D.1 uses a separate smaller-scale configuration.

A.3 RECONSTRUCTION AND UNDERSTANDING EVALUATION

Reconstruction. Original and decoded audio are compared at 16 kHz. Wideband PESQ and STOI evaluate quality and intelligibility on Chinese and English Seed-TTS audio (Anastassiou et al., 2024; Rix et al., 2001; Taal et al., 2010). Mel and STFT distances evaluate spectral differences on AudioSet and MUSDB18; Mel-16k and STFT-16k distances are computed using the DAC toolkit (Kumar et al., 2023). Higher PESQ/STOI and lower spectral distances are better. Tables 3 and 4 give all reconstruction results.

Understanding. The linear evaluation protocol follows X-ARES for external models and the Semantic, Acoustic, and Unified features (Zhang et al., 2025). Discrete codecs use continuous pre-quantization features. Acoustic and Unified outputs of MingTok-Audio are listed separately. Speech tasks cover emotion recognition (CREMA-D, RAV), intent classification (FSC), speaker counting (LibCnt), speech recognition (LS100h), gender classification (LSMF), keyword recognition (SPV1), vocal imitation (VocI), vocal category classification (VocS), speaker identification (VoxC), language identification (VoxL33), and spoof detection (ASV2015, abbreviated ASV in the source evaluation).

General-sound tasks include bidirectional audio–text retrieval (Clo), domestic sound event detection (DES), multilabel sound event recognition (FSD50, F18K), and environmental sound classification (ESC, Urb8). Music tasks cover genre classification (FMA, GTZAN), note recognition (MAESTRO), and instrument classification (NSynth).

Classification uses accuracy, FSD50K and FSD18-Kaggle use mAP, and DESED and MAESTRO use Segment-F1 over one-second segments. Clotho averages Recall@1 for audio-to-text and text-to-audio retrieval. LS100h uses $\text{iWER} = \max(1 - \text{WER}, 0)$ with WER expressed as a proportion. All understanding scores are displayed on a percentage scale, with higher values better; scores from different tasks are not averaged. X-ARES defines the benchmark and evaluation procedures (Zhang et al., 2025). Appendix B provides external and within-model comparisons separately.

A.4 CONDITIONAL GENERATION CONFIGURATIONS AND EVALUATION

Configurations. The main full-sequence comparison separately models 1280-dimensional Dasheng-Tokenizer features and 128-dimensional acoustic VAE latents with the same 32-layer, width-1536

DiT configuration, trained on an approximately 77k-hour private extension of ACAVCaps (Anonymous, 2026). Each scheme uses its corresponding decoder. The autoregressive application uses a 768-dimensional variant and predicts five continuous frames per step (Sun et al., 2026); Table 7 summarizes these settings.

Single-domain supplementary experiments train a generator per task, with approximately 750M parameters for each representation. The acoustic VAE uses a 24-layer, width-1024 DiT; DashengTokenizer uses 11 layers with the reported width of 1532. Text-to-audio and music train on WavCaps and LP-MusicCaps and evaluate on AudioCaps and MusicCaps, respectively (Mei et al., 2024; Doh et al., 2023; Kim et al., 2019; Agostinelli et al., 2023). Training runs for 200K updates with global batch size 192; inference uses 25 steps and classifier-free guidance of 5.0. The single-domain speech comparison trains and evaluates on LibriTTS, also adapting generator width and depth to achieve approximately equal parameter counts. Appendices C.1 and C.2 give the results.

Metrics. AudioCaps and MusicCaps use FAD, FD, KL, and CLAP. The first three measure distributional differences from reference audio (lower is better); CLAP measures audio–text relevance (higher is better) (Kilgour et al., 2019; Wu et al., 2023). CLAP is displayed on a percentage scale throughout. LibriTTS uses WER for consistency with target text and UTMOS-v2 for predicted speech quality across all generation settings.

B Complete audio understanding results

Following Section 4.3, external comparisons establish the fused representation's performance relative to other methods, and within-model comparisons reveal changes from the actual semantic features. Appendix A.3 defines the tasks and readout protocol.

B.1 External comparisons

Tables B1–B3 report 19 models or output configurations on 22 tasks, including discrete codecs, understanding encoders, and continuous tokenizers. Architectures and pretraining are not fully matched; discrete codecs use pre-quantization features. Baselines include DAC, EnCodec, SemantiCodec, FlexiCodec, WavTokenizer, OpenBEATs, Whisper, HuBERT, MSM-MAE, WavLM, SPEAR, Dasheng, EzAudio, UniFlow-Audio, and MingTok-Audio (Kumar et al., 2023; Défossez et al., 2022; Liu et al., 2024a; Li et al., 2025; Ji et al.; Bharadwaj et al., 2025; Radford et al., 2023; Hsu et al., 2021; Niizumi et al., 2022; Chen et al., 2022; Yang et al., 2025; Dinkel et al., 2024; Hai et al., 2024; Xu et al., 2025; Yan et al., 2025).

Table B1: External understanding comparison on speech tasks. LS100h uses iWER; other columns use accuracy. Higher is better. Metrics are defined in Appendix A.3.

Model	CREMA-D	FSC	LibCnt	LS100h	LSMF	RAV	SPV1	VocI	VocS	VoxC	VoxL33	ASV2015
DAC	43.90	2.55	43.93	0.00	91.58	36.45	15.50	2.23	42.17	13.24	7.96	93.46
Encodec	39.16	3.48	33.78	0.00	88.44	27.71	26.70	2.16	43.06	10.92	7.89	95.14
SemantiCodec	60.30	46.40	66.05	0.00	98.14	55.00	88.02	23.82	84.28	60.99	26.66	93.69
FlexiCodec	58.00	99.00	43.00	0.00	0.00	50.00	0.00	0.00	79.00	0.00	72.00	93.00
WavTokenizer	33.00	2.00	38.00	0.00	0.00	27.00	0.00	0.00	27.00	0.00	9.00	95.00
OpenBeats	65.90	58.50	69.90	0.00	97.30	63.00	94.40	21.40	87.70	0.00	0.00	0.00
Whisper-Large-V3	71.32	97.78	64.40	90.00	94.93	68.47	97.75	29.30	91.48	24.76	97.38	97.94
HuBERT	58.06	98.73	49.72	82.45	85.94	47.71	95.67	18.82	81.71	3.39	45.06	96.24
MSM-MAE	68.64	0.47	71.22	0.00	97.58	66.39	92.77	0.34	87.56	71.64	66.63	93.19
WavLM	45.16	96.07	50.37	67.06	75.95	32.57	89.77	10.71	71.19	4.51	40.89	95.37
SPEAR-Base	72.90	96.90	62.40	90.00	98.60	68.80	97.10	29.10	90.20	65.50	86.10	99.20
SPEAR-XLarge	78.41	98.20	69.00	88.43	98.73	76.59	97.30	31.28	92.62	79.42	90.49	99.81
Dasheng-Base	77.20	94.38	68.80	72.40	98.47	72.50	96.70	25.05	91.00	77.98	81.29	96.29
Dasheng 0.6B	79.39	97.79	69.67	69.64	98.58	78.13	97.22	29.67	92.21	84.00	86.33	99.24
EzAudio	38.80	1.63	37.19	0.00	84.90	25.56	12.47	1.21	35.26	5.99	6.96	91.44
UniFlow-Audio	39.78	1.32	37.15	0.00	74.57	23.26	11.70	1.37	34.73	6.18	6.34	93.77
MingTok-Audio (Acoustic)	41.22	2.82	39.42	0.00	86.91	30.83	20.34	2.45	38.63	6.97	9.88	92.41
MingTok-Audio (Unified)	66.85	98.58	62.92	93.45	97.10	57.08	96.14	20.64	88.59	35.28	77.44	98.70
DashengTokenizer	80.56	83.39	79.98	68.71	98.06	83.06	97.35	23.92	93.15	51.16	93.60	96.45

Table B2: External understanding comparison on general-sound tasks. Column names and metrics follow Appendix A.3; higher is better.

Model	Clo	DES	ESC	FSD50	F18K	Urb8
DAC	0.73	20.20	31.20	7.80	16.44	50.26
Encodec	0.57	3.82	16.80	4.63	6.94	40.34
SemantiCodec	1.89	48.57	81.75	34.84	26.69	84.54
FlexiCodec	0.00	17.00	30.00	0.00	0.00	0.00
WavTokenizer	0.00	16.00	19.00	0.00	0.00	0.00
OpenBeats	4.00	55.20	86.80	38.00	68.90	86.30
Whisper-Large-V3	3.10	22.64	62.45	32.00	49.56	75.74
HuBERT	1.74	3.64	37.35	17.40	26.50	57.40
MSM-MAE	3.86	0.00	88.70	44.15	48.63	84.35
WavLM	0.79	16.76	31.45	10.06	19.63	53.13
SPEAR-Base	3.65	48.63	82.25	37.96	48.00	81.46
SPEAR-XLarge	3.37	53.42	88.45	44.55	64.82	83.78
Dasheng-Base	3.30	53.20	86.90	40.90	55.70	83.50
Dasheng 0.6B	4.00	55.01	87.95	44.79	61.63	84.71
EzAudio	0.40	17.01	18.05	4.86	13.38	34.97
UniFlow-Audio	0.57	16.90	19.85	4.83	12.19	35.16
MingTok-Audio (Acoustic)	0.63	15.98	23.15	5.82	13.88	44.85
MingTok-Audio (Unified)	2.00	39.80	62.25	24.83	40.81	72.87
DashengTokenizer	5.64	55.40	96.40	59.98	86.81	83.80

Table B3: External understanding comparison on music tasks. MAESTRO uses Segment-F1; the other columns use accuracy. Higher is better.

Model	FMA	GTZAN	MAESTRO	NSynth
DAC	35.43	54.06	12.78	38.57
Encodec	28.80	35.14	23.37	32.59
SemantiCodec	57.94	66.67	8.62	68.14
FlexiCodec	0.00	0.00	0.00	0.00
WavTokenizer	0.00	0.00	0.00	0.00
OpenBeats	62.40	85.90	0.00	58.90
Whisper-Large-V3	58.90	71.77	0.00	63.50
HuBERT	42.81	51.85	0.00	48.22
MSM-MAE	62.74	84.39	0.32	73.02
WavLM	36.11	48.14	0.00	40.14
SPEAR-Base	62.06	81.08	12.65	63.55
SPEAR-XLarge	64.16	84.59	13.63	67.85
Dasheng-Base	64.00	86.90	43.25	69.29
Dasheng 0.6B	64.34	89.49	46.81	78.39
EzAudio	30.74	40.54	17.57	32.18
UniFlow-Audio	30.17	40.74	16.44	31.73
MingTok-Audio (Acoustic)	34.74	39.94	21.38	27.81
MingTok-Audio (Unified)	54.45	71.17	25.00	57.00
DashengTokenizer	66.29	89.99	57.65	77.91

The complete results expose differences beyond the selected main-text tasks. DashengTokenizer scores highly on emotion recognition and speaker counting but falls below some understanding encoders and unified tokenizers on FSC. ESC and MAESTRO are strong, whereas ASV2015, Urb8, and NSynth have higher-scoring external references. Downstream suitability therefore depends on the specific task.

B.2 Branch changes after acoustic adaptation

Table B4 compares the actual semantic features with the acoustic branch of the complete model and their fusion. Semantic, Acoustic, and Unified denote these three outputs; $\Delta = \text{Unified} - \text{Semantic}$ measures within-task changes in displayed scores.

Table B4: Complete 22-task branch comparison. ASV2015 is grouped with speech tasks. Scores follow Appendix A.3.

Domain	Task	Semantic	Acoustic	Unified	Δ
Speech	CREMA-D	80.10	44.44	80.55	+0.45
Speech	FSC	98.26	2.71	83.39	-14.87
Speech	LibCnt	80.79	52.34	79.98	-0.81
Speech	LS100h	80.32	0.00	68.70	-11.62
Speech	LSMF	98.47	88.14	98.06	-0.41
Speech	RAV	87.15	33.54	83.06	-4.09
Speech	SPV1	97.61	19.62	97.35	-0.26
Speech	VocI	30.32	2.86	23.92	-6.40
Speech	VocS	93.15	47.73	93.15	0.00
Speech	VoxC	55.51	19.62	51.16	-4.35
Speech	VoxL33	75.40	34.34	74.06	-1.34
Speech	ASV2015	98.37	89.73	96.45	-1.92
Sound	Clo	6.12	0.37	5.63	-0.49
Sound	DES	52.02	21.49	55.40	+3.38
Sound	ESC	96.95	34.05	96.40	-0.55
Sound	FSD50	61.46	8.93	59.98	-1.48
Sound	F18K	89.19	27.06	86.81	-2.38
Sound	Urb8	86.23	50.02	83.80	-2.43
Music	FMA	65.14	41.71	66.29	+1.15
Music	GTZAN	91.19	58.85	89.99	-1.20
Music	MAESTRO	38.65	58.84	57.65	+19.00
Music	NSynth	77.59	43.16	77.91	+0.32

Fusion improves 5 scores, reduces 16, and ties 1 relative to the semantic features. Improvements occur on CREMA-D, DES, FMA, MAESTRO, and NSynth; VocS ties, while tasks including FSC decline. The acoustic branch exceeds fusion on MAESTRO, further illustrating task-dependent suitability. These comparisons expose adaptation trade-offs; the scores are point estimates without reported variation across random seeds.

C Supplementary generation experiments

Section 4.4 evaluates multi-domain full-sequence flow matching and autoregression. This appendix adds separately trained single-domain generators and comparisons between mixed audio and single-type content. Appendix A.4 gives the configurations.

C.1 Single-domain audio and music generation

Generators are trained separately for text-to-audio and text-to-music, comparing the acoustic VAE and DashengTokenizer on AudioCaps and MusicCaps. The two representations use different DiT widths and depths, each with approximately 750M parameters.

Table C1: Single-domain audio and music generation. CLAP uses the same percentage scale as Table 8.

Task	Representation	FAD ↓	FD ↓	KL ↓	CLAP ↑
TTA	Acoustic VAE	4.27	24.72	1.84	41.2
TTA	DashengTokenizer	3.06	21.23	1.56	43.8
TTM	Acoustic VAE	4.35	23.68	1.94	25.8
TTM	DashengTokenizer	3.47	22.74	2.00	26.1

All four text-to-audio metrics improve, matching the directions in the main multi-domain comparison. For music, FAD, FD, and CLAP improve, while KL rises from 1.94 to 2.00. Some generation benefits thus persist under single-domain training, with different responses across metrics.

C.2 Single-domain speech generation

The speech comparison trains and evaluates on LibriTTS, using approximately equal parameter counts with different generator widths and depths to complement the main multi-domain comparison.

Table C2: Single-domain LibriTTS generation. WER is a percentage; UTMOS-v2 predicts speech quality.

Train → test	Representation	WER (%) ↓	UTMOS-v2 ↑
LibriTTS → LibriTTS	Acoustic VAE	6.40	1.48
LibriTTS → LibriTTS	DashengTokenizer	12.20	3.02

DashengTokenizer raises UTMOS-v2 from 1.48 to 3.02 but increases WER from 6.40% to 12.20%, unlike the joint improvements in the main multi-domain setting. Content consistency and predicted quality must therefore be examined separately. As generator configurations differ, this contrast cannot be attributed solely to the training domain (Anonymous, 2026).

C.3 Mixed audio and single-type references

Mixed-audio evaluation tests co-occurring content within a single clip. MECAT-SMA contains speech, music, and sound effects, while MECAT-0M0 provides a music-only reference within the same evaluation framework. Results come from the multi-domain DiT comparison trained on the private ACAVCaps extension (Anonymous, 2026; Niu et al., 2025).

Table C3: Mixed-audio and single-type generation. ACAVCaps denotes the approximately 77k-hour private extension; WER is a percentage.

Train → test	Metric	Acoustic VAE	DashengTokenizer
ACAVCaps → MECAT-SMA	FAD ↓	2.81	2.17
ACAVCaps → MECAT-SMA	WER (%) ↓	38.17	28.98
ACAVCaps → MECAT-0M0	FAD ↓	2.48	1.68
ACAVCaps → MECAT-0M0	FD ↓	8.32	15.44

On SMA, lower FAD and WER indicate improved overall audio distribution and speech content consistency. On music-only 0M0, FAD decreases while FD increases. Content composition and metric choice can therefore produce different directions of change. The single-domain MusicCaps results appear in Table C1.

D TRAINING OBJECTIVES AND DECODER-INPUT ANALYSIS

The analyses follow Section 4.5: varying the training objectives probes effects on reconstruction and understanding, whereas fixing the trained decoder probes its response to different feature inputs.

D.1 OBJECTIVE EFFECTS ON RECONSTRUCTION AND UNDERSTANDING

The weight scan uses an 86M semantic backbone trained only on LibriSpeech for 200K updates. It fixes $\lambda_{\text{mel}} = 45$ and sets λ_{sem} to 0, 10, 30, 45, and 60, with an additional $\lambda_{\text{sem}} = 45$, $\lambda_{\text{mel}} = 0$ setting. Every configuration retains semantic and acoustic pathways. A zero weight removes only the corresponding loss; feature-matching and adversarial objectives remain.

Table D1: Complete reconstruction results of the alignment-weight scan. Columns report Mel distance at 16 kHz; lower is better.

λ_{sem}	λ_{mel}	LibriTTS-clean Mel ↓	AudioSet Mel ↓	MusicCaps Mel ↓
0	45	0.33	0.37	0.41
10	45	0.35	0.44	0.52
30	45	0.35	0.47	0.56
45	45	0.35	0.49	0.59
60	45	0.48	0.61	0.68
45	0	1.47	4.15	4.96

Table D2: Complete understanding results for the configurations in Table D1. LS100h uses iWER; Appendix A.3 defines the metrics. Higher is better.

λ_{sem}	λ_{mel}	ESC	FSD50	GTZAN	LS100h	NSynth	SPV1	Urb8	VocS	VoxC	VoxL33
0	45	35.20	8.20	54.16	0.00	38.67	24.77	51.97	50.29	21.62	10.07
10	45	85.25	53.11	81.78	78.43	76.56	97.23	79.11	92.99	30.94	90.30
30	45	90.25	55.26	85.89	80.23	76.90	97.29	81.18	93.13	33.67	90.55
45	45	91.85	55.94	86.89	81.89	77.05	97.29	81.83	93.15	37.20	91.11
60	45	91.60	56.09	86.88	85.39	77.17	97.40	82.12	92.79	35.40	91.30
45	0	95.35	60.43	89.09	83.67	77.86	97.57	85.49	93.71	32.61	91.86

Alignment affects readout and reconstruction differently. Raising λ_{sem} from 0 to 10 increases NSynth from 38.67 to 76.56 while increasing all three Mel distances. Further alignment does not benefit all understanding tasks: from 45 to 60, FSD50 increases slightly while ESC decreases. Removing the Mel term at $\lambda_{\text{sem}} = 45$ substantially increases all reconstruction distances while improving several understanding scores. These complete results complement Figure 3 by exposing trade-offs across objectives and uses.

D.2 INPUT DIAGNOSTIC WITH A FIXED DECODER

The trained decoder G is held fixed and receives acoustic features a, semantic features s, or fusion $s + a$. No separate decoder is trained for these inputs; a comes from the complete model with its frozen semantic pathway and alignment objective.

Table D3: Input replacement with a fixed decoder. Mel and STFT are spectral distances, and PESQ evaluates speech reconstruction. Seed-TTS Chinese and English audio is used here for reconstruction.

Decoder input	AudioSet Mel ↓	AudioSet STFT ↓	MUSDB18 Mel ↓	MUSDB18 STFT ↓	SeedTTS-ZH PESQ ↑	SeedTTS-EN PESQ ↑
Acoustic a	0.268	1.497	0.267	1.410	4.178	4.131
Semantic s	3.310	4.978	3.290	5.540	1.040	1.040
Unified $s + a$	0.320	1.585	0.293	1.458	4.163	4.125

Acoustic-only input improves all six listed metrics over fusion, with similar Chinese/English PESQ, while semantic-only input reconstructs poorly. This describes the current trained decoder's response to input changes. Together with Appendix B, the diagnostic reveals different relative performance for sound recovery and downstream readout. It does not train separate semantic or acoustic autoencoders and therefore does not determine whether independently adapted pretrained semantic features could be decoded effectively.

E LATENT-SPACE SPEECH ENHANCEMENT

Conditional generation synthesizes sound by predicting and decoding fused features. Latent-space speech enhancement follows the same usage pattern, replacing text with noisy audio and predicting corresponding clean features. This appendix expands the application in Section 4.5.

E.1 FEATURE DENOISING AND EVALUATION

Noisy speech is formed by mixing clean speech with additive noise. The tokenizer encodes both noisy and clean audio, and a three-layer Transformer denoiser learns the noisy-to-clean feature mapping with MSE. The waveform decoder then synthesizes enhanced speech, following the latent-space denoising application of pretrained audio features (Sun et al., 2025).

Evaluation uses Valentini and DNS1, with noisy inputs and enhancement systems as references (Valentini-Botinhao, 2017; Strake et al., 2020). PESQ and STOI measure speech quality and intelligibility, DNSMOS and NISQAv2 predict quality, and SpkSim measures speaker similarity (Rix et al., 2001; Taal et al., 2010; Reddy et al., 2022; Mittag et al., 2021). DNSMOS retains P808, SIG, BAK, and OVL; the latter three correspond to speech signal, background, and overall quality. Higher values are better for all reported metrics.

E.2 COMPLETE ENHANCEMENT RESULTS

Table E1: Speech enhancement on Valentini and DNS1. LMS is the log-Mel spectrogram reference; Noisy is the unprocessed input. External enhancement results follow Sun et al. (2025). P808, SIG, BAK, and OVL are DNSMOS columns.

Dataset	System	PESQ	STOI	P808	SIG	BAK	OVL	NISQAv2	SpkSim
Valentini	Noisy	1.97	0.92	3.09	3.32	3.11	2.68	3.04	0.888
Valentini	LMS	1.77	0.85	3.45	3.12	4.07	2.90	4.25	0.718
Valentini	Whisper	1.31	0.80	3.39	3.47	4.12	3.22	3.83	0.406
Valentini	WavLM	1.27	0.81	3.49	3.49	4.13	3.26	4.12	0.408
Valentini	Dasheng Denoiser	2.32	0.90	3.52	3.45	4.14	3.22	4.51	0.783
Valentini	DashengTokenizer	2.62	0.92	3.50	3.33	4.08	3.09	4.58	0.814
DNS1	Noisy	1.58	0.92	3.16	3.39	2.62	2.48	2.58	0.924
DNS1	LMS	1.90	0.90	3.87	3.34	4.07	3.10	3.85	0.847
DNS1	Whisper	1.20	0.81	3.87	3.60	4.14	3.36	4.09	0.489
DNS1	WavLM	1.20	0.83	3.91	3.61	4.18	3.40	4.41	0.486
DNS1	Dasheng Denoiser	2.24	0.92	4.03	3.58	4.17	3.37	4.41	0.881
DNS1	DashengTokenizer	2.66	0.95	4.02	3.52	4.16	3.30	4.46	0.902

DashengTokenizer improves PESQ, every DNSMOS score, and NISQAv2 over noisy inputs on both datasets. STOI ties on Valentini and improves on DNS1. Relative to the listed enhancement systems, PESQ, STOI, NISQAv2, and SpkSim are strong, but individual DNSMOS components have better-scoring references. SpkSim remains below the noisy input on both datasets, separating quality improvement from speaker-feature preservation. The results demonstrate fused-feature prediction and waveform decoding for audio restoration.