Neural Imitation of Human Perceptual Response to High-Resolution Audio

Baicheng Huang[1], Xinyi Pan[2], Haiwei Chai[2] (Oral Presenter), Feng Zhu[2], Dong Liu[2], Xiaoyong Pan[2]

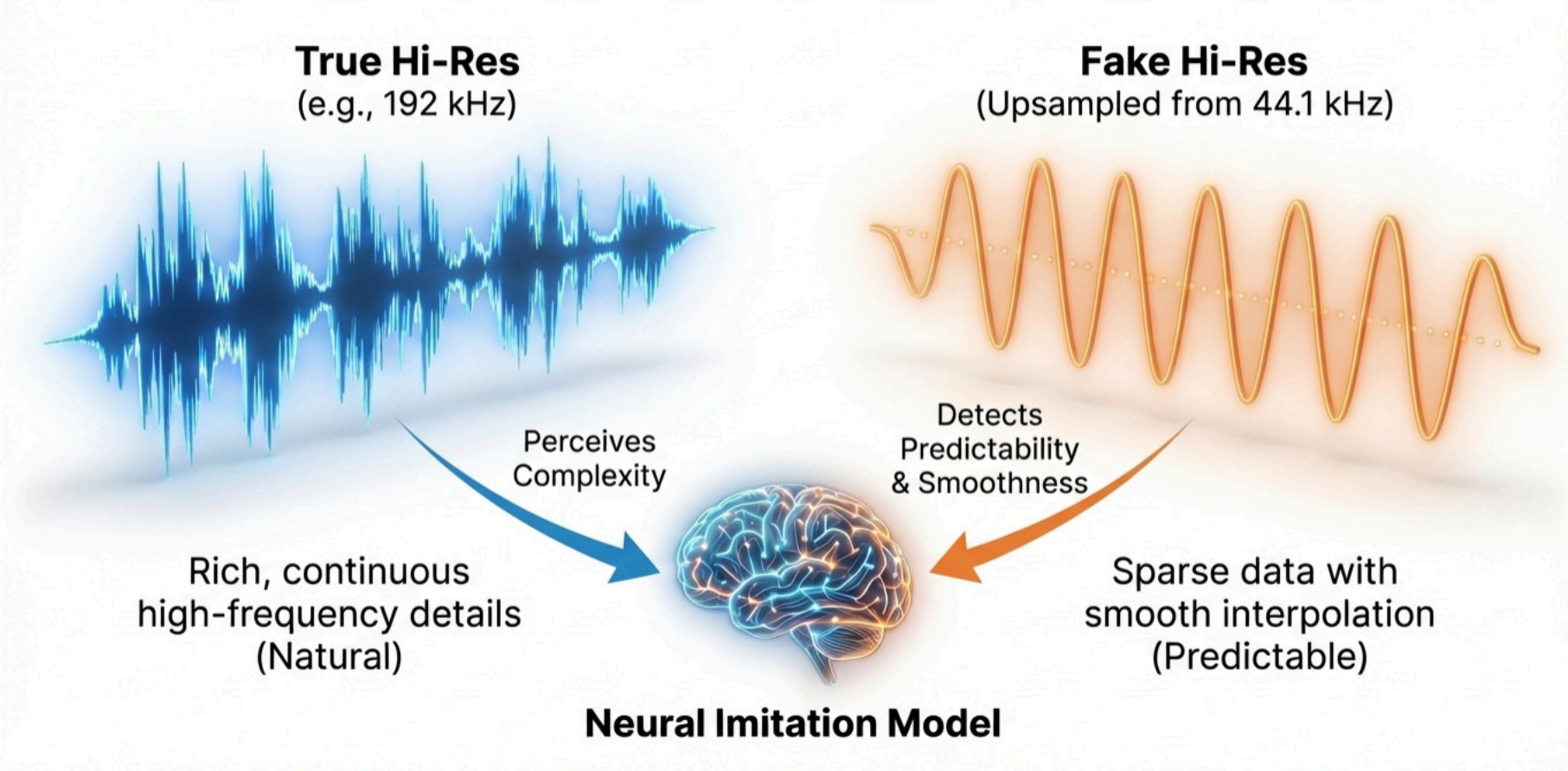
1. *Tsinghua University, Beijing, China*
2. *Conceptual Computing, Cambridge, MA, USA*

Introduction & Motivation

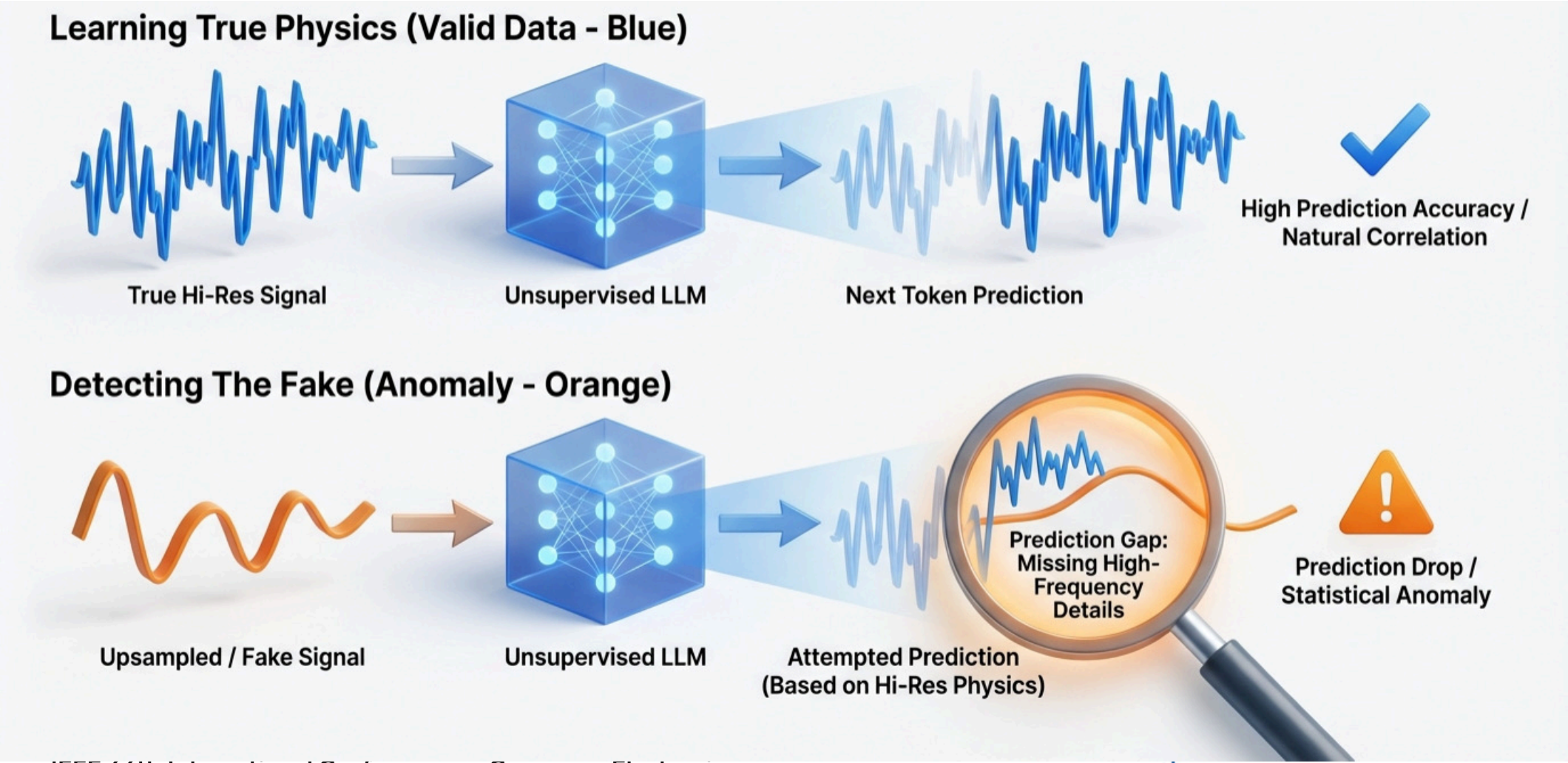
•**Context:** High-Resolution (Hi-Res) audio is becoming a standard in consumer electronics.

•**The Problem:** Many "Hi-Res" tracks are actually artificially up-sampled from CD-quality (44.1kHz).
•Human ears struggle to distinguish true Hi-Res from synthesized versions.
•Manual annotation of audio quality is expensive and subjective.

•**Objective:** Develop an unsupervised framework that imitates human perception to detect "fake" Hi-Res audio.



Conceptual Framework for Signal Authenticity Analysis



Methodology: Unsupervised Framework

Detail: Audio Tokenization & Serialization

Hierarchical Quantization (DAC)

•**Multi-Codebook Structure**: Decomposes each audio frame into **9 parallel discrete codes**.

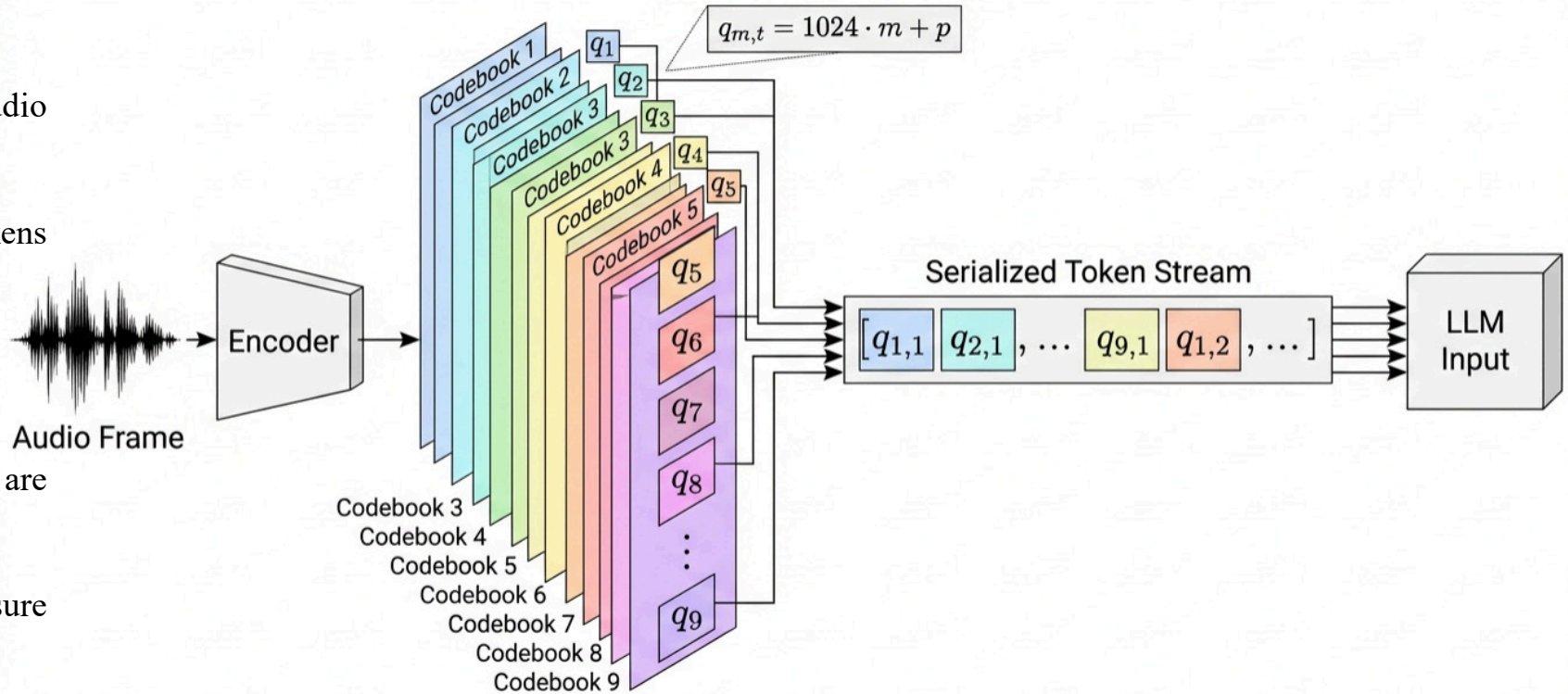
•**Vocabulary Mapping**: Maps 9 codebooks (1,024 tokens each) into a **9,216 non-overlapping integer space**.

1D Sequence Generation

•**Interleaving**: Parallel codes from a single timestep are flattened sequentially.

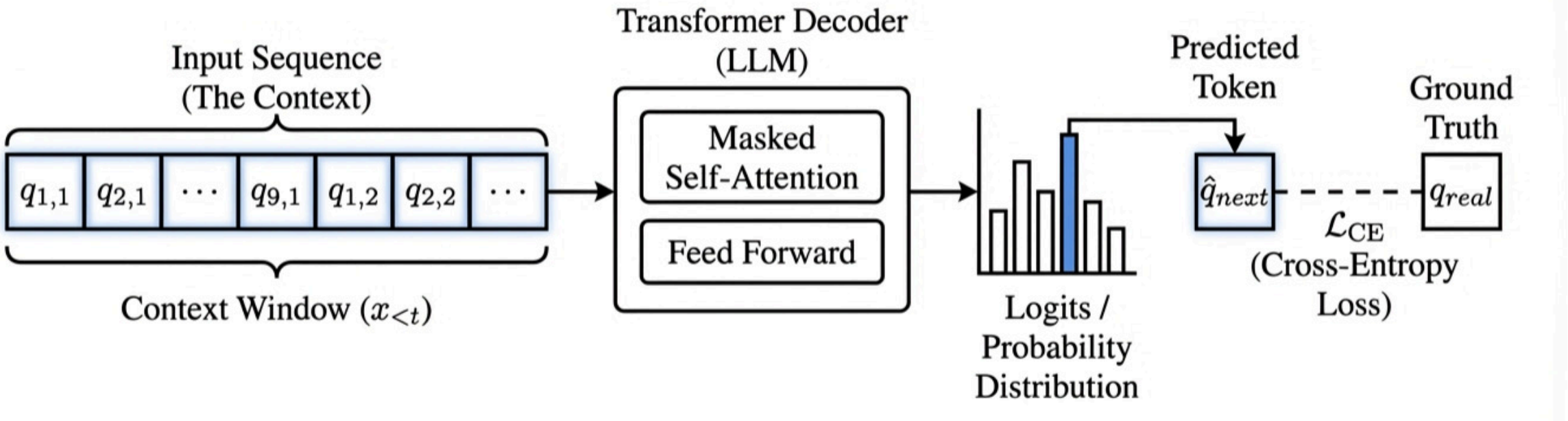
•**Frame Alignment**: Audio is precisely padded to ensure length is a multiple of the model stride (512 samples).

•**Result**: A continuous "token stream" that serves as the final input for LLM training.



Modeling via Large Language Models

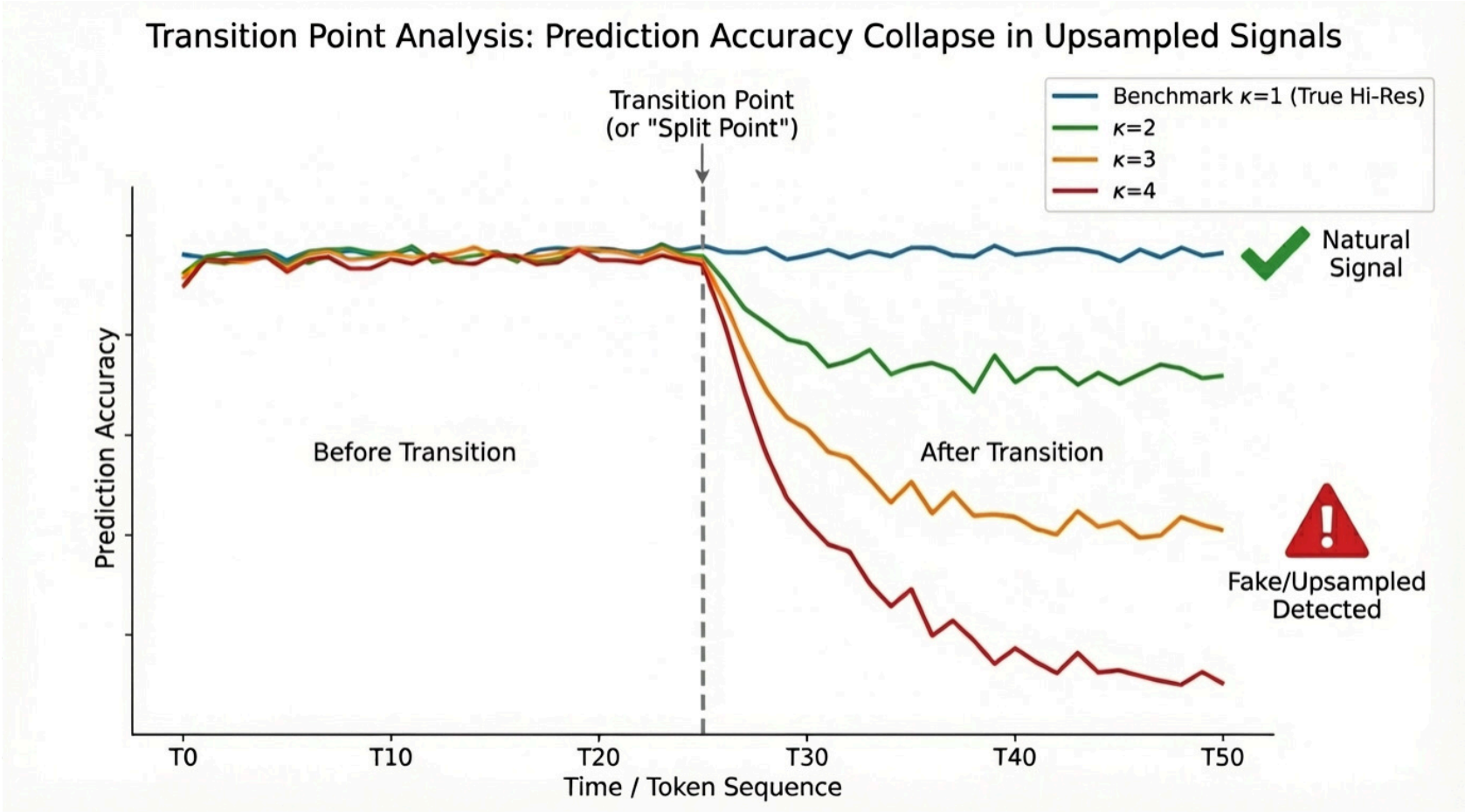
- **The Task: Next-token prediction.**



Experimental Design & Expected Results

Experimental Design:
•We constructed a hybrid audio stream: The first half is **True Hi-Res** (192 kHz), and the second half is **Fake/Upsampled** (degraded to κ=2, 3, 4).

Visual Analysis (The Graph):
Phase 1 (Left): The model predicts with high accuracy, indicating it understands the "grammar" of the natural signal.
Phase 2 (Right): At the "Transition Point" (where fake data begins), the model's accuracy **collapses**.



Result I: Distinguishing True vs. Fake Hi-Res

•**Observation:** The model shows a significant drop in prediction accuracy when processing artificially up-sampled content.

•**Why?** Up-sampling creates "mathematical" high frequencies that lack the natural harmonic phase relationships found in true Hi-Res recordings.

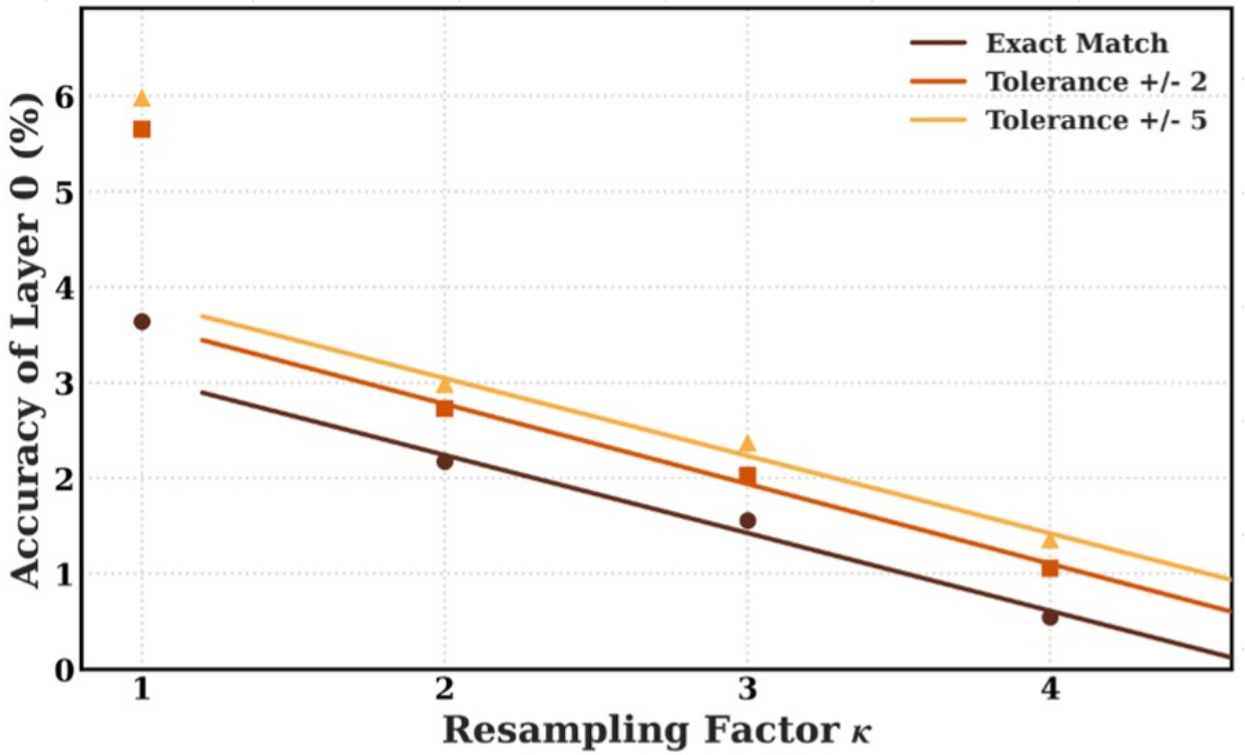


Fig. 2. Analysis of model sensitivity to original versus reconstructed high-frequency signals. Comparison of the model's prediction accuracy under different resampling factors.

Result II: Model Sensitivity to High-Res

This sharp drop in model confidence aligns to our expect.

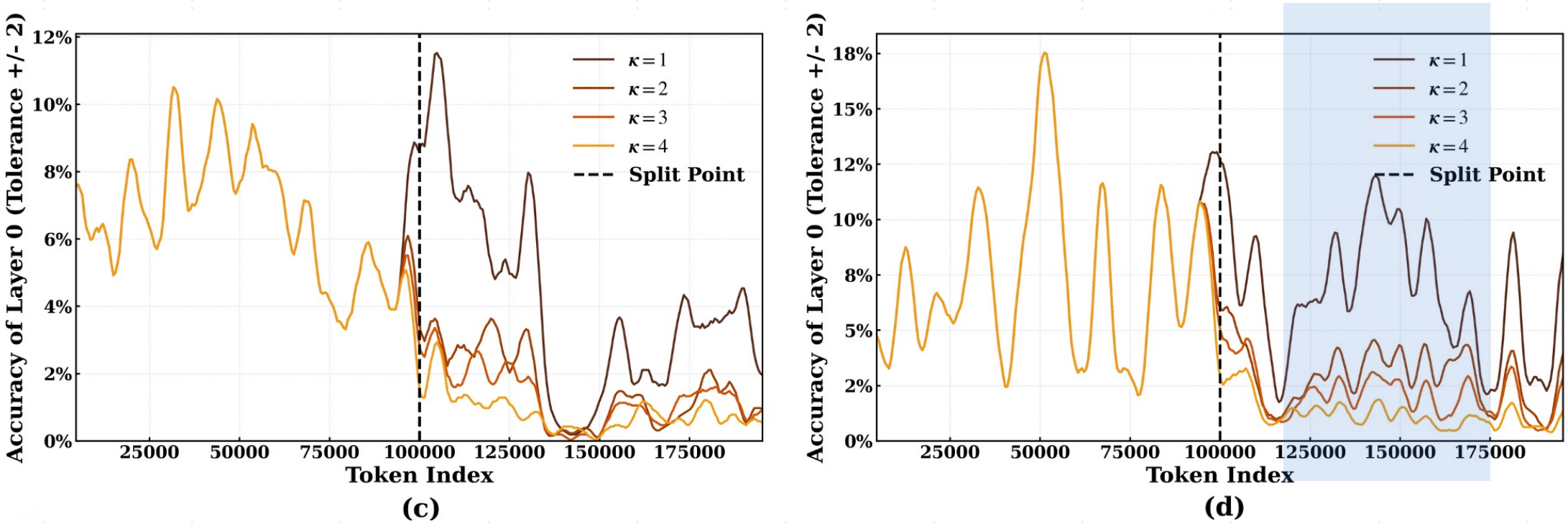


Fig.3. Model sensitivity to high-frequency information loss, evaluated across multiple audio segments. Subplots (a) - (h) each illustrate the experiment on an independent audio segment. Within each plot, the model's sliding window MLM accuracy is compared across four hybrid streams, corresponding to resampling factors κ =1 (darkest line, baseline) to κ =4 (lightest line). All streams are identical (baseline κ =1) before the "Split Point" (vertical dashed line) and are degraded to their respective n factor after it.

Result II: Model Sensitivity to High-Res

This sharp drop in model confidence aligns to our expect.

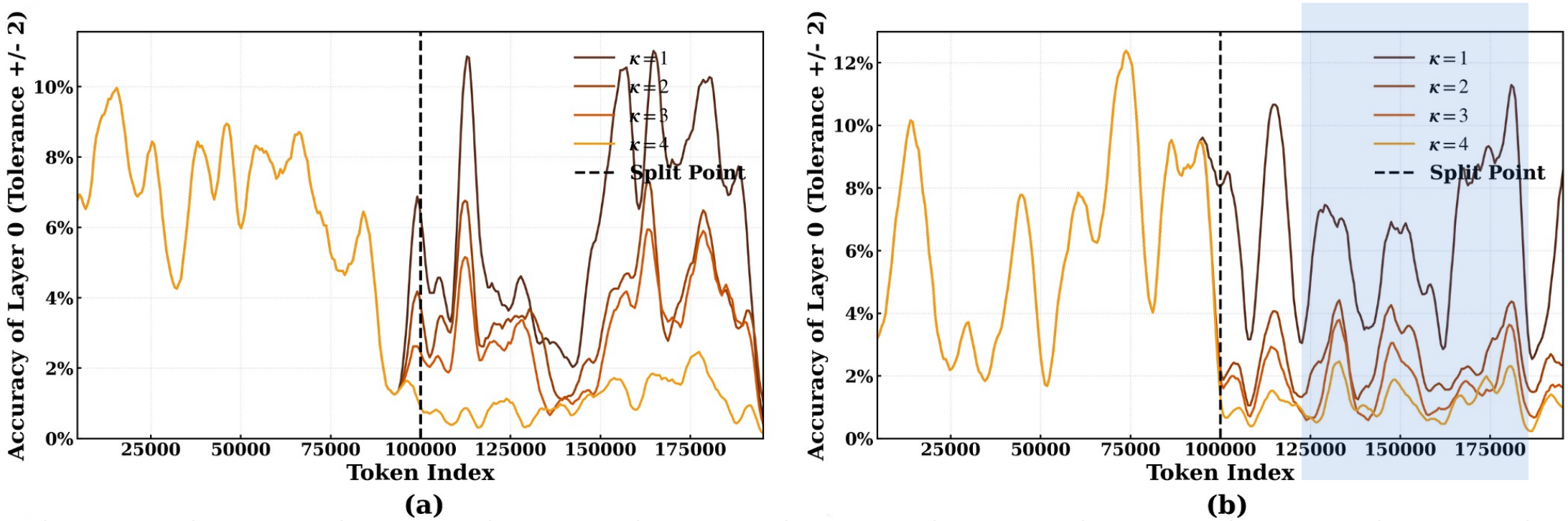


Fig.3. Model sensitivity to high-frequency information loss, evaluated across multiple audio segments. Subplots (a) - (h) each illustrate the experiment on an independent audio segment. Within each plot, the model's sliding window MLM accuracy is compared across four hybrid streams, corresponding to resampling factors κ =1 (darkest line, baseline) to κ =4 (lightest line). All streams are identical (baseline κ =1) before the "Split Point" (vertical dashed line) and are degraded to their respective n factor after it.

Result II: Model Sensitivity to High-Res

This sharp drop in model confidence aligns to our expect.

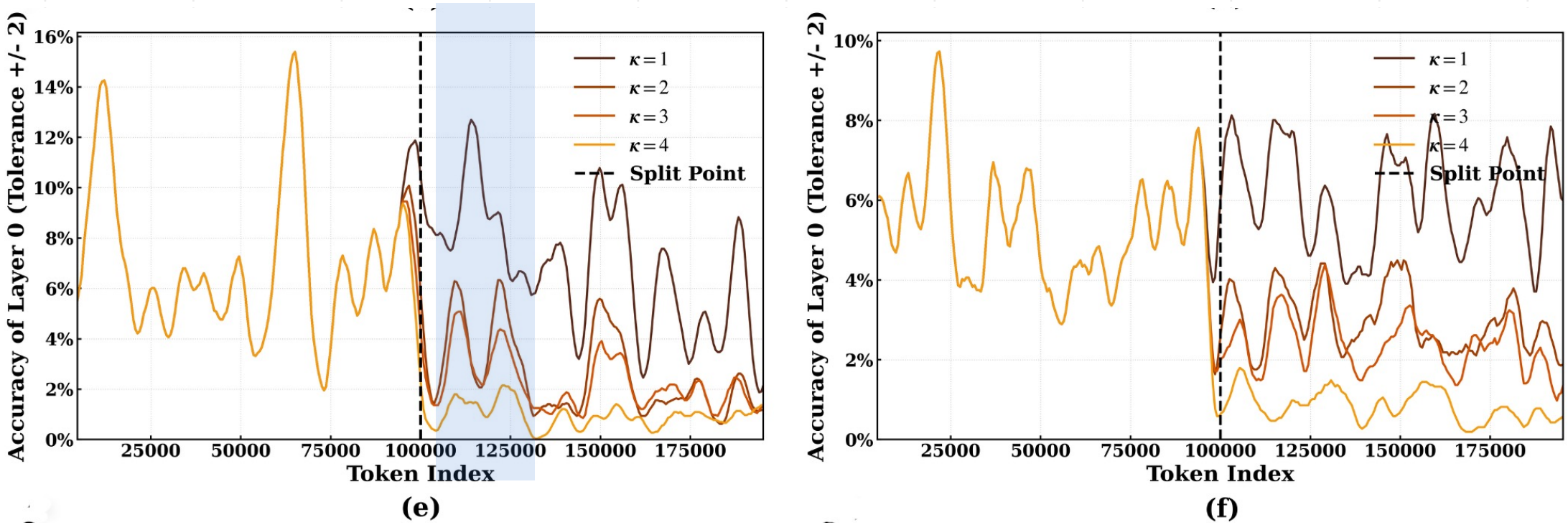


Fig.3. Model sensitivity to high-frequency information loss, evaluated across multiple audio segments. Subplots (a) - (h) each illustrate the experiment on an independent audio segment. Within each plot, the model's sliding window MLM accuracy is compared across four hybrid streams, corresponding to resampling factors κ =1 (darkest line, baseline) to κ =4 (lightest line). All streams are identical (baseline κ =1) before the "Split Point" (vertical dashed line) and are degraded to their respective n factor after it.

Result II: Model Sensitivity to High-Res

Non-ideal results were observed in a few audio clips.

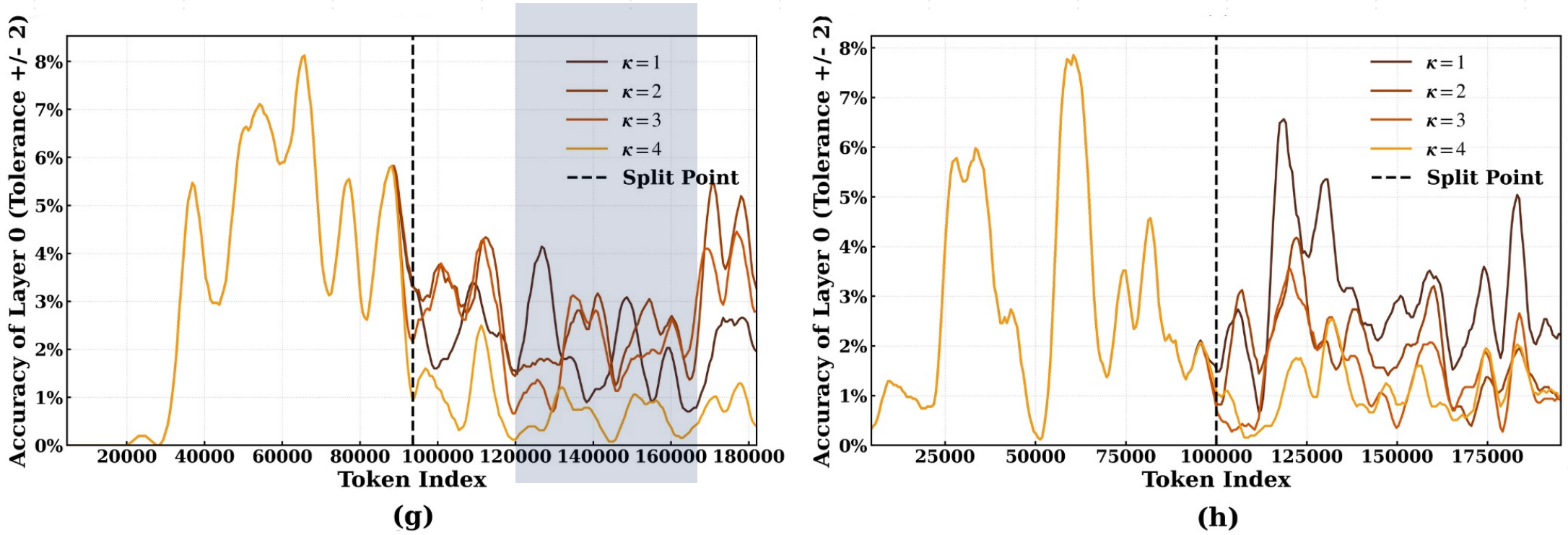


Fig.3. Model sensitivity to high-frequency information loss, evaluated across multiple audio segments. Subplots (a) - (h) each illustrate the experiment on an independent audio segment. Within each plot, the model's sliding window MLM accuracy is compared across four hybrid streams, corresponding to resampling factors κ =1 (darkest line, baseline) to κ =4 (lightest line). All streams are identical (baseline κ =1) before the "Split Point" (vertical dashed line) and are degraded to their respective n factor after it.

Discussion: The Role of LLMs in Audition

Beyond Speech: LLMs are not just for text; they can serve as powerful "Acoustic World Models."

Perceptual Imitation: By modeling the statistical structure of audio, the LLM effectively "learns" the "physics" of high-resolution sound.

Contributions:

Proposed an unsupervised framework for Hi-Res audio quality assessment.

Eliminated the need for manual annotation.

Demonstrated high sensitivity in detecting artificial up-sampling.

A new tool for digital forensics and quality control in the music industry.

Future Directions & Applications

- **Broadening the Scope of Artifact Detection**

Extend the unsupervised framework to detect other subtle degradations beyond upsampling, such as Lossy Compression Artifacts (MP3/AAC) and Generative AI Hallucinations.

- **From "Detection" to "Restoration"**

Utilize the model's "perceptual intuition" as a Perceptual Loss Function to guide generative models in restoring genuine high-frequency details (Super-Resolution).

- **Efficient Edge Implementation**

Optimize the Llama-based architecture via Model Quantization & Distillation.

Enable real-time "Fake Hi-Res" alerts on consumer devices (DAPs, Smartphones) without cloud dependency.

Thank You & Q&A

Thank you for your attention!

•Contact Info: Haiwei Chai (chaihw@conceputing.com)