The WER Trap: Shattering the Illusion of Unified Tokens in Speech Language Models

Anonymous ACL submission

Abstract

The pursuit of a "unified" discrete token for both speech understanding and generation has led the Speech Language Model (SLM) community to heavily rely on Word Error Rate (WER)—the core metric for Whisper-style tokenizers—as the definitive proxy for representation quality. This fosters the assumption that low-WER tokens inherently preserve the information necessary for intelligible acoustic synthesis. We argue this is fundamentally deceptive. While high-frequency tokens succeed in generation tasks due to implicit information leakage, isolating pure semantic information at ultra-low frame rates strips away the fine-grained articulation and micro-dynamics essential for ODE-based generation. Empirically validating this requires extreme compression without sacrificing WER—a methodological bottleneck, as standard fixed-stride downsampling arbitrarily truncates phonetic boundaries. To overcome this, we develop a dynamic compression tokenizer that intelligently aligns representations with semantic boundaries, achieving ultra-low frame rates with exceptionally low WER. Using these isolated "pure" semantic tokens, we expose the WER trap: when conditioning generative models—even with oracle duration alignments—the reconstructed speech suffers from severe articulation blur and is rendered acoustically unintelligible. Our findings demonstrate that semantic categorization rewarded by low WER is inherently orthogonal to the continuous phonetic trajectories required for synthesis, shattering the illusion of the unified token and advocating for explicitly decoupled speech representations.

1 Introduction

The integration of speech modalities into large language models (LLMs) has fundamentally bifurcated the landscape of audio processing (Team et al., 2023; Yang et al., 2025a). On the comprehension side, state-of-the-art models often leverage continuous representations from pre-trained encoders (e.g., Whisper (Radford et al., 2023)) to maintain rich semantic density (Chu et al., 2024; Xu et al., 2025; Wu et al., 2025; Tian et al., 2025). Conversely, the generative frontier has increasingly pivoted towards discrete speech tokens to enable efficient autoregressive modeling within standard NLP architectures (Du et al., 2024, 2025). This architectural divergence has fueled a persistent quest for a "unified" discrete token: a singular, highly-compressed representation capable of bridging this chasm by facilitating both robust language understanding and high-fidelity acoustic synthesis (Défossez et al., 2024; Zhang et al., 2025a). To maximize LLM efficiency—extending context windows and accelerating inference—there is an aggressive

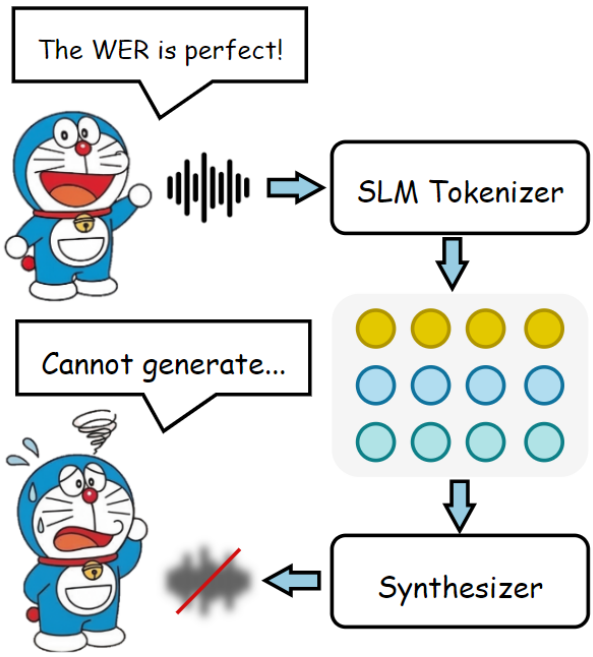

Figure 1: A conceptual illustration of the **WER Trap** It highlights a fundamental paradox in Speech Language Models: discrete tokens that achieve perfect Word Error Rate (WER) in semantic comprehension (top) inherently discard the fine-grained acoustic details required for generation, resulting in synthesis failure (bottom)

push to drive these unified tokens toward low frame rates (Zhang et al., 2025a).

Within this paradigm, the community has come to rely on Word Error Rate (WER)—the primary optimization objective for semantic-driven encoders—as the definitive proxy for token quality (Xu et al., 2025; Wu et al., 2025). This reliance fosters a pervasive assumption in the SLM community: that tokens optimized for minimal WER inherently preserve the essential information for intelligible acoustic generation (Xu et al., 2025; Wu et al., 2025). We argue this is fundamentally deceptive. While high-frequency discrete tokens may appear to support generation, their success is often a byproduct of implicit information leakage afforded by dense sampling. We hypothesize that as tokens are pushed toward the semantic limits required by modern SLMs, the process of isolating pure linguistic states at ultra-low frame rates inherently strips away the fine-grained articulation and micro-dynamics essential for continuous generative models.

Empirically validating this hypothesis—that semantic purity at ultra-low frame rates is fundamentally incompatible with acoustic generation—presents a significant methodological bottleneck. To observe this generative collapse, one must achieve extreme token compression without sacrificing the WER of the tokenizer itself. However, standard fixed-stride downsampling mechanisms widely used in current codecs arbitrarily truncate phonetic boundaries (Huang et al., 2025; Wu et al., 2025; Du et al., 2025). As the compression rate increases (i.e., lowering the frame rate), this rigid downsampling aggressively corrupts the topological sequence of speech, artificially degrading the WER (Zhang et al., 2025b; Yang et al., 2025b). Consequently, the community has been precluded from conducting a fair assessment; the degradation in generation quality at ultra-low frequencies is often mistakenly attributed to alignment errors or boundary corruption, rather than the fundamental mismatch between discriminative and generative representations.

To bypass this bottleneck and truly isolate "pure" semantic tokens at ultra-low frame rates, we develop a novel dynamic compression tokenizer. Rather than relying on rigid, fixed-stride time windows, our method intelligently aligns continuous representations with their underlying acoustic-semantic boundaries through a macroscopic fixed-ratio, microscopic dynamic alignment mechanism. This methodological shift allows us to achieve ultra-low frame rates while preserving exceptionally low WER, effectively providing the controlled experimental condition that has been absent from prior investigations.

With this capability, we are positioned to directly confront the central question motivating this work: *does a discrete speech token that is demonstrably sufficient for language understanding—as certified by low WER and strong reasoning performance—also preserve the information necessary for high-fidelity continuous acoustic generation?* To answer this, we propose a Dual-Probing Protocol that evaluates the *same* ultra-low-rate token sequence through two independent pathways: a discriminative understanding probe built on frozen LLMs, and a generative probe based on continuous Flow Matching conditioned with oracle temporal alignment. By explicitly controlling for both semantic integrity and temporal confounds, this protocol isolates the representational adequacy of the tokens themselves as the sole variable under test. Through this framework, we expose what we term the *WER Trap*—demonstrating that the gap between discriminative sufficiency and generative viability is not an engineering artifact to be resolved with better alignment, but a structural consequence of extreme semantic compression—and argue that the SLM community should move toward explicitly decoupled speech representations.

2 Speech Representation in LLMs: Continuous vs. Discrete

Recent advancements in SLMs generally bifurcate into two distinct architectural paradigms regarding how audio is integrated into the text-centric space of LLMs.

The first paradigm, exemplified by models such as Qwen-Audio, Step-Audio, and SALMONN (Chu et al., 2024; Tian et al., 2025; Tang et al., 2023), relies on **continuous acoustic representations**. These architectures typically employ a robust, pre-trained acoustic encoder (e.g., Whisper (Radford et al., 2023) or Wav2Vec2 (Baevski et al., 2020)) to map raw audio waveforms into a high-dimensional, continuous latent space $\mathcal{H} \in \mathbb{R}^{T \times d}$. These dense embeddings are then projected directly into the LLM's input space. While this continuous integration preserves extremely rich semantic and paralinguistic information—yielding exceptional performance on speech understanding

tasks such as ASR and Audio Question Answering (AQA)—it inherently clashes with the discrete, autoregressive generative nature of LLMs. Consequently, native end-to-end speech generation within this continuous framework is highly non-trivial, often requiring external, cascaded TTS modules or complex continuous-space diffusion adapters (Ho et al., 2020; Lipman et al., 2022).

To circumvent this generation bottleneck, a second paradigm has rapidly emerged, driving the development of fully **discrete SLMs** (e.g., Mini-Audio (Zhang et al., 2025a), SpeechGPT (Zhang et al., 2023), and AudioLM (Borsos et al., 2023)). This approach employs a neural tokenizer to quantize the continuous speech signal into a sequence of discrete tokens $Z = \{z_1, z_2, \ldots, z_N\}$ from a finite vocabulary $\mathcal{V}$, effectively casting audio as a "foreign language". This discrete formulation elegantly aligns with the standard next-token prediction objective, enabling native, autoregressive speech-to-speech and text-to-speech generation.

However, within this discrete paradigm, a further dichotomy exists regarding the nature of the tokens: acoustic versus semantic. Acoustic neural codecs (e.g., EnCodec (Défossez et al., 2022) and SoundStream (Zeghidour et al., 2021)) excel at high-fidelity waveform reconstruction but suffer from high frame rates and multi-layer codebooks (e.g., Residual Vector Quantization (Lee et al., 2022)), which severely burden LLM context windows and exhibit low semantic density, hindering complex language comprehension (Défossez et al., 2024; Zhang et al., 2023).

To balance comprehension capabilities with the autoregressive efficiency of 1D sequences, the community increasingly favors **semantic-driven discrete tokens** (Du et al., 2024, 2025). Typically derived by applying quantization bottlenecks (e.g., FSQ (Mentzer et al., 2023)) to robust encoders like Whisper (Radford et al., 2023), these tokens are rigidly optimized for linguistic content and universally evaluated using the WER. Crucially, because these tokens empirically succeed in driving continuous generative models (e.g., Flow Matching) to reconstruct high-fidelity speech at *standard* frame rates (Du et al., 2024, 2025), the practice of using WER as the sole proxy for overall token quality has become deeply entrenched. Recent unified SLMs routinely adopt a single set of semantic tokens—optimized and validated primarily through WER—for both language comprehension and speech synthesis (Défossez et al., 2024; Zhang et al., 2025a; Wu et al., 2025), without explicitly verifying whether the information preserved by low-WER tokens is also sufficient for continuous acoustic generation. This implicit conflation of discriminative adequacy with generative viability underpins the **"unified" discrete token** paradigm: the assumption that extreme compression, so long as it maintains low WER, will not compromise synthesis quality.

3 The Methodological Bottleneck: Fixed-Stride Compression

To empirically validate or refute the "unified token" assumption, one must push semantic tokens to extreme compression rates (ultra-low frequencies) to isolate the purely linguistic state. However, conducting a rigorous stress test reveals a severe methodological bottleneck in current SLM architectures: the reliance on rigid, **fixed-stride downsampling**.

Standard compression mechanisms applied to semantic encoders (e.g., strided convolutions or fixed-window pooling) operate uniformly across the time axis (Radford et al., 2023; Baevski et al., 2020). This rigid approach fundamentally conflicts with the physical reality of human speech, which is inherently **asynchronous** (Ladefoged, 2001). Phonetic events and acoustic micro-dynamics vary drastically in duration; for instance, a transient plosive (like /p/ or /t/) spans only a few milliseconds, whereas a stressed vowel may extend over hundreds of milliseconds (Ladefoged, 2001; Stevens, 2000).

When fixed-stride compression is aggressively scaled to achieve ultra-low frame rates, it acts as a blind temporal filter. It arbitrarily truncates and fractures these underlying phonetic boundaries, forcefully merging distinct acoustic events or splitting single phonemes across multiple discrete tokens (Lu et al., 2016; Dong and Xu, 2020). Consequently, this temporal misalignment artificially corrupts the semantic integrity of the sequence, causing the downstream WER to explode before the generative limits of the tokens can be fairly evaluated.

This architectural flaw creates a confounding variable: when TTS generation fails at low frame rates, it is often misattributed to the corrupted alignment (the exploded WER) rather than the intrinsic generative deficiency of the purely semantic tokens (Gong et al., 2025; Cheng et al., 2025). There-

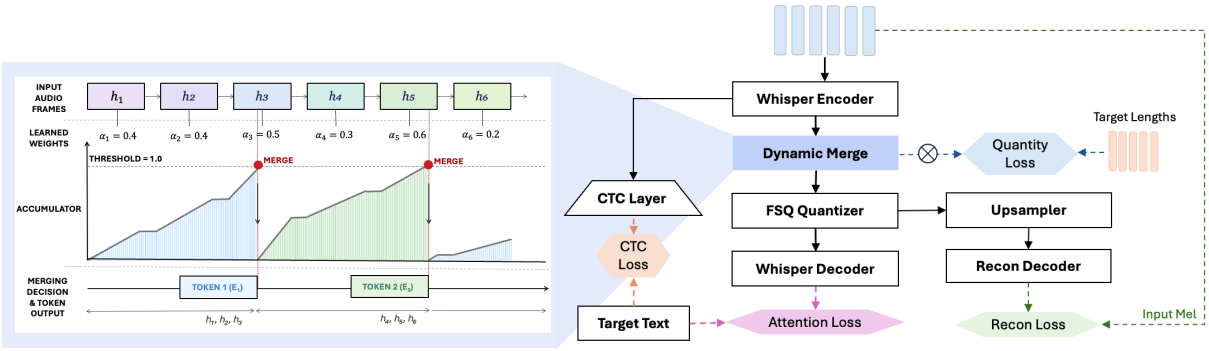

Figure 2: The Dynamic Compression Tokenizer. **Left:** The soft-accumulation mechanism of the Dynamic Merge Module. Learned frame weights α_t are progressively summed until the threshold $\Theta = 1.0$ is reached, triggering a token boundary. Frames h_1, h_2, h_3 are soft-aggregated into Token 1 (E_1), and h_4, h_5, h_6 into Token 2 (E_2). **Right:** The complete dual-probing architecture.

fore, to truly isolate a "pure" semantic representation without sacrificing comprehension accuracy, we require a paradigm shift in tokenization. We must abandon fixed-stride constraints in favor of a **dynamic compression mechanism**—one that intelligently integrates information and aligns token emissions strictly with the asynchronous semantic boundaries of the continuous speech signal (Dong and Xu, 2020; Gao et al., 2022).

4 Methodology: Dynamic Compression Tokenizer

To explicitly isolate pure semantic representations at ultra-low frame rates while overcoming the arbitrary truncation of fixed-stride mechanisms, we propose a novel dynamic tokenization architecture. The rationale behind this design is deeply rooted in the physical reality of human speech: acoustic transitions are inherently asynchronous. Drawing inspiration from the linguistic theory of distinctive features (Liu, 1996; Stevens, 2002; He et al., 2019; Zhang et al., 2025d,c,e) and dynamic sequence merging techniques in other domains (Dong and Xu, 2020; Gao et al., 2022; Baade et al., 2025), our architecture is meticulously designed to align token emissions with actual semantic boundaries rather than rigid temporal grids.

As illustrated in Figure 2, our pipeline realizes this through four components, each motivated by a specific requirement identified in the preceding analysis. First, we employ a pre-trained Whisper encoder as the continuous semantic backbone, leveraging its established linguistic robustness to avoid introducing encoder quality as a confounding variable. Second, a Dynamic Merge Module soft-accumulates these continuous features according to their acoustic transitions, directly addressing the fixed-stride boundary corruption identified in Section 3. Third, a Finite Scalar Quantization bottleneck enforces strict discrete compression, providing the informational constraint necessary to push tokens toward the ultra-low-rate regime under test. Finally, a multi-objective decoding framework branches into independent discriminative and generative pathways, enabling the Dual-Probing Protocol that evaluates semantic integrity and generative viability from the same token sequence.

4.1 Information Weight Prediction and Soft-Accumulation

Given an input Log-Mel Spectrogram, the semantic encoder extracts a sequence of continuous acoustic features $H = \{h_1, h_2, \ldots, h_T\} \in \mathbb{R}^{T \times D}$, where T is the number of acoustic frames. Instead of rigidly downsampling H by a fixed temporal window, our Dynamic Merge Module predicts a frame-level information weight $\alpha_t \in (0, 1)$ for each frame h_t.

A 1D convolutional layer followed by a linear projection and a sigmoid activation is employed to estimate these weights. To enhance sparsity and suppress acoustic noise, we apply a smoothing factor λ_s and a noise threshold λ_n:

$$\alpha_t = \mathrm{ReLU}(\sigma(\mathrm{Conv1D}(H)_t \cdot \mathbf{W} + b) \cdot \lambda_s - \lambda_n) \quad (1)$$

The sequence of weights $\alpha = \{\alpha_1, \ldots, \alpha_T\}$ represents the semantic boundary probabilities. In standard dynamic formulations, a new token boundary is dynamically triggered whenever the accumulated weight $S_t = \sum_{\tau=1}^{t} \alpha_\tau$ exceeds an integer threshold (typically 1.0).

4.2 Macroscopic Fixed-Ratio with Microscopic Dynamic Alignment

Conventional mechanisms determine the final sequence length based dynamically on the latent semantic content, rendering the overall compression rate unpredictable. However, to empirically stress-test the "unified token" assumption, we must evaluate the tokens at a strictly controlled, ultra-low frame rate. To bridge this gap, we introduce an **Information Scaling Paradigm**.

We impose a strict, predefined global compression ratio R (e.g., $R = 8$ converts 50Hz features to 6.25Hz). For an input sequence of length T, we explicitly compute the exact target compressed length $N = \max(1, \lfloor T/R \rceil)$. Prior to the integrate-and-fire operation, we globally scale the predicted information weights α_t such that their total sum strictly equals the target length N:

$$\hat{\alpha}_t = \alpha_t \cdot \frac{N}{\sum_{i=1}^{T} \alpha_i} \tag{2}$$

This scaling constraint (**Macroscopic Fixed-Ratio**) forces the module to compress the entire utterance into exactly N tokens. Crucially, the network retains complete freedom over the distribution of $\hat{\alpha}_t$ (**Microscopic Dynamic Alignment**). It learns to reallocate its constrained "token budget," allocating higher weights to dense semantic transitions and lower weights to steady-state acoustics. The continuous frames h_t are then softly aggregated into a sequence of dynamically merged continuous tokens $C = \{c_1, \ldots, c_N\}$.

4.3 Quantization and Dual-Decoder Architecture

To enforce a rigorous discrete informational bottleneck, the merged sequence C is passed through an FSQ Quantizer, mapping each c_j to a discrete codebook vector to obtain the final semantic tokens $Z = \{z_1, z_2, \ldots, z_N\}$.

To expose the generative limitations of these highly compressed tokens, our architecture bifurcates into a dual-decoding pathway. First, Z is fed into a **Whisper Decoder** and a **CTC Layer** (Watanabe et al., 2017) for autoregressive generation and alignment, ensuring the tokens maintain absolute semantic integrity (measured by WER). Second, Z is optionally passed to a lightweight **Recon Decoder**. To completely eliminate temporal misalignment as a confounding variable in generation, we utilize the cumulative weights $\hat{S}_t = \sum_{\tau=1}^{t} \hat{\alpha}_\tau$ to perform an Oracle Target Length Upsampling. The compressed sequence Z is mapped exactly back to the original temporal resolution T via $u_t = z_{\lfloor \hat{S}_t \rfloor}$, which explicitly aligns the discrete tokens to their physical acoustic boundaries before acoustic reconstruction.

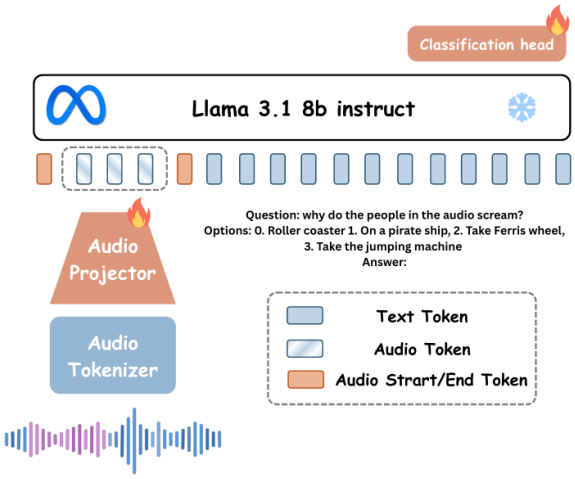

Figure 3: Downstream evaluation framework. Discrete VQ codes from the audio tokenizer are mapped to language model embeddings via a trainable Audio Projector. The Llama 3.1 8B model (frozen) processes audio tokens with textual questions for classification. Only the Audio Projector and classification head are trainable, isolating tokenization quality effects.

4.4 Overall Optimization Objectives

The entire architecture is optimized end-to-end using a multi-task learning framework. The total loss $\mathcal{L}_{total}$ comprises fundamental objectives for semantic alignment and compression, along with an optional generative probe:

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{CTC} + \mathcal{L}_{Attn} \\ + \lambda_{qua}\mathcal{L}_{Qua} + \lambda_{recon}\mathcal{L}_{Recon}\end{aligned} \tag{3}$$

Semantic Understanding Losses. To guarantee that the ultra-low frequency tokens capture pure linguistic content, we apply a Connectionist Temporal Classification loss ($\mathcal{L}_{CTC}$) on the latent projections and an Attention-based cross-entropy loss ($\mathcal{L}_{Attn}$) from the Whisper Decoder. Both are strictly supervised by the ground-truth target text.

Quantity Loss. To guide the Dynamic Merge Module in predicting reasonable semantic boundaries before the forced scaling is applied, we penalize the L_1 distance between the predicted unscaled token sum and the exact target compressed length N:

$$\mathcal{L}_{Qua} = \frac{1}{\sum N} \sum \left| \sum_{t=1}^{T} \alpha_t - N \right| \quad (4)$$

Acoustic Reconstruction Loss. As an *optional* generative probe to explicitly test the "unified token" hypothesis, we train the Recon Decoder to reconstruct the original Log-Mel Spectrogram from the upsampled discrete tokens. This process is supervised by a Mean Squared Error loss ($\mathcal{L}_{Recon}$) applied directly to the continuous acoustic space.

5 Evaluation Framework: The Dual-Probing Protocol

The dynamic tokenizer developed in Section 4 provides the necessary control variable: it isolates pure semantic states via extreme compression without introducing the confounding alignment errors typical of fixed-stride methods. To test whether these isolated tokens are sufficient for both understanding and generation, we propose a Dual-Probing Protocol that evaluates the exact same token sequence Z through two independent pathways.

Discriminative Understanding Probe. To validate semantic integrity beyond WER, we formulate audio understanding as a multiple-choice AQA classification task (Sakshi et al., 2024). As illustrated in Figure 3, the discrete FSQ codes are projected into a frozen LLM's embedding space via a trainable Audio Projector. By freezing the LLM backbone, any downstream accuracy is exclusively attributable to the informational density of the tokens themselves, not the reasoning capacity of the language model.

Generative Probe. To stress-test generative viability, we employ a Flow Matching decoder (Lipman et al., 2022) conditioned on the oracle upsampled sequence U (Section 4.3), which guarantees perfect temporal alignment. This forces the generative model to rely entirely on the micro-temporal dynamics *within* the tokens. If the tokens excel in the Discriminative Probe but fail here, it constitutes direct evidence that extreme semantic compression eradicates the continuous gradients required for ODE-based synthesis. Full architectural details and training objectives for both probes are provided in Appendix B.

6 Experiments

6.1 Experimental Setup

Datasets. We train our dynamic tokenizer using a large-scale, multi-lingual corpus comprising thousands of hours of speech from LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), and Aishell (Fu et al., 2021), among other open-source datasets, ensuring robust phonetic coverage across English and Mandarin. For evaluation, we report CER on two standard Mandarin test sets from WenetSpeech (Zhang et al., 2022): **Test_Net** (23 hours, internet speech, matched domain) and **Test_Meeting** (15 hours, real meeting recordings, mismatched domain with far-field conditions).

Model and Compression. Our tokenizer backbone follows a Whisper-style Transformer architecture. The input 128-dimensional log-Mel spectrograms are first downsampled to 50Hz. We enforce a compression ratio of R=10 within the Dynamic Merge Module, yielding an ultra-low frame rate of 5.0Hz. The merged representations are quantized via FSQ with 7 dimensions and 4 levels per dimension (codebook size 4^7=16,384). The model is optimized end-to-end with Adam (peak LR 2×10^{-5}, 12k warmup steps) using the multi-task objective in Section 4.4. Full architectural specifications and hyperparameter configurations are detailed in Appendix A.

6.2 Discriminative Probe: Establishing the Semantic Upper Bound

The primary prerequisite for testing our "ODE Flatline" hypothesis is obtaining a sequence of tokens that is extremely compressed yet undeniably semantically rich. If the tokens used for our generative probe are semantically deficient, any subsequent acoustic collapse could be trivially dismissed as an artifact of information loss rather than a fundamental representational mismatch. We evaluate this semantic sufficiency through two dimensions: Character Error Rate (CER) on the WenetSpeech corpus (Table 1) and complex language reasoning via the AVQA task (Table 2).

The Confounding Variable of Fixed-Stride. As shown in Table 1 and Table 2, scaling standard fixed-stride downsampling to an ultra-low 5.5Hz severely truncates necessary semantic information (CERs exploding to 29.80% and 31.50%). Consequently, its downstream AVQA accuracy collapses to 0.5777. This confirms that fixed-stride truncation inherently corrupts the linguistic state, rendering

Table 1: Character Error Rate (CER %) comparison on WenetSpeech test sets. The results highlight the catastrophic semantic degradation of fixed-stride compression at ultra-low frequencies (∼5.5Hz) and the robust semantic preservation of our Dynamic Compression Tokenizer. †Trained without the acoustic reconstruction probe loss ($\mathcal{L}_{Recon}$), isolating the purely linguistic state.

Model	Rate	Test-Net	Test-Meeting
Continuous Reference (Uncompressed)			
Whisper v3 Baseline	50Hz	9.68	18.54
Fixed-Stride Compression			
Whisper + FSQ	12.5Hz	18.66	20.46
Whisper + FSQ	5.5Hz	29.80	31.50
Dynamic Compression (Ours)			
Dynamic FSQ	6.0Hz	14.47	16.14
Dynamic FSQ (w/ Recon)	5.0Hz	14.32	15.94
Dynamic FSQ (Pure Semantic)†	5.0Hz	**11.98**	**12.50**
Dynamic FSQ	4.0Hz	15.61	17.44

Table 2: Downstream LLM Comprehension Performance (AVQA Accuracy). Evaluated using a frozen LLaMA 3.1 8B backbone. Our 5.0Hz dynamic tokens set a rigorous empirical baseline for semantic saturation. †**Trained without the acoustic reconstruction probe loss ($\mathcal{L}_{Recon}$).**

Model	Approx. Rate	AVQA Acc.
Established SOTA Baselines		
SYLLABLELM	∼4-5Hz	0.5526
Speech Tokenizer	50Hz	0.5839
DAC Tokenizer	75-100Hz	0.6561
WavTokenizer	40/75Hz	0.6732
Fixed-Stride Compression		
Fixed-Stride FSQ	∼5.5Hz	0.5777
Dynamic Compression (Ours)		
Dynamic FSQ	6.0Hz	0.7015
Dynamic FSQ (w/ Recon)	5.0Hz	0.7139
Dynamic FSQ (Pure Semantic)†	5.0Hz	**0.7246**
Dynamic FSQ	4.0Hz	0.6526

it an invalid control variable for testing generative limits at extreme compression.

Isolating the Pure Semantic State. In contrast, our Dynamic Compression Tokenizer completely bypasses this methodological bottleneck. To prove these tokens possess sufficient informational depth for generation, we compare them against established high-framerate SOTA tokenizers (e.g., WavTokenizer at ∼50Hz) used in standard generative SLMs.

Operating at an ultra-low frame rate of **5.0Hz**, our dynamic tokens (w/ Recon) achieve a remarkable AVQA accuracy of **0.7139**, eclipsing the reasoning capabilities of standard 50-75Hz representations (0.65-0.67) and recent syllable-based models like SYLLABLELM (0.5526). **This comparison is not to claim a new state-of-the-art tokenizer**, but to establish a rigorous empirical baseline: our 5.0Hz tokens are unequivocally saturated with dense semantic and reasoning information.

The Paradoxical Trap. The most profound insight emerges from our ablation study. When the dynamic tokenizer is trained strictly without the acoustic reconstruction probe loss ($\mathcal{L}_{Recon}$), its downstream AVQA accuracy peaks at **0.7246** (Dynamic FSQ, Pure Semantic). This sets up a profound paradox. Relieved of the burden to preserve continuous acoustic micro-dynamics, the tokens' semantic categorization capability reaches its absolute upper bound. We have successfully isolated a discrete representation that exceeds the semantic capacity of current generative benchmarks, yet it has been systematically stripped of acoustic constraints. Having established these 5.0Hz tokens as flawless, ultra-dense vehicles for language comprehension, we have completely eliminated "information scarcity" as a potential excuse. We now push these highly capable semantic tokens into the Generative Flow Matching Probe to expose the true nature of the unified token illusion.

6.3 Generative Probe: Exposing the ODE Flatline

Having established in Tables 1 and 2 that the ultra-low-rate dynamic tokens remain semantically competent, we now turn to the critical question of generation. The central purpose of the Generative Probe is not to evaluate reconstruction quality in isolation, but to determine whether a representation that is demonstrably sufficient for discriminative understanding can also support continuous acoustic synthesis. This distinction is essential. If generation fails only because token boundaries are corrupted by rigid downsampling, then the failure says little about the intrinsic adequacy of semantic tokens themselves. However, once dynamic compression removes this alignment confound and oracle upsampling restores the target temporal resolution, any remaining collapse must be attributed to the representational content of the tokens rather than to boundary mismatch.

Figure 4 makes this distinction explicit. The reconstructed utterance is nearly perfectly matched in duration to the source (`duration_ratio` $= 0.9995$), and the reconstructed log-Mel spectrogram still retains visible correspondence to the original at the level of the coarse spectral envelope (`mel_corr` $= 0.6285$). These two observations are important: they indicate that the token sequence

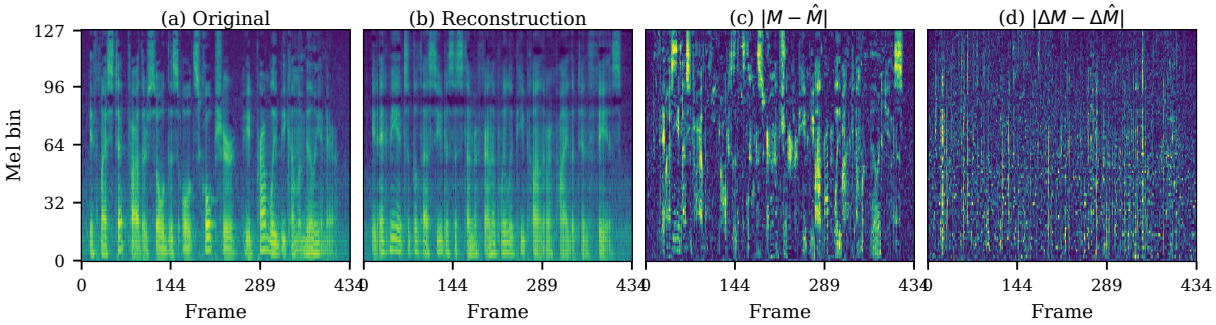

Figure 4: A representative reconstruction failure case from the dynamic tokenizer. Panels (a) and (b) show that the reconstructed log-Mel spectrogram roughly preserves the coarse semantic envelope and utterance duration, but panels (c) and (d) reveal substantial frame-wise and temporal-difference errors. This pattern indicates that ultra-low-rate semantic tokens retain enough information for coarse content recovery while failing to preserve the micro-dynamics required for clear articulation.

has not undergone trivial semantic collapse, nor is the failure primarily explained by temporal misalignment. At a macroscopic level, the utterance remains structurally recognizable.

Yet this apparent success is precisely where the WER trap becomes visible. The decisive failure emerges not in the coarse envelope, but in the local dynamics. The reconstruction exhibits substantial deviations in the Mel domain (`mel_mae` $= 11.11$), and, more critically, severe errors in the temporal-difference domain (`delta_mae` $= 7.90$ and `flux_mae` $= 4.04$). These error maps show that the model fails to recover the rapid frame-to-frame spectral transitions that encode articulation, including the short-timescale modulations needed to realize phonetic boundaries and continuous acoustic motion. Perceptually, this manifests as catastrophic articulatory loss and severe acoustic muffling—the generated speech becomes effectively unintelligible. The degradation in these micro-dynamic metrics is so stark that it precludes the necessity of subjective Mean Opinion Score (MOS) evaluations to confirm the generative collapse. This systemic failure is rigorously validated across diverse utterance lengths and verified over the entire test set in Appendix C and Appendix D. In other words, the compressed token sequence still carries enough information to preserve a semantically interpretable global structure, but it no longer provides the micro-dynamic gradients required to steer a continuous generative trajectory faithfully.

Taken together with the discriminative results—14.32/15.94 CER and 0.7139 AVQA accuracy at 5.0Hz, rising to 0.7246 under purely semantic training—these observations rule out information scarcity as an explanation for the generative failure. The tokens retain sufficient content for robust language understanding, yet the micro-dynamic gradients required by ODE-based synthesis have been irrevocably discarded. This confirms that low CER and generative adequacy are not interchangeable objectives: the very compression that sharpens semantic categorization strips away the continuous phonetic trajectories on which acoustic generation depends.

7 Conclusion

This work exposes a fundamental flaw in the prevailing assumption that a single discrete token, validated by low WER, can universally serve both speech understanding and generation. By developing a dynamic compression tokenizer that achieves ultra-low frame rates without corrupting semantic boundaries, we isolate pure semantic representations as a rigorous control variable. Our Dual-Probing Protocol reveals a stark divergence: at 5.0Hz, these tokens attain strong discriminative performance—surpassing standard high-framerate representations in complex audio reasoning—yet cause severe articulatory collapse when conditioning ODE-based generative models under oracle temporal alignment. This *ODE Flatline* is not an engineering artifact but an information-theoretic consequence: the categorical abstraction rewarded by WER optimization is structurally orthogonal to the continuous micro-dynamic gradients demanded by acoustic synthesis. Our findings carry a direct implication: the pursuit of a unified, ultra-compressed speech token is fundamentally misguided. We advocate instead for explicitly decoupled representations—semantic tokens for understanding, acoustic tokens for generation—as the principled path toward truly capable speech language models.

Limitations

Our generative probe employs a single synthesis paradigm, ODE-based Flow Matching, which is currently the dominant framework for discrete-token-conditioned speech generation. While alternative architectures such as autoregressive waveform models or GAN-based vocoders may exhibit different failure modes, the structural root cause we identify—the erasure of continuous micro-dynamic gradients under extreme discrete compression—is upstream of any particular decoder; it resides in the token representation itself, and is therefore expected to constrain any generation method that requires sub-token temporal detail. All evaluations are conducted on Mandarin (WenetSpeech), a tonal language whose lexical pitch contours impose stringent micro-dynamic demands. This arguably makes our experimental setting a strong test case rather than a narrow one: if the representational gap already manifests in a language where fine pitch variation is semantically critical, it is unlikely to vanish in phonetically simpler contexts. Finally, this work is diagnostic rather than constructive—we expose the incompatibility but do not propose a concrete decoupled architecture. However, precisely delineating the failure boundary is a necessary prerequisite for principled design, and we believe the Dual-Probing Protocol introduced here provides a reusable evaluation framework for future work on decoupled speech representations.

References

Alan Baade, Puyuan Peng, and David Harwath. 2025. SyllableLM: Learning coarse semantic units for speech language models. In *The Thirteenth International Conference on Learning Representations*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Yue Cheng, Bochen Lyu, Weiwei Xing, and Zhanxing Zhu. 2025. Dycast: Learning dynamic causal structure from time series. In *The Thirteenth International Conference on Learning Representations*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Transactions on Machine Learning Research*.

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

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Linhao Dong and Bo Xu. 2020. Cif: Continuous integrate-and-fire for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6079–6083. IEEE.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, and 1 others. 2021. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proc. Interspeech 2022*, pages 2063–2067.

Yitian Gong, Luozhijie Jin, Ruifan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li, and Xipeng Qiu. 2025. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs. *arXiv preprint arXiv:2506.23325*.

Di He, Xuesong Yang, Boon Pang Lim, Yi Liang, Mark Hasegawa-Johnson, and Deming Chen. 2019. When ctc training meets acoustic landmarks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5996–6000. IEEE.

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

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Peter Ladefoged. 2001. *A course in phonetics*. Heinle & Heinle.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11523–11532.

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

Sharlene A Liu. 1996. Landmark detection for distinctive feature-based speech recognition. *The Journal of the Acoustical Society of America*, 100(5):3417–3430.

Liang Lu, Lingpeng Kong, Chris Dyer, Noah A Smith, and Steve Renals. 2016. Segmental recurrent neural networks for end-to-end speech recognition. In *Proc. Interspeech 2016*, pages 385–389.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. In *The Twelfth International Conference on Learning Representations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Kenneth N Stevens. 2000. *Acoustic phonetics*, volume 30. MIT press.

Kenneth N Stevens. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, and 1 others. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xusheng Yang, Long Zhou, Wenfu Wang, Kai Hu, Shulin Feng, Chenxing Li, Meng Yu, Dong Yu, and Yuexian Zou. 2025b. U-codec: Ultra low frame-rate neural speech codec for fast high-fidelity speech generation. *arXiv preprint arXiv:2510.16718*.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. 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)*, pages 6182–6186. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, and 1 others. 2025a. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Haoyang Zhang, Hexin Liu, Xiangyu Zhang, Qiquan Zhang, Yuchen Hu, Junqi Zhao, Fei Tian, Xuerui Yang, Leibny Paola Garcia, and Eng Siong Chng. 2025b. Impact of frame rates on speech tokenizer: A case study on mandarin and english. *arXiv preprint arXiv:2505.17076*.

Xiangyu Zhang, Beena Ahmed, and Julien Epps. 2025c. Why pre-trained models fail: Feature entanglement in multi-modal depression detection. *arXiv preprint arXiv:2503.06620*.

Xiangyu Zhang, Daijiao Liu, Tianyi Xiao, Cihan Xiao, Tünde Szalay, Mostafa Shahin, Beena Ahmed, and Julien Epps. 2025d. Auto-landmark: Acoustic landmark dataset and open-source toolkit for landmark extraction. In *Proc. Interspeech 2025*, pages 4263–4267.

Xiangyu Zhang, Hexin Liu, Qiquan Zhang, Beena Ahmed, and Julien Epps. 2025e. Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10019–10030.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. 2020. Neural networks fail to learn periodic functions and how to fix it. *Advances in Neural Information Processing Systems*, 33:1583–1594.

A Detailed Experimental Configurations

To facilitate reproducibility, we detail the complete hyperparameter configurations used for training the Dynamic Compression Tokenizer, the quantization bottleneck, and the generative reconstruction probe.

Acoustic Features. The input signals are processed into 128-dimensional log-Mel spectrograms, extracted using a 400-point FFT and a hop length of 160. The sampling rate is standardized to 16kHz.

Tokenizer Backbone. Both the encoder and decoder utilize a robust Transformer architecture:

- **Encoder:** 32 layers, 20 attention heads, 1280 hidden dimension, and 5120 linear units. The input layer uses a `conv1d2` module, which provides an initial $2\times$ temporal downsampling (resulting in 50Hz features).
- **Decoder:** 32 layers, 20 attention heads, 1280 hidden dimension, utilizing Whisper-style subword tokenization (vocabulary size: 51,866, including special tokens like `<sot>` and `<eot>`).

Dynamic Merge Module. The dynamic compression operates with the following hyperparameters to strictly enforce the ultra-low frame rate:

- **Integration Threshold:** 1.0, with a smooth factor of 1.0 and a noise threshold of 0.0.
- **Custom Initialization:** The output bias of the dynamic predictor is initialized to -1.5. This negative bias ensures smaller initial information weights (α_t), stabilizing the early stages of training under extreme compression constraints.
- **Train-Infer Consistency:** Scaling is strictly enforced during both training and inference to guarantee exactly aligned $N = \lfloor T/R \rceil$ token lengths.

Quantization (FSQ). The discrete bottleneck is applied immediately after the 32nd encoder layer. We employ Finite Scalar Quantization (FSQ) mapped to 7 dimensions, with 4 levels per dimension (i.e., $[4, 4, 4, 4, 4, 4, 4]$). Positional embeddings are injected post-quantization to retain global temporal awareness for the downstream probes.

Generative Reconstruction Probe. The lightweight continuous decoder used to verify acoustic collapse comprises a 3-layer 1D ResNet architecture. It takes the 1280-dimensional upsampled discrete tokens and reconstructs the 128-dimensional log-Mel spectrograms. The network utilizes Snake activations (Ziyin et al., 2020) and employs hierarchical transposed convolutions with strides of $[2, 2, 2]$ to achieve an $8\times$ upsampling back to the acoustic resolution.

Optimization. The model is optimized using Adam with a peak learning rate of 2×10^{-5} and 12,000 warmup steps. We use dynamic batching (max 24,000 frames per batch) with a gradient accumulation step of 2 and gradient clipping capped at 5.0.

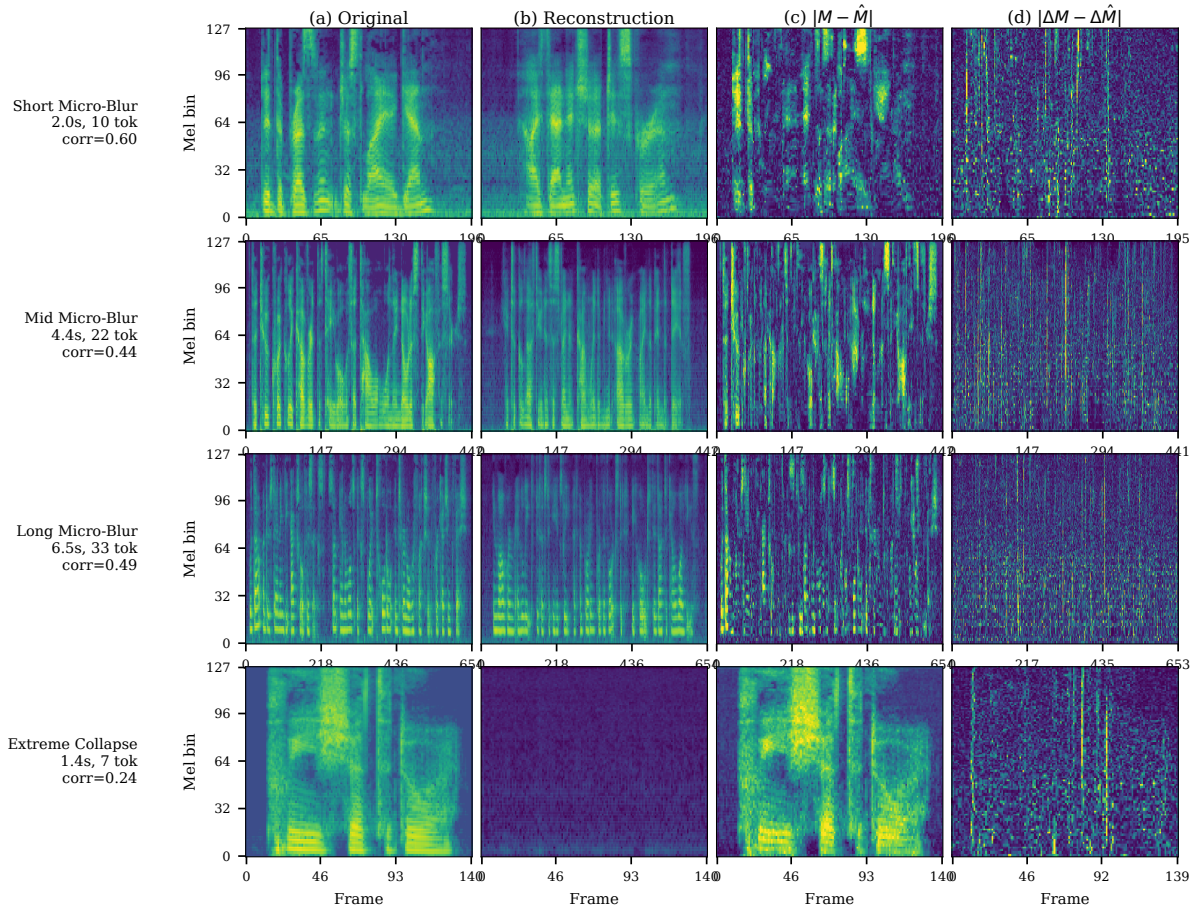

Figure 5: Additional reconstruction failures from the dynamic tokenizer under oracle duration alignment. Rows correspond to short-, medium-, and long-duration utterances, plus an extreme collapse case. Across diverse utterance lengths, the reconstructed spectrograms retain gross utterance extent but still exhibit large absolute and temporal-difference deviations, indicating that the generative deficiency is systematic rather than confined to a single sample.

B Dual-Probing Protocol: Architectural Details

This appendix provides the complete architectural and training details for the two evaluation pathways introduced in Section 5.

B.1 Discriminative Understanding Probe via Frozen LLMs

Relying solely on WER is insufficient for modern SLMs; therefore, drawing inspiration from recent advances in discrete token projection techniques and LLM-based audio evaluation paradigms (Deshmukh et al., 2023), we formulate audio understanding as a multiple-choice AQA classification task (Sakshi et al., 2024). To strictly isolate the representational quality of Z from the intrinsic reasoning capacity of the LLM, we employ a completely frozen language model backbone.

The discrete FSQ codes are mapped into the continuous semantic space via a trainable Audio Projector:

$$E_{audio} = \text{Network}(\text{Embed}(Z)) \in \mathbb{R}^{N \times d_{llm}} \quad (5)$$

The projected embeddings E_{audio} are concatenated with the text embeddings of the question and its options. By freezing the LLM and only training the projector and classification head, any success in downstream accuracy is exclusively attributable to the informational density of the underlying tokenizer. This probe is explicitly designed to confirm that our dynamic compression preserves deep linguistic reasoning capabilities, establishing the tokens as highly effective for comprehension tasks.

B.2 Generative Probe via Continuous Flow Matching

We employ a Flow Matching decoder, the current standard for Whisper-style discrete generation. As hypothesized, high-fidelity acoustic generation via FM requires the conditioning signal to provide temporally dynamic gradients to the ODE solver. To

test this without the confounding variable of temporal misalignment, we condition the FM model on the oracle upsampled sequence $U \in \mathbb{R}^{T \times D}$ (as defined in Section 4.3). The generative process is formulated as an ODE, $\frac{dx_t}{dt} = v_\theta(x_t, t, U)$, where v_θ is trained to match the target vector field between a Gaussian prior x_0 and the continuous log-Mel spectrogram x_1:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_1,x_0} \left[\| v_\theta(x_t, t, U) - (x_1 - x_0) \|_2^2 \right] \quad (6)$$

By explicitly feeding the FM decoder with U—which guarantees perfect physical acoustic boundaries—we force the generative model to rely entirely on the micro-temporal dynamics *within* the tokens. If the tokens excel in the Discriminative Probe but fail here, resulting in severe articulation blur, it serves as definitive empirical proof of our hypothesis.

C Additional Qualitative Evidence for the ODE Flatline

Figure 5 extends the main-text qualitative evidence by presenting reconstruction results across short, medium, and long utterances, as well as an extreme collapse case. The recurrence of the same failure pattern across substantially different durations confirms that the phenomenon is systematic rather than sample-specific. Across all 495 reconstruction cases, the duration ratio is tightly concentrated around 1.0 (median 0.9978, interquartile range 0.9965–0.9989), indicating that oracle upsampling successfully eliminates temporal mismatch as a confounding factor. The failure therefore cannot be attributed to duration error or token-frame misalignment.

In the short-, medium-, and long-duration examples, the reconstructed log-Mel spectrograms retain a rough correspondence to the source at the level of the coarse spectral envelope—broad energy placement, phrase-level segmentation, and approximate frequency-band occupancy remain visible. This confirms that the compressed tokens still encode sufficient information to preserve a global semantic scaffold, consistent with their strong CER and AVQA performance reported in the main text. However, the dominant deviations are consistently concentrated in the fine-grained structures that reflect rapid spectral transitions and short-lived articulatory events. The temporal-difference and spectral-flux error maps reveal that the largest errors arise precisely in the frame-to-frame change patterns that encode phonetic boundaries and continuous acoustic motion. Notably, even the short-duration example ($\sim$2 seconds) exhibits the same qualitative signature, ruling out the explanation that micro-dynamic blur is merely a byproduct of long-horizon error accumulation. The bottleneck lies in the representational content of the token stream itself.

The extreme collapse case in the final row serves a complementary role: it demonstrates that once the conditioning signal becomes sufficiently impoverished, even the coarse spectral scaffold breaks down entirely. Together, these examples delineate a continuum of failure severity—from semantically scaffolded but articulatorily blurred reconstructions to near-total spectral collapse—providing further evidence that the gap between discriminative adequacy and generative adequacy is a persistent structural property of ultra-low-rate semantic tokens, not an artifact of any single experimental condition.

Metric	Median	IQR
Mel MAE ↓	14.99	13.59–16.68
Mel Corr ↑	0.471	0.390–0.543
Delta-Mel MAE ↓	6.41	6.05–6.80
Flux MAE ↓	3.28	3.09–3.48
Duration Ratio ↑	0.9978	0.9965–0.9989

Table 3: Distribution of reconstruction metrics over all 495 dynamic-tokenizer reconstruction cases. We report the median and interquartile range (IQR) to avoid over-emphasizing a small number of extreme collapse samples. Duration Ratio close to 1.0 indicates that the reconstructed utterances largely preserve their global temporal extent, whereas the persistent Mel- and delta-domain errors reveal substantial spectral and micro-dynamic mismatch.

D Quantitative Reconstruction Statistics

To complement the qualitative examples in Figures 4 and 5, Table 3 summarizes the reconstruction statistics over all 495 evaluated utterances. Since the dynamic tokenizer removes boundary corruption introduced by fixed-stride compression, and oracle duration alignment eliminates gross temporal mismatch during upsampling, the remaining errors can be attributed more directly to representational insufficiency rather than architectural confounds.

The most revealing pattern is the sharp asymmetry between temporal preservation and spectral fidelity. The Duration Ratio is tightly concentrated near 1.0 (median 0.9978, IQR 0.9965–0.9989), confirming that reconstructed utterances preserve nearly correct temporal extent in the vast majority

of cases. The decoder is not failing because the utterance is globally misaligned—this rules out the familiar alignment pathology of low-rate tokenization as the primary explanation.

However, this temporal fidelity does not translate into acoustic intelligibility. The median Mel correlation is only 0.471, indicating at best a partial coarse resemblance to the source. More critically, the delta-domain errors remain persistently high: the median Delta-Mel MAE is 6.41 and the median Flux MAE is 3.28. These metrics probe frame-to-frame spectral evolution rather than static spectral occupancy, and their magnitude confirms that the dominant failure lies in the rapid local transitions encoding articulation—precisely the micro-dynamic gradients that a continuous generative decoder requires to steer its acoustic trajectory.

This asymmetry is exactly what one would predict if the token stream retains categorical semantic states but has lost the continuous micro-dynamic cues required for generation. A low-rate semantic tokenizer can preserve utterance-level structure, broad phonetic regions, and a coarse spectral envelope—information that is highly useful for discriminative objectives, consistent with the strong CER and AVQA results in the main text. But continuous synthesis additionally requires knowledge of *how* the spectrum should locally evolve between articulatory configurations, and this is precisely where the representation fails. The quantitative footprint thus mirrors the qualitative evidence: global timing survives, local dynamics do not. This dissociation provides aggregate statistical confirmation that the failure of ultra-low-rate semantic tokens is not one of generic quality degradation, but a structural mismatch between discriminative sufficiency and generative adequacy.