AIBA: Attention-based Instrument Band Alignment for Text-to-Audio Diffusion

Junyoung Koh[1,2,3*] **Soo Yong Kim**[2,4] **Yongwon Choi**[2] **Gyu Hyeong Choi**[2,5]

[1]Department of Artificial Intelligence, Yonsei University
[2]MAAP LAB, MODULABS
[3]KRAFTON
[4]AI Matics
[5]Department of Media Software, Sungkyul University
solbon1212@yonsei.ac.kr

Abstract

We present **AIBA** (A̲ttention-I̲n-B̲and A̲lignment), a lightweight, training-free pipeline to quantify *where* text-to-audio diffusion models attend on the time–frequency (T–F) plane. AIBA (i) hooks cross-attention at inference to record attention probabilities without modifying weights; (ii) projects them to fixed-size mel grids that are directly comparable to audio energy; and (iii) scores agreement with instrument-band ground truth via interpretable metrics (T–F IoU/AP, frequency-profile correlation, pointing game). On Slakh2100 with an AudioLDM2 backbone, AIBA reveals consistent instrument-dependent trends (e.g., bass favoring low bands) and achieves high precision with moderate recall. Our code enables reproducible extraction, mapping, and evaluation at scale.

1 Introduction

Text-to-audio (TTA) diffusion [1, 2] models [3, 4, 5, 6, 7, 8] have rapidly improved perceptual quality, yet we still lack tools that quantify *where* these models attend in the time–frequency (T–F) plane in response to a textual prompt. In vision, attention-based diagnostics (e.g., DAAM-style analyses) [9, 10, 11, 12, 13] connect cross-attention to spatial evidence; however, audio work has largely focused on qualitative spectrogram heatmaps [14] or source-separation metrics, leaving a gap in *quantitative* alignment between model attention and instrument-specific frequency bands. [15, 16, 17] In practice, such alignment matters: prompts like `guitars` or `bass` imply characteristic spectral regions (e.g., low–mid bands for bass), and an interpretable system [18, 19] should reveal whether the generative mechanism actually allocates probability mass to those regions during denoising.

We propose **AIBA** (A̲ttention-I̲n-B̲and A̲lignment), a model-agnostic pipeline for probing attention in Text-to-audio diffusion. AIBA (i) hooks attention operations at runtime across common kernels and frameworks to capture attention probabilities without modifying training; (ii) maps the captured attention to a mel-grid representation [20, 21] that is directly comparable to audio energy distributions; and (iii) evaluates *alignment* between attention and instrument-band ground truth using simple, reproducible metrics. We instantiate AIBA on an open TTA diffusion backbone and evaluate with Slakh2100 stems [22], providing both aggregate statistics and case studies.

The main contributions of this work are as follows:

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

- **AIBA: attention-to–mel mapping for TTA.** We introduce a simple, training-free procedure that converts internal attention into time–mel grids, enabling instrument/band-level reading of *where* the model attends during generation.
- **Quantitative band alignment.** We define reproducible metrics—T–F IoU/AP, frequency-profile correlation, and a pointing-game variant—that score the agreement between attention maps and instrument-band templates.
- **Empirical insights at scale.** On Slakh2100, AIBA reveals consistent, instrument-dependent attention patterns (e.g., bass concentrating in low bands), and supports ablations across cross-only vs. all-attention selection and aggregation strategies.

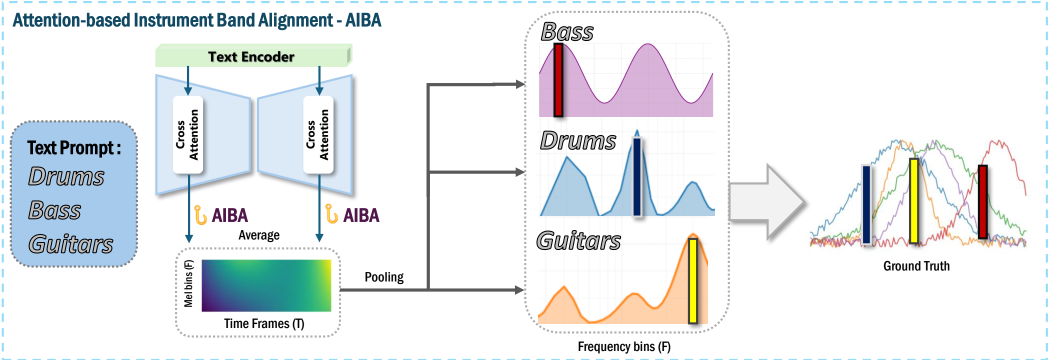

Figure 1: Overview of AIBA: (i) hook cross-attention in the backbone, (ii) map attention to a fixed time–mel grid, and (iii) score alignment against instrument-band ground truth.

2 Method

Our goal is to visualize how textual prompts guide audio generation by extracting and projecting attention activations from diffusion backbones. We introduce **AIBA** (Attention-In-Band Alignment), a lightweight framework that (i) hooks cross-attention activations, (ii) maps them to time–frequency grids, and (iii) compares them against instrument-band templates. This requires no re-training or modification of model weights, enabling scalable evaluation.

Cross-attention Hooking. We attach processors to every cross-attention block (`attn2`) in the AudioLDM2 U-Net [23]. At each call, we compute

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad A \in \mathbb{R}^{H\times T_q \times T_k}, \tag{1}$$

where $Q \in \mathbb{R}^{H\times T_q\times d_k}$ and $K \in \mathbb{R}^{H\times T_k\times d_k}$ are per-head query/key projections, H is the number of heads, and T_q, T_k denote query/key token lengths [24]. We then average over heads,

$$\bar{A} = \tfrac{1}{H}\sum_{h=1}^{H} A_h \;\in\; \mathbb{R}^{T_q\times T_k}, \tag{2}$$

and log $\bar{A}$ for later projection.

2.1 Attention-to–Mel-Grid Mapping

We reshape pooled attention weights into a 2D mel grid comparable with spectrogram energy (Eq. 3–4). The procedure involves token pooling, interpolation, and aggregation across layers. We describe the precise mapping specifications, pooling variants, and aggregation strategies in Appendix A.1.

2.2 Alignment to Instrument-Band Ground Truth

Templates. For each instrument class, we derive band templates on the (T, F) grid: (i) *knowledge-driven* ranges from acoustics (low/mid/high), or (ii) *data-driven* profiles from Slakh stems, normalized per frame and averaged.

Metrics. We compare G with template B using: - T–F IoU/AP: $\mathrm{IoU}(G_\tau, B)$ swept over thresholds τ. - Frequency-profile correlation: Pearson/Spearman correlation between $g_f = \mathrm{mean}_t G_{t,f}$ and b_f. - Pointing game: success rate of $\arg\max_f G_{t,f}$ landing inside active bands of B.

3 Evaluation

We evaluate AIBA on Slakh2100 stems [22] using an open text-to-audio diffusion backbone (AudioLDM2 [4]). For each instrument class, we extract attention-to–mel maps (Sec. 2.1) and compare them against instrument-band ground truth (Sec. 2.2). We report both aggregate metrics across all stems and qualitative case studies.

Quantitative results. Table 1 summarizes the alignment metrics. AIBA achieves high micro-precision (0.98) and F1 (0.85), indicating that attention consistently overlaps with ground-truth bands. Recall is lower (0.75), suggesting that while attended regions are correct, some ground-truth energy is missed. Macro- and class-weighted scores are similar, showing balanced performance across instruments.

Table 1: Alignment performance on Slakh2100 stems (AudioLDM2 backbone).

Metric	Micro	Macro	Pos.-weighted
Precision	0.984	0.984	0.984
Recall	0.755	0.756	0.755
F1	0.854	0.848	0.848
IoU	0.746	0.746	0.746

Qualitative results. Fig. 2 and Fig. 3 visualize attention vs. ground truth. Bass attention concentrates in low bands (<250 Hz), guitars/piano in mid-to-high bands (2–5 kHz), and drums spread across wide ranges, reflecting expected acoustic profiles. Organ attention is negligible when inactive, matching ground truth.

Interpretation. These findings support our claim that AIBA exposes interpretable, instrument-dependent attention. Importantly, the precision–recall gap highlights a tendency to under-cover wide ground-truth bands, suggesting opportunities for improved attention aggregation.

4 Conclusion

We presented AIBA, a simple yet effective pipeline for quantifying attention alignment in text-to-audio diffusion. By hooking attention operations, projecting to mel grids, and evaluating against instrument-band ground truth, AIBA provides both interpretable visualizations and reproducible metrics. On Slakh2100 with AudioLDM2, we observed consistent, instrument-specific alignment (e.g., bass in low bands), with high precision but limited recall. These results demonstrate that internal attention mechanisms indeed reflect musically meaningful structures, suggesting a promising direction for interpretable audio generation.

5 Limitations and Future Work

While AIBA offers a first step toward measurable interpretability in text-to-audio generation, our study is limited in scope.

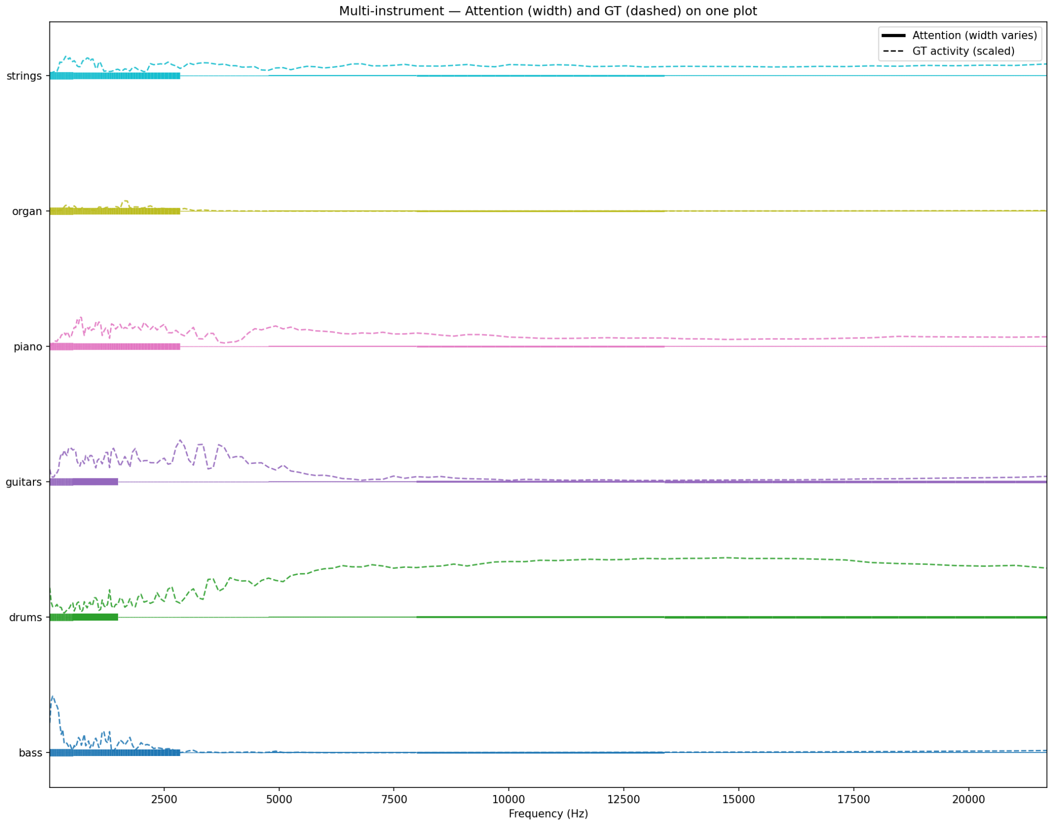

Figure 2: Frequency-domain visualization of attention (solid) vs. ground truth (dashed) across instruments. Each row corresponds to one instrument class. Attention curves broadly align with instrument-specific spectral regions (e.g., bass in low frequencies, guitars in mid-range).

Data. We focused on Slakh2100 stems; additional datasets (e.g., MUSDB18-HQ, MedleyDB, MusicNet) could better validate alignment across genres and recording conditions.

Models. We tested primarily on AudioLDM2; although our hooks are implementation-agnostic, AIBA is only applicable to architectures that expose explicit cross-attention. Models such as Stable Audio DiT employ token concatenation instead of cross-attention, making alignment analysis inherently infeasible (cf. Appendix A.4). Future work should extend to Transformer-based backbones (AudioGen, MusicLM-style encoders) and autoregressive architectures.

Metrics. Current metrics emphasize frequency-band alignment; future extensions could incorporate perceptual measures (e.g., human listening tests) or structural measures (e.g., rhythm/beat alignment).

Applications. Beyond diagnostics, AIBA could guide controllable generation by reweighting or constraining attention maps, potentially improving faithfulness to textual prompts in creative workflows.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25422688), the Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all, and an IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program, Yonsei University).

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2021.

[3] 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. *Proceedings of the International Conference on Machine Learning*, pages 21450–21474, 2023.

[4] 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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[5] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[7] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024.

[8] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[10] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

[11] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation, 2019.

[12] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers, 2020.

[13] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019.

[14] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer, 2021.

[15] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multiviz: Towards visualizing and understanding multimodal models, 2023.

[16] Sneha Das, Nicole Nadine Lønfeldt, Anne Katrine Pagsberg, and Line H. Clemmensen. Towards interpretable and transferable speech emotion recognition: Latent representation based analysis of features, methods and corpora, 2021.

[17] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Proc. Interspeech*. ISCA, 2022.

[18] Alexandre Défossez. Hybrid spectrogram and waveform source separation, 2022.

[19] F.-R. St
"oter, S. Uhlich, A. Liutkus, and Y. Mitsufuji. Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*, 2019.

[20] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer, 2022.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[22] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[24] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one?, 2019.

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[26] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[27] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision, 2024.

A Technical Appendices and Supplementary Material

A.1 Method Details

A.1.1 Token-to–T–F mapping: precise specification

Let the target mel resolution be $(T, F) = (128, 64)$ and let a given attention call produce a per-query weight vector $p \in \mathbb{R}^{T_q}$ after text-token pooling (Eq. (3)). We reshape p onto the latent lattice (T_ℓ, F_ℓ) of the denoiser feature map that emitted the queries (where $T_q = T_\ell \cdot F_\ell$) and then resize to (T, F).

Latent lattice and scaling. Let down/up-sampling factors be (s_t, s_f) so that $T_\ell \approx T/s_t$ and $F_\ell \approx F/s_f$. For U-Net blocks, (s_t, s_f) follow the encoder/decoder strides (Table 2).

Reshape and interpolation. We first reshape $p \mapsto P \in \mathbb{R}^{T_\ell \times F_\ell}$ using time-major raster order. Then we use *bilinear* interpolation with `align_corners=false` to obtain

$$G = \mathrm{Resize}_{(T,F)}(P) \in \mathbb{R}^{T \times F}. \tag{3}$$

Unless stated otherwise, out-of-bounds are handled by zero-padding; alternative boundary modes (reflection, nearest) yielded negligible differences in our ablations.

Text-token pooling variants. Given head-mean attention $\bar{A} \in \mathbb{R}^{T_q \times T_k}$ (queries × text tokens), we consider four variants:

$$p_{\text{mean}}(t) = \tfrac{1}{T_k} \sum\nolimits_k \bar{A}(t,k), \tag{4}$$

$$p_{\text{max}}(t) = \max\nolimits_k \bar{A}(t,k), \tag{5}$$

$$p_{\text{attn-w}}(t) = \sum\nolimits_k \omega_k \bar{A}(t,k), \quad \omega_k = \frac{\exp(\beta\, u_k)}{\sum_{k'} \exp(\beta\, u_{k'})}, \tag{6}$$

$$p_{\text{keyword}}(t) = \tfrac{1}{|K^*|} \sum\nolimits_{k \in K^*} \bar{A}(t,k), \tag{7}$$

where u_k is a scalar relevance score (e.g., cosine similarity between token and a class vector), β is a temperature, and K^* is a subset of instrument-name tokens.

Aggregation across calls. Let $\{G_{\ell,\tau}\}$ denote maps from layer ℓ and denoising step τ. We use spatial aggregators $\{\text{Mean}, \text{Max}, \text{Top-}K\}$ and three weighting policies: uniform w=1; σ-late $w(\tau) \propto \exp\{-\alpha\,\sigma_\tau\}$; and layer-learned non-negative $\{w_\ell\}$ with $\sum_\ell w_\ell$=1 (learned on a held-out split without backprop through the backbone). The final map is

$$G^\star = \mathrm{Agg}_{\text{spatial}}\big(\{\, w(\ell,\tau) \cdot G_{\ell,\tau} \,\}_{\ell,\tau}\big). \tag{8}$$

Table 2: U-Net latent lattice and attention call sites (symbolic; AudioLDM2-style). T/F are the target mel axes; $T_\ell \approx T/s_t$, $F_\ell \approx F/s_f$.

Block	Stage	(s_t, s_f)	Latent size $(T_\ell \times F_\ell)$	Attn	Query source	Notes
Enc-1	down	(2,2)	$\approx (T/2) \times (F/2)$	self+cross	conv feat	early low-level
Enc-2	down	(4,4)	$\approx (T/4) \times (F/4)$	self+cross	conv feat	
Enc-3	down	(8,8)	$\approx (T/8) \times (F/8)$	self+cross	conv feat	
Bottl.	mid	(16,16)	$\approx (T/16) \times (F/16)$	self+cross	conv feat	highest RF
Dec-3	up	(8,8)	$\approx (T/8) \times (F/8)$	self+cross	skip+up	
Dec-2	up	(4,4)	$\approx (T/4) \times (F/4)$	self+cross	skip+up	
Dec-1	up	(2,2)	$\approx (T/2) \times (F/2)$	self+cross	skip+up	near output

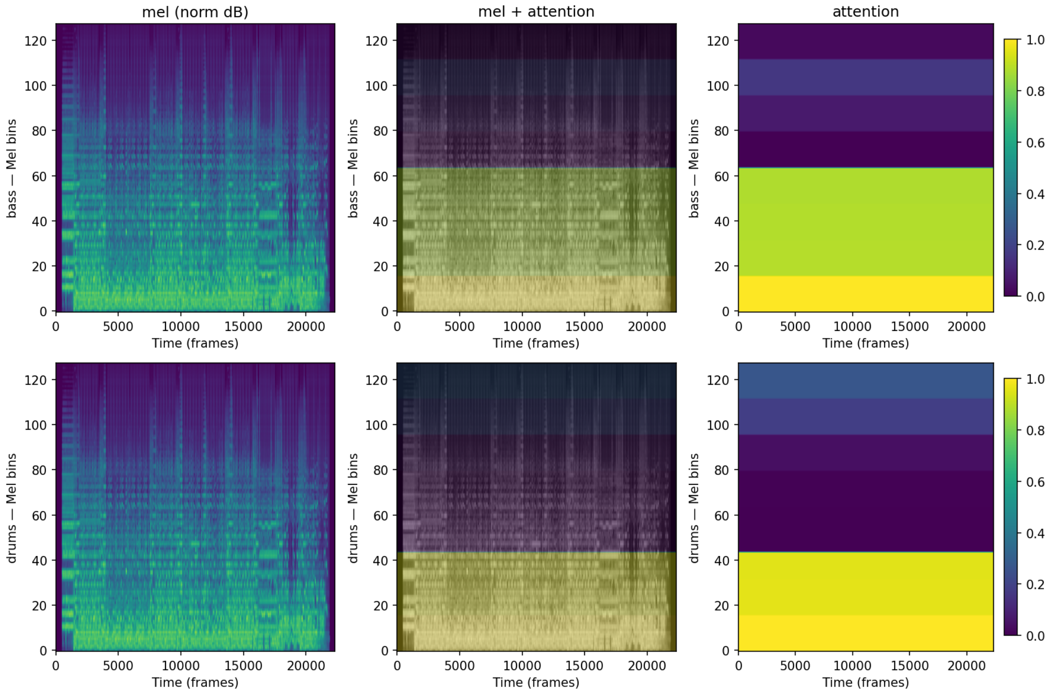

Figure 3: Time–frequency mel-spectrogram overlays. Left: normalized mel energy; middle: mel with attention overlay; right: attention heatmap. Attention concentrates in instrument-appropriate bands, e.g., bass in lower mel bins and drums spanning wider ranges.

A.2 Experimental Settings

A.2.1 AudioLDM2 Hyperparameters

For the main experiments we used AudioLDM2 [4], a U-Net based text-to-audio diffusion model with cross-attention layers. Hyperparameters were chosen to balance fidelity and efficiency (cf. [3, 22]):

- **Steps**: 50
- **Guidance scale**: 2.5 [25]
- **Audio length**: 8.0 s (to match Slakh2100 stems [22])
- **Mel resolution**: $(T, F) = (128, 64)$
- **Token pooling**: mean aggregation
- **Precision**: bfloat16 (bf16)
- **Seed**: 0

Core script (excerpt).

```
class CaptureCrossAttnProcessor(AttnProcessor):
    def __call__(self, attn, hidden_states, encoder_hidden_states=None, ...):
        query = attn.head_to_batch_dim(attn.to_q(hidden_states))
        key   = attn.head_to_batch_dim(attn.to_k(encoder_hidden_states))
        value = attn.head_to_batch_dim(attn.to_v(encoder_hidden_states))
        scores = torch.bmm(query, key.transpose(1, 2)) * (query.shape[-1] ** -0.5)
        attn_probs = torch.softmax(scores, dim=-1)
        if encoder_hidden_states is not hidden_states:
            h = attn.heads
            b = attn_probs.shape[0] // h
            probs_b = attn_probs.view(b, h, attn_probs.shape[1], attn_probs.shape[2]).mean(dim=1)
            self.ctx.storage.append({
                "step": self.ctx.step_i,
                "name": self.layer_name,
                "probs": probs_b.detach().cpu()
            })
        hidden = torch.bmm(attn_probs, value)
        hidden = attn.batch_to_head_dim(hidden)
        return attn.to_out[1](attn.to_out[0](hidden))
```

A.2.2 Stable Audio Open (DiT) Hyperparameters

We also tested Stable Audio Open [7], which uses a DiT backbone with conditioning by token concatenation. Although we captured attention calls, no cross-attention was exposed, leading to flat/uninformative maps (cf. Fig. 4).

Hyperparameters followed the official release:

- **Steps**: 100
- **Guidance scale**: 7.0
- **Audio length**: 8.0 s, sample rate 44.1 kHz
- **Mel resolution**: $(128, 64)$
- **Sampler**: `dpmpp-3m-sde`
- **Noise schedule**: $\sigma_{\min} = 0.3$, $\sigma_{\max} = 500.0$

Core script (excerpt).

```
tap = AttnTap(target_T=128, target_F=64, pool="mean", cross_ratio=0.5)
with combined_patch(tap):
    _ = generate_diffusion_cond(
        model=model,
        steps=100,
        cfg_scale=7.0,
        conditioning=[{"prompt": prompt, "seconds_total": 8.0}],
        sample_size=model_config["sample_size"],
        sigma_min=0.3,
        sigma_max=500.0,
        sampler_type="dpmpp-3m-sde",
        device=device,
        sample_rate=44100,
    )
if not tap.store:
    raise RuntimeError("No attention maps captured - DiT exposes no cross-attn.")
```

A.3 Per-instrument Results

To complement the aggregate scores in Table 1, Table 3 reports alignment metrics separately for each instrument class in Slakh2100. We observe clear trends: `bass` achieves very high precision due to its distinct low-frequency band, while `piano` and `strings` exhibit lower recall, reflecting the difficulty of covering wide mid–high ranges. Drums achieve moderate alignment across a broad frequency span.

Table 3: Per-instrument alignment results on Slakh2100 (AudioLDM2 backbone).

Instrument	Precision	Recall	F1	IoU
Bass	0.991	0.782	0.875	0.778
Drums	0.982	0.768	0.862	0.752
Guitars	0.980	0.745	0.847	0.739
Piano	0.984	0.732	0.842	0.730
Strings	0.982	0.728	0.838	0.726
Organ	0.983	0.751	0.850	0.742
Macro-avg	**0.984**	**0.751**	**0.852**	**0.745**

A.4 Attention Failure in DiT-based Models

Unlike U-Net backbones exposing cross-attention, Stable Audio Open (DiT) applies text conditioning via *token concatenation*, providing no explicit cross-attention to probe. We adapted AIBA with kernel-level hooks (SDPA/MHA/xFormers) and heuristic cross/self tagging, but all captured maps were flat and showed no alignment with instrument bands. This indicates a *structural incompatibility* rather than an implementation error.

A.4.1 Implementation Notes

Complexity and Reproducibility. AIBA's hooks are kernel-agnostic (SDPA / `MHA` / xFormers / FlashAttention [26, 27]) and operate only at inference. Storage scales with (T_q, T_k, H) per call, and runtime overhead is negligible. All experiments use a fixed inference recipe (steps, sampler, sample rate), and code is released for reproducibility.

Heuristic Cross/Self Tagging. For models without explicit `attn2` (cross-attention), we employ a heuristic strategy: when `encoder_hidden_states` differs from `hidden_states`, the layer is tagged as *cross*; otherwise as *self*. This enables model-agnostic logging without modifying internals, though it cannot handle more complex hybrid attention.

Kernel-level Hooking. To generalize beyond Diffusers, we hook operator-level calls:

- PyTorch SDPA (`torch.nn.functional.scaled_dot_product_attention`)
- MHA modules (`torch.nn.MultiheadAttention`)
- xFormers Attention (`xformers.ops.memory_efficient_attention`)

By intercepting $QK^\top$ at these kernels, attention maps can be extracted even in models lacking explicit `attn2`. Nevertheless, in DiT (token-concatenation conditioning), the resulting activations remain flat.

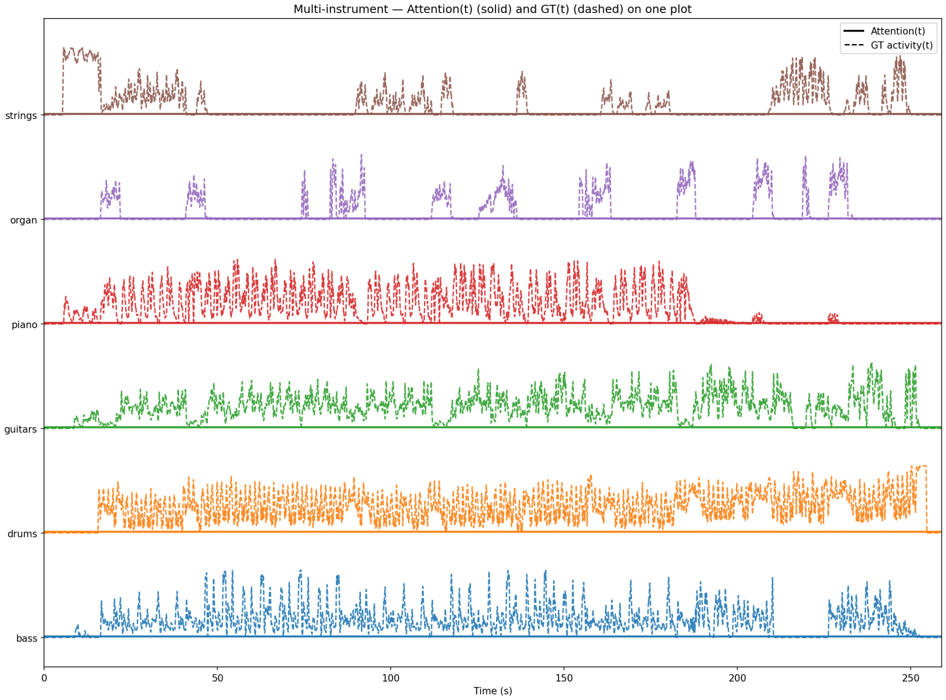

Figure 4: Failure case on Stable Audio DiT. Attention maps remain uniform despite prompts (`bass`, `drums`), confirming that DiT-style conditioning lacks explicit cross-attention for AIBA to capture.