

RAG-Audio: Retrieval-Augmented Generation for Faithful Brain-to-Audio Reconstruction

Ambuj Mehrish
CVML Lab
Ca' Foscari University of Venice
ambuj.mehrish@unive.it

Sebastiano Vascon
CVML Lab
Ca' Foscari University of Venice
sebastiano.vascon@unive.it

Abstract

Brain-to-audio reconstruction is limited by *prior domination*: when a pretrained generator is conditioned on a weak neural signal, it produces realistic but stimulus-inaccurate audio. We introduce RAG-Audio, which decodes fMRI into a semantic audio embedding, retrieves a matching real-audio exemplar, and initializes the frozen generator’s sampling trajectory from that exemplar while retaining the decoded embedding as conditioning. On Brain2Music, RAG-Audio improves 10-way stimulus identification from 0.14–0.18 for direct generation, near the 0.10 chance level, to 0.40–0.43, comparable to retrieval. It also reduces Fréchet Audio Distance by roughly an order of magnitude, from 13.49 to 1.25 for AudioLDM. RAG-Audio approaches nearest-neighbor retrieval in identification while remaining generative; its higher FAD is expected because retrieval directly replays real audio. An autoregressive negative control, which lacks an initializable latent trajectory, shows no comparable gain, attributing the improvement to trajectory initialization. These results suggest that retrieval-guided initialization can mitigate prior domination in brain-to-audio generation.

1 Introduction

Reconstructing perceptual experience from brain activity is a demanding challenge for modern generative models. In vision, recent methods can recover recognizable images from functional magnetic resonance imaging (fMRI), even though fMRI measures stimulus-related changes in blood oxygenation rather than neural activity directly (Nishimoto et al., 2011; Takagi and Nishimoto, 2023; Scotti et al., 2023; Huo et al., 2024). Audio reconstruction has received far less attention, especially for music. Most brain-to-audio systems use a two-stage pipeline: they first map the fMRI response to a semantic audio embedding and then use this embedding to guide a pretrained audio generator (Denk et al., 2023; Ciferri et al., 2025; Park et al., 2023). Brain2Music (Denk et al., 2023), for example, predicts a joint music–language embedding (Huang et al., 2022) from fMRI and supplies it to a text-to-audio model. This design appears reasonable because it separates neural decoding from waveform generation. Yet it also assumes that the generator will preserve whatever stimulus information survives the decoding stage.

Our results suggest that this assumption often fails. We refer to this failure as *prior domination*: when a powerful pretrained generator receives a weak or noisy brain-derived condition, its learned distribution over plausible audio can outweigh the evidence carried by that condition. This is the conditioning-strength

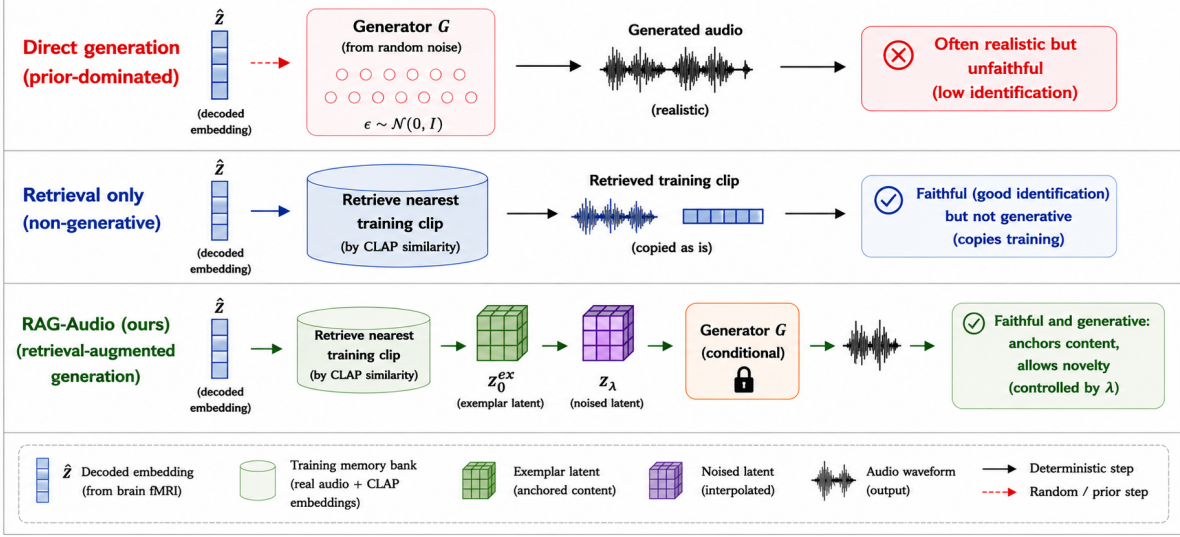


Figure 1: Overview of RAG-Audio and two baselines. Given a brain-decoded CLAP embedding $\hat{z}$, direct generation samples from noise and may be dominated by the generator prior; retrieval returns the nearest training clip but is non-generative. RAG-Audio initializes the frozen generator from the exemplar latent z_0 perturbed to an intermediate time, z_{t_0} , preserving stimulus-related structure while the anchoring strength s (with $t_0 = sT$) controls the trade-off between faithfulness and novelty.

problem familiar from diffusion guidance, where a guidance weight balances a conditional signal against the unconditional prior (Dhariwal and Nichol, 2021; Ho and Salimans, 2022); here the brain-derived condition is fixed and weak, so no reweighting of a text prompt can recover it. The generated audio may sound realistic but still fail to match the music the participant actually heard. This gap is clear on the public Brain2Music dataset (Denk et al., 2023), which contains fMRI recordings from five subjects listening to GTZAN music clips (Tzanetakis and Cook, 2002). In our re-implementation, decoding fMRI into a CLAP embedding identifies the heard clip with 0.43 accuracy in a 10-way test, well above the 0.10 chance level. However, when the same embedding is passed to AudioLDM (Liu et al., 2023a), TangoFlux (Hung et al., 2024), or MusicGen (Copet et al., 2023), identification accuracy for the generated audio falls to 0.14–0.18, including in the published Brain2Music generation setting (Denk et al., 2023). The conditioning signal therefore contains recoverable stimulus information, but direct generation retains little of it. Brain-to-image studies have reported a related qualitative pattern, with realistic reconstructions that drift from the target (Takagi and Nishimoto, 2023; Scotti et al., 2024) (Fig. 7), although it remains unclear whether the same behavior can be measured systematically in audio or whether it depends on a particular generator.

We investigate a targeted way to reduce prior domination (Figure 1). Instead of asking the generator to reconstruct the stimulus from the decoded embedding alone, we first retrieve the training audio clip whose CLAP embedding is closest to the decoded representation. We then use this clip to initialize the generator at an intermediate point in its sampling trajectory. For a diffusion-based generator such as AudioLDM (Liu et al., 2023a), we follow the SDEdit principle (Meng et al.) by adding partial noise to the exemplar latent before reverse diffusion. For a flow-based generator such as TangoFlux (Hung et al., 2024), we use the corresponding deterministic interpolation along the rectified-flow trajectory (Liu et al., 2023b; Lipman et al.; Rout et al., 2024; Kulikov et al., 2025). The generator can therefore refine the retrieved audio rather than synthesize entirely from its prior. The anchoring strength controls how closely the output follows the exemplar: weaker perturbation preserves more of it, while stronger

perturbation allows greater variation and novelty. This approach differs from plain retrieval because the final output does not simply copy the stored clip. It also differs from direct brain-conditioned generation because sampling begins from stimulus-related audio structure. We therefore view exemplar anchoring as a tunable trade-off between faithfulness and novelty, rather than as a replacement for retrieval or unconstrained generation (Lewis et al., 2020; Blattmann et al., 2022; Yuan et al., 2024).

These design choices yield a consistent improvement: anchoring recovers identification to the level of a strong non-generative retrieval baseline (Ciferri et al., 2025) while keeping the output newly generated, and an autoregressive negative control (MusicGen (Copet et al., 2023)), which exposes no initializable trajectory, shows no comparable gain isolating trajectory initialization rather than retrieval as the mechanism. Section 6 reports the full results. In summary, our contributions are:

- We identify and quantify *prior domination*, a failure mode the decode-then-generate recipe inherits and that sharpens as pretrained generators grow stronger.
- We introduce exemplar anchoring, a retrieval-augmented scheme that initializes the generator partway along its sampling trajectory from a retrieved real-audio exemplar, recovering retrieval-level identification while keeping generation genuinely generative and reducing FAD by $\sim 10.8\times$ for AudioLDM and $\sim 3.3\times$ for TangoFlux.
- We use the autoregressive generator MusicGen (Copet et al., 2023) as a negative control, attributing the recovery to latent-trajectory initialization rather than to retrieval alone.
- We provide an anatomical validity check that localizes the decoder signal to auditory cortex.

2 Related Work

Brain-to-audio reconstruction. Perceived sound has been decoded from ECoG, MEG, EEG, and fMRI, spanning speech, music, and semantic language (Anumanchipalli et al., 2019; Metzger et al., 2023; Bellier et al., 2023; Défossez et al., 2023; Postolache et al., 2025; Tang et al., 2023). For fMRI music, most methods decode an audio representation and condition a pretrained generator (Denk et al., 2023; Park et al., 2023; Liu et al., 2024). Existing systems thus range from nearest-neighbour retrieval (Ferrante et al., 2024) to unconstrained generation, including prior-guided diffusion (Ciferri et al., 2025). We re-implement both endpoints in a common evaluation framework and bridge them through exemplar anchoring.

Prior domination and cortical localization. Brain-to-image reconstruction exhibits a closely related tension. Latent-diffusion decoders reconstruct images from fMRI on the Natural Scenes Dataset (Allen et al., 2022) by mapping activity into a generator’s conditioning space (Takagi and Nishimoto, 2023; Chen et al., 2023), and the MindEye family separates a retrieval pathway from a diffusion-prior reconstruction pathway (Scotti et al., 2023, 2024); a recurring observation is that outputs look realistic yet drift from the stimulus when the brain signal is weak relative to the generator prior. We term this regime prior domination and, unlike the largely qualitative image-domain reports, quantify it for audio and show it holds across generator families. Our decoder-localization analysis follows voxelwise modeling of auditory cortex, where deep-network features predict responses along the auditory hierarchy and music-selective populations occupy superior temporal cortex (Kell et al., 2018; Norman-Haignere et al., 2015; Tuckute et al., 2023).

Audio generators and embeddings. Brain-to-image reconstruction faces a similar trade-off. Diffusion-based fMRI decoders can produce realistic images but often drift from the stimulus when the generator prior overwhelms weak neural evidence (Allen et al., 2022; Takagi and Nishimoto, 2023; Chen et al., 2023; Scotti et al., 2023, 2024). We quantify this *prior domination* for audio across generator families and localize decoder contributions using established models of auditory-cortex organization (Kell et al., 2018; Norman-Haignere et al., 2015; Tuckute et al., 2023).

Editing and retrieval-augmented generation. SDEdit starts from a guide signal, adds noise up to a chosen intermediate diffusion step, and then denoises from that point to produce a new sample (Meng et al.). The noise level controls the balance between preserving the guide and allowing the generator to introduce new content (Ho et al., 2020; Song et al., 2020b; Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Rombach et al., 2022). Related methods guide frozen diffusion models with reference images (Choi et al., 2021; Lugmayr et al., 2022), while retrieval-augmented generators condition on external examples across language, image, and audio synthesis (Khandelwal et al., 2019; Lewis et al., 2020; Borgeaud et al., 2022; Blattmann et al., 2022; Yuan et al., 2024). RAG-Audio combines these ideas by initializing generation from a retrieved audio exemplar, preserving stimulus-relevant content without copying the retrieved clip verbatim.

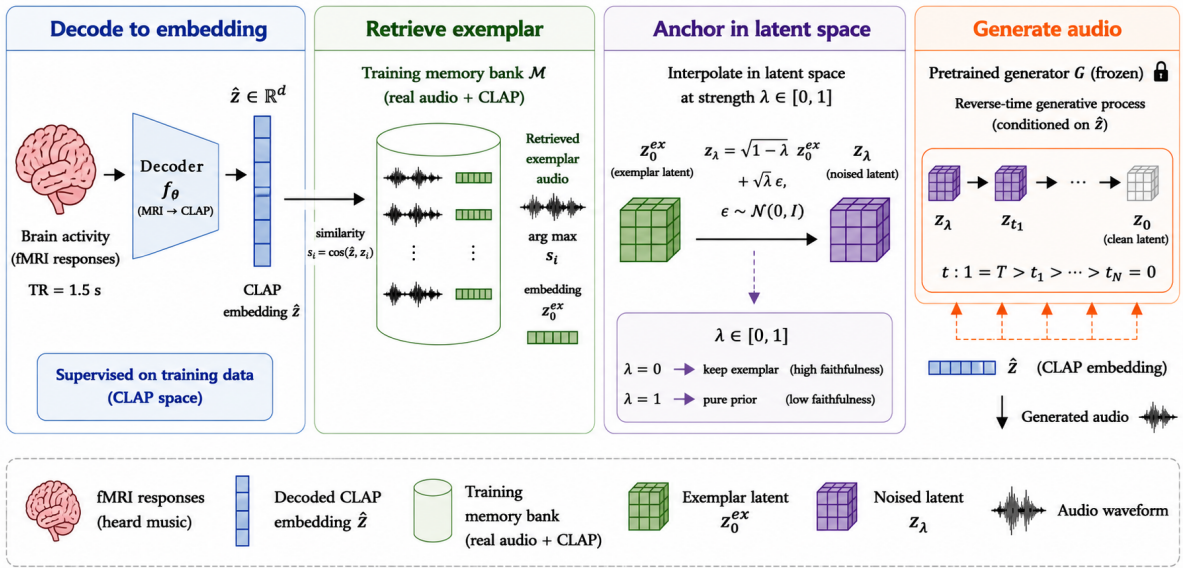


Figure 2: The RAG-Audio pipeline. The decoder f_θ maps an fMRI response x to a CLAP embedding $\hat{z}$; the nearest exemplar a^* is retrieved and encoded to z_0 , perturbed to an intermediate time $t_0 = sT(z_{t_0})$, and the frozen generator denoises from z_{t_0} conditioned on $\hat{z}$ to produce $\hat{a}$. The strength s trades faithfulness against novelty.

3 Problem Setup

Let a denote an audio clip heard by a subject while functional magnetic resonance imaging (fMRI) records the corresponding brain response x . Given a paired dataset $\mathcal{D} = (x_i, a_i)_{i=1}^M$, the goal is to reconstruct the stimulus associated with a held-out scan. Since fMRI is an indirect and temporally coarse measurement of neural activity, we formulate reconstruction in a semantic audio space rather than at the waveform level. Each stimulus is represented using CLAP (Wu et al., 2023):

$$z_a = \text{CLAP}(a) \in \mathbb{R}^{512}. \quad (1)$$

Following the standard decode-then-generate pipeline (Denk et al., 2023), a learned decoder f maps the fMRI response to a predicted semantic embedding:

$$\hat{z} = f(x). \quad (2)$$

A frozen pretrained audio generator G then synthesizes the reconstruction:

$$\hat{a} = G(\hat{z}). \quad (3)$$

We defer the architectures and training procedures for f and G to the Method section.

We measure semantic faithfulness using N -way identification. Let a^+ be the true stimulus, and let $\mathcal{A}^- = a_1^-, \dots, a_{N-1}^-$ contain $N - 1$ distractors. Writing $s(\cdot, \cdot)$ for cosine similarity, the identification accuracy is

$$\begin{aligned} \text{acc}_N = \mathbb{E} \left[\mathbb{I} \left(s(\text{CLAP}(\hat{a}), \text{CLAP}(a^+)) \right. \right. \\ \left. \left. > \max_{a^- \in \mathcal{A}^-} s(\text{CLAP}(\hat{a}), \text{CLAP}(a^-)) \right) \right]. \end{aligned} \quad (4)$$

Chance performance is $1/N$. This metric evaluates whether the reconstruction preserves the semantic content of the heard stimulus; it does not measure sample-accurate waveform recovery.

The two stages can contribute unequally to the final output. The generator G represents a strong prior $p(a)$ learned from large audio corpora, whereas the brain-derived condition $\hat{z}$ is comparatively weak and noisy. We use *prior domination* to describe the regime in which $G(\hat{z})$ remains acoustically plausible but depends only weakly on $\hat{z}$, behaving approximately like an unconditional sample from the generator prior. Generation may therefore discard stimulus information retained by the decoded embedding. Our objective is to preserve the realism supplied by G while improving the semantic faithfulness of its output.

4 Method

We propose *RAG-Audio*, a retrieval-augmented pipeline for brain-to-audio reconstruction (Fig. 2). RAG-Audio first decodes an fMRI response into a CLAP embedding, retrieves the nearest real training exemplar from a memory bank, and then initializes a frozen generator’s sampling trajectory from that exemplar at an intermediate time. For AudioLDM, this follows the SDEdit principle of partially noising the exemplar latent before reverse diffusion (Meng et al.), while TangoFlux (Hung et al., 2024) applies the corresponding deterministic interpolation along its rectified-flow trajectory. The retrieved clip supplies plausible audio structure that can survive the generator’s prior, turning prior domination into a controllable trade-off between faithfulness to the retrieved content and generative novelty.

4.1 fMRI-to-CLAP Decoder

For each stimulus audio clip a , we use its CLAP representation (Wu et al., 2023) as the decoding target, with $z_a = \text{CLAP}(a) \in \mathbb{R}^{512}$. We apply rigid motion correction with ANTsPy (Avants et al., 2011) and

retain each subject’s data in native space. To account for the haemodynamic response, we use a lag of $l = 3$ repetition times (TRs), corresponding to 4.5 seconds, and a temporal window of $w = 8$ TRs.

We rank voxels according to their association with the CLAP target. For voxel v , let x_v denote its temporal response and let $z^{(d)}$ denote the d -th CLAP dimension. We define the relevance score as

$$s_v = \max_{d \in \{1, \dots, 512\}} |\text{corr}(x_v, z^{(d)})|. \quad (5)$$

We retain the $K = 2000$ highest-scoring voxels. The same scores define the localization map analyzed in Figure 3.

The decoder $f_\theta : \mathbb{R}^{2000} \rightarrow \mathbb{R}^{512}$ maps the selected fMRI features to a predicted embedding $\hat{z} = f_\theta(x)$. Rather than minimizing dimension-wise reconstruction error, we train f_θ with InfoNCE (Oord et al., 2018) so that decoded scans remain close to their paired audio embeddings and separated from other clips in the batch. For a batch of size B and temperature $\tau = 0.07$, the objective is

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\cos(f_\theta(x_i), z_{a_i})/\tau)}{\sum_{j=1}^B \exp(\cos(f_\theta(x_i), z_{a_j})/\tau)}. \quad (6)$$

This objective directly optimizes the retrieval geometry used for semantic identification. We compare it with ridge, linear, and MLP decoders in Figure 8; the contrastive decoder is selected because its training criterion matches the downstream identification objective.

4.2 Memory Bank and Retrieval

We construct a memory bank $\mathcal{B} = \{z_a : a \in \mathcal{A}_{\text{train}}\}$ from CLAP embeddings of real training-set audio, where $\mathcal{A}_{\text{train}}$ denotes the training clips. Test stimuli are excluded. Each bank entry retains a reference to its corresponding audio clip.

Given a held-out fMRI response x , the decoder produces $\hat{z} = f_\theta(x)$. We retrieve the training exemplar whose CLAP embedding has the largest cosine similarity to $\hat{z}$:

$$a^* = \arg \max_{a \in \mathcal{B}} \cos(\hat{z}, z_a). \quad (7)$$

Here, $a \in \mathcal{B}$ indexes the audio clip associated with a bank embedding. Returning a^* directly defines the RETRIEVAL baseline, corresponding to the published linear-contrastive fMRI-to-CLAP retrieval setting (Ferrante et al., 2024). This baseline provides a strong non-generative reference: it can preserve stimulus-level content, but its output is a verbatim training clip. RAG-Audio instead treats a^* as an anchor from which to generate a new reconstruction.

4.3 Exemplar-Anchored Generation via Intermediate-Time Initialization

Let E denote the encoder of a frozen latent generative model. We first map the retrieved exemplar to the generator latent, $z_0 = E(a^*)$. A anchoring strength $s \in (0, 1]$ determines the intermediate initialization time $t_0 = sT$, where T is the full denoising horizon.

For a latent-diffusion generator such as AudioLDM (Liu et al., 2023a), the anchoring strength $s \in (0, 1]$ determines the diffusion initialization step

$$t_0 = \lfloor sT \rfloor, \quad (8)$$

where T denotes the total number of diffusion steps. Following the SDEdit principle (Meng et al.), we partially corrupt the exemplar latent z_0 according to the forward diffusion process and then run the reverse process from t_0 , conditioned on the decoded embedding $\hat{z}$:

$$z_{t_0} = \sqrt{\bar{\alpha}_{t_0}} z_0 + \sqrt{1 - \bar{\alpha}_{t_0}} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \quad (9)$$

Flow based generator such as TangoFlux (Hung et al., 2024) follows a normalized rectified-flow trajectory. For this model, the anchoring strength directly determines the initialization time:

$$\tau_0 = s. \quad (10)$$

We initialize the trajectory by deterministically interpolating between the exemplar latent and Gaussian noise (Song et al., 2020a):

$$z_{\tau_0} = (1 - \tau_0) z_0 + \tau_0 \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (11)$$

and then integrate the learned flow from τ_0 to 0.

The strength s controls how much exemplar structure remains at the start of generation. As $s \rightarrow 0$, the output approaches a^* , favoring faithfulness at the cost of novelty. As $s \rightarrow 1$, the initialization increasingly erases the exemplar and approaches direct generation from the model prior. Sweeping s therefore traces the faithfulness–novelty continuum shown in Figure 4; empirically, we consider the working range $s \approx 0.25$ – 0.40 .

This mechanism requires a continuous latent trajectory that can be initialized through Equation (9). Autoregressive MusicGen (Copet et al., 2023) does not provide such a latent. Its retrieved exemplar can enter only through melody conditioning, which remains subject to the autoregressive token prior. We use this case as a negative control in the Results section to distinguish the effect of intermediate-time trajectory initialization from the mere availability of an exemplar.

Table 1: Main comparison on Brain2Music ($n=300$; 10-way chance 0.10), all methods run in a single harness. Exemplar anchoring lifts identification from 0.14–0.18 (direct) to the retrieval level (0.40–0.43) and cuts FAD by up to $\sim 10\times$ for AudioLDM and TangoFlux, but not for autoregressive MusicGen.

Method	Type	Gen	10-way $\uparrow$	FAD $\downarrow$
<i>Published baselines (re-implemented)</i>				
R&B (Ferrante et al., 2024)	retrieval	no	0.40	0.89
MusicGen (Copet et al., 2023)(decode $\rightarrow$ gen)	direct	yes	0.18	8.06
<i>Direct generation, prior-only (ours)</i>				
AudioLDM (Liu et al., 2023a)	direct	yes	0.14	13.5
TangoFlux (Hung et al., 2024)	direct	yes	0.14	7.89
MusicGen (Copet et al., 2023)	direct	yes	0.18	8.06
<i>RAG-Audio (ours)</i>				
AudioLDM + RAG	retr.-aug.	yes	0.43	1.25
TangoFlux + RAG	retr.-aug.	yes	0.40	2.36
MusicGen + RAG	retr.-aug.	yes	0.20	6.21

5 Experimental Setup

Dataset. We evaluate on Brain2Music, released as OpenNeuro ds003720 (Nakai et al., 2022; Denk et al., 2023). The dataset contains fMRI recordings from five subjects listening to 15-second music clips

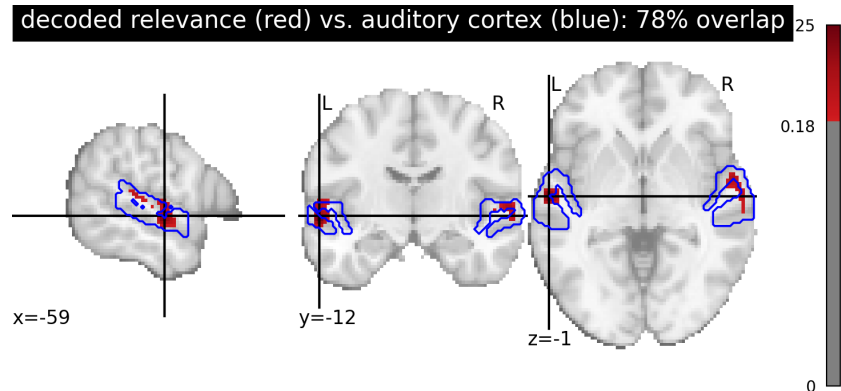


Figure 3: Decoder localization: voxel relevance scores (red) overlap the Harvard–Oxford auditory-cortex mask (blue), with 78% of top voxels in bilateral superior temporal gyrus.

drawn from the ten GTZAN genres (Tzanetakis and Cook, 2002), with a repetition time of $TR = 1.5$, s. We follow the official subject-wise split of 480 training clips and 60 held-out test clips. All metrics are averaged across subjects, corresponding to a pooled set of $n = 300$ test generations.

Audio generators. We consider three pretrained generators: AudioLDM (cvssp/audioldm-s-full-v2) (Liu et al., 2023a), TangoFlux (declare-lab/TangoFlux) (Hung et al., 2024), and MusicGen (facebook/musicgen-melody) (Copet et al., 2023). All generator parameters remain frozen. AudioLDM and TangoFlux use latent-diffusion and rectified-flow formulations, respectively, and therefore support intermediate-time trajectory initialization. MusicGen is autoregressive and serves as a mechanism control for which exemplar initialization is unavailable.

Comparison arms. We compare three reconstruction strategies. *Direct* performs brain-conditioned generation using only the pretrained generative prior. *Retrieval* returns the nearest real clip from the training bank (Ferrante et al., 2024). *RAG-Audio* applies our intermediate-time exemplar-anchoring procedure. All methods, including published baselines, are run within a single evaluation harness using the same test scans, candidate sets, checkpoints, and metric implementations. Their reported values are therefore directly comparable.

Faithfulness. Semantic faithfulness is measured by the N -way identification of Eq. 4, computed in CLAP space using the LAION checkpoint laion/clap-htsat-unfused. We report $N \in \{2, 5, 10, 50\}$, with corresponding chance levels of 0.50, 0.20, 0.10, and 0.02.

Realism and novelty. Audio realism is evaluated using FAD computed with VGGish features (Kilgour et al., 2019; Hershey et al., 2017), together with FD and KL divergence computed from the PANNs Cnn14 audio tagger (Kong et al., 2020). Lower values indicate better agreement with real audio. We quantify novelty as one minus the cosine similarity between each generated clip and its nearest clip in the training bank. Plain retrieval has novelty near zero by construction because it returns a training clip verbatim. We additionally report pairwise diversity among generated samples.

Implementation. We apply rigid-motion correction with ANTsPy (Avants et al., 2011), select the top 2000 voxels, sweep the anchoring strength over $\{0.2, \dots, 0.5\}$, and train the contrastive decoder with

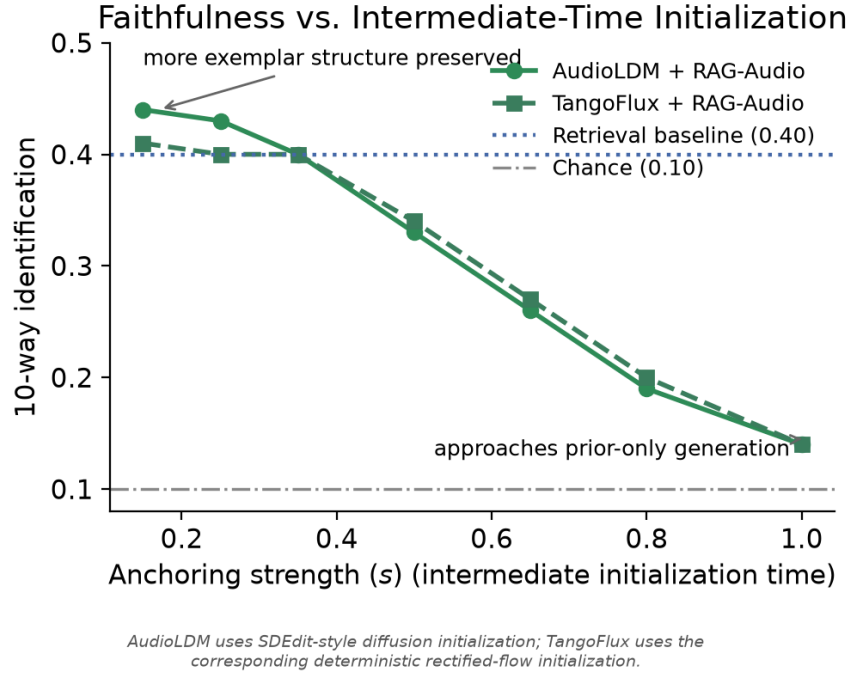


Figure 4: Ten-way identification versus anchoring strength s : low s stays near the retrieval bound, high s decays to the 0.10 chance level; working range $s \approx 0.25$ – 0.40 .

temperature $\tau = 0.07$. Complete preprocessing, optimization, and generation settings are provided in Appendix C.

6 Results

Decoder Performance The contrastive fMRI-to-CLAP decoder achieves the strongest stimulus identification despite the lowest dimension-wise embedding correlation. Averaged across five subjects, it reaches 0.43 accuracy in 10-way identification (chance 0.10) and 0.80 in 2-way identification, with every subject above 10-way chance. It outperforms ridge, linear, and MLP decoders even though its Pearson correlation is only 0.20, versus 0.68–0.71 for regression-based alternatives. This contrast likely reflects the objectives: InfoNCE (Oord et al., 2018) directly preserves the relative geometry required

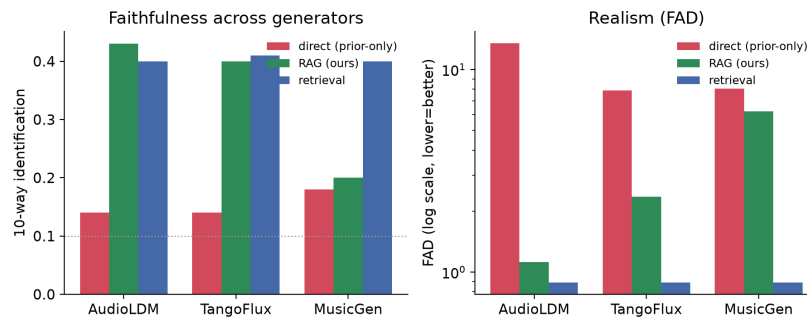


Figure 5: Ten-way identification and FAD per generator: anchoring restores AudioLDM and TangoFlux to retrieval level but not autoregressive MusicGen.

for identification, whereas correlation and mean-squared error emphasize dimension-wise agreement. Per-subject results, peak relevance, and the full decoder comparison appear in Tables 3 and 4 and Fig. 8. The selected voxels also exhibit an anatomically plausible distribution. Figure 3 shows that the relevance

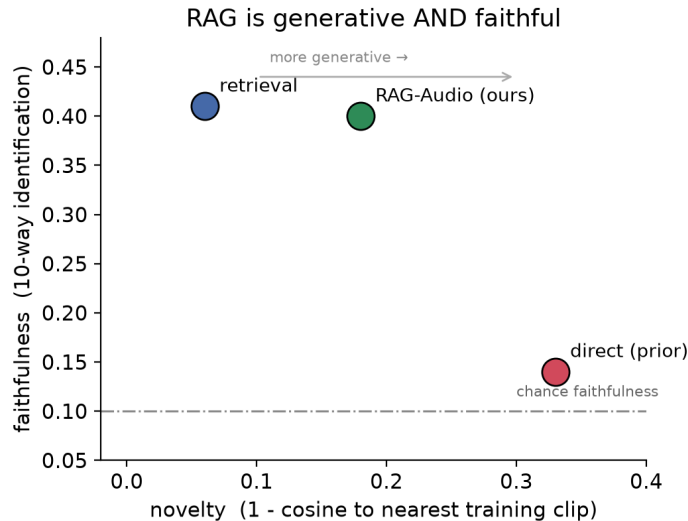


Figure 6: Faithfulness (identification) versus novelty: direct is novel but unfaithful, retrieval faithful but verbatim, RAG-Audio faithful and generative.

scores concentrate in the bilateral superior temporal gyrus, a region associated with auditory processing. Under the Harvard–Oxford atlas (Desikan et al., 2006), 78% of the top-ranked voxels fall within auditory cortex across the two hemispheres. This localization supports the interpretation that the decoder relies primarily on stimulus-related auditory activity rather than scanner artefacts or head motion.

Prior Domination Direct brain-conditioned generation discards much of the stimulus information recovered by the decoder. Although the decoded embedding reaches 0.43 identification accuracy, generation reduces performance to 0.14 with AudioLDM, 0.14 with TangoFlux, and 0.18 with MusicGen, only modestly above the 0.10 chance level. Table 1 reports the full comparison, with the complete N -way breakdown in Table 2. The MusicGen result re-implements the decode-then-generate formulation of (Denk et al., 2023). The degradation is not confined to one generator family. It appears in latent diffusion, rectified flow, and autoregressive generation, while direct-generation FAD ranges from 7.89 to 13.50. Taken together, these results are consistent with prior domination as a property of the decode-then-generate recipe: the decoded condition remains informative, but the pretrained generator produces audio governed largely by its own prior.

Exemplar Anchoring Exemplar anchoring restores identification to the level of the retrieval reference for latent-diffusion generators. RAG-Audio reaches 0.43 with AudioLDM and 0.40 with TangoFlux, compared with 0.14 for their directly conditioned counterparts and 0.10 chance. These values approximately match both the 0.40 retrieval baseline and the 0.43 decoder-level identification score, corresponding to an approximately $3\times$ improvement over direct generation. The same intervention reduces AudioLDM FAD from 13.50 to 1.25, an approximately order-of-magnitude change, and TangoFlux FAD from 7.89 to 2.36 (Table 1); the same reductions hold for FD and KL (Table 5). This comparison should not be read as RAG-Audio outperforming retrieval. Retrieval directly returns real training audio and is therefore the

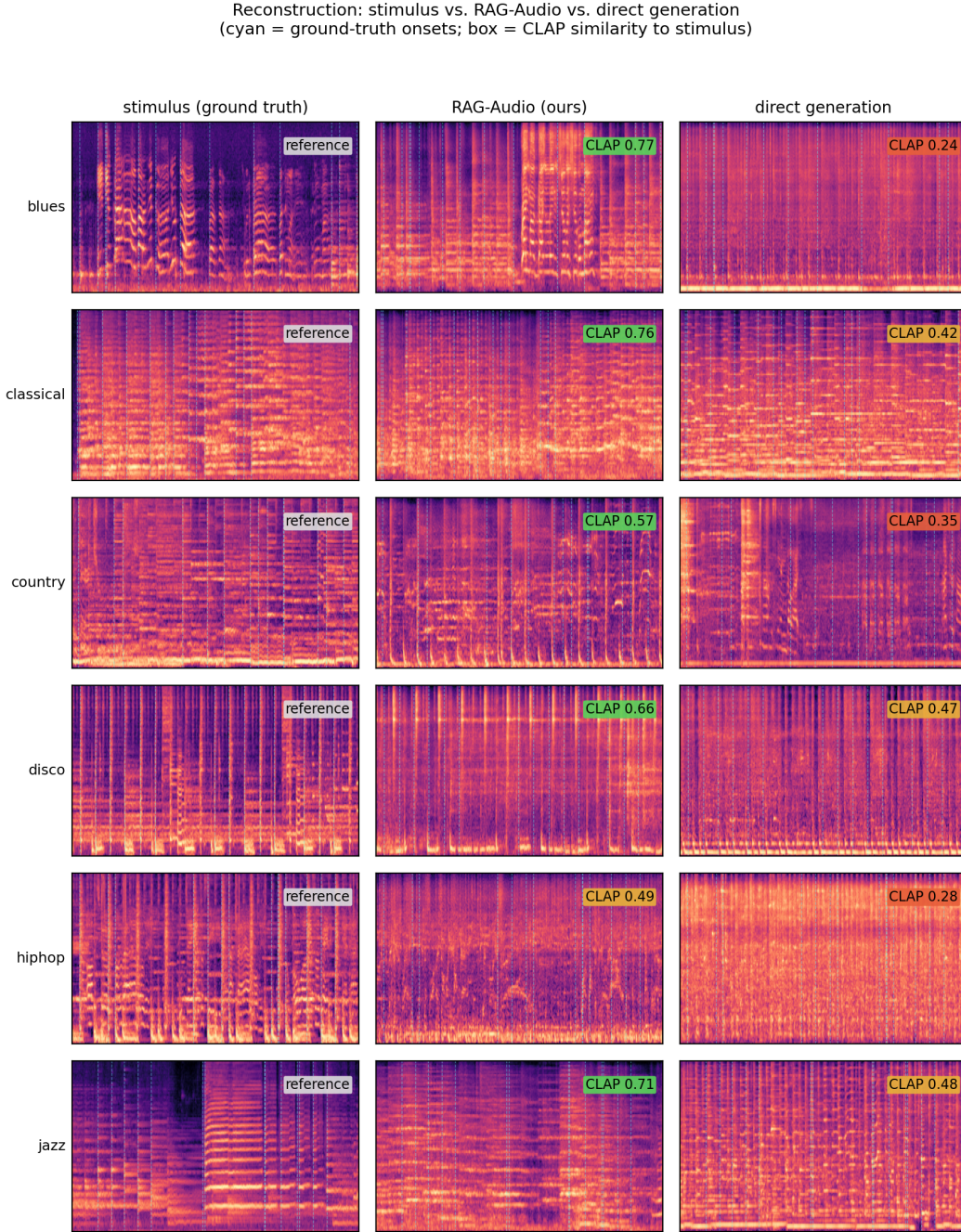


Figure 7: Per-genre reconstructions: stimulus, RAG-Audio, and direct generation (log-mel spectrograms). Cyan lines are stimulus onsets (shared per row); the corner box is CLAP similarity to the stimulus (green/orange/red). Direct generation is realistic but off-grid and low-similarity; RAG-Audio recovers onset and harmonic structure (on-grid, high) without copying any retrieved clip. Structural, not sample-accurate, closeness.

strongest non-generative reference, with an FAD of 0.89. RAG-Audio does not surpass this value, nor do we claim that it does. Its advantage is instead that it matches retrieval-level identification while producing an edited sample rather than returning the retrieved clip verbatim. Figure 6 reflects this distinction:

novelty increases from approximately 0.06 for retrieval to 0.18 for RAG-Audio while identification remains at the retrieval level (Table 6). A genre-level breakdown of the residual errors (Appendix G, Table 8) shows they remain musically coherent, concentrating among acoustically similar GTZAN genres (Tzanetakis and Cook, 2002).

The anchoring strength controls this balance. As shown in Fig. 4 (full sweep in Table 7), identification decreases smoothly from the retrieval regime toward the 0.10 chance level as the strength approaches 1, while novelty increases as more exemplar structure is removed. The range $s \approx 0.25\text{--}0.40$ provides the most useful operating region in our experiments, retaining stimulus-related structure without reducing generation to direct retrieval. The anchoring advantage is also stable across memory-bank sizes (Appendix F, Table 8).

Mechanism Control The autoregressive negative control indicates that exemplar availability alone is insufficient. On MusicGen, adding the retrieved exemplar changes identification only from 0.18 to 0.20, in contrast to the approximately $3\times$ gains observed for AudioLDM and TangoFlux in Fig. 5. MusicGen has no continuous latent trajectory that admits intermediate-time initialization; the exemplar can enter only through melody conditioning, which remains governed by the autoregressive token prior.

FAD nevertheless improves modestly, from 8.06 to 6.21. This pattern suggests that the exemplar can nudge acoustic realism without restoring stimulus identity. The cross-generator comparison therefore points to intermediate-time trajectory initialization, rather than the mere presence of a retrieved example, as the mechanism responsible for the faithfulness gains.

Qualitative reconstructions. Figure 7 makes prior domination and its mitigation directly visible. Direct brain-conditioned generation produces spectrograms with plausible spectrotemporal texture, yet their energy drifts off the stimulus onset grid (cyan) and their CLAP similarity to the stimulus stays low across all six examples (red/orange boxes). The failure is thus easy to miss from realism or casual listening alone precisely the regime we call prior domination. Exemplar anchoring recovers the structure the prior discards: RAG-Audio reconstructions align with the onset grid and reproduce the harmonic banding of the stimulus, and the CLAP-similarity boxes rise correspondingly (green), roughly doubling over direct generation across the shown genres. At the same time, the RAG panels differ in fine detail from both the stimulus and any single retrieved clip, consistent with the novelty results in Table 6 and confirming that anchoring edits rather than copies.

7 Discussion

Our results support a simple account of when and why brain-to-audio generation fails. A pretrained generator encodes a strong prior over natural audio, whereas the brain-derived condition is weak and noisy; combined through direct conditioning, the prior dominates and the output drifts toward a fluent but stimulus-agnostic sample. Anchoring the sampling trajectory on a retrieved real-audio exemplar tethers the output to stimulus-related structure, converting prior domination into a controllable faithfulness-novelty trade-off governed by the anchoring strength.

This mitigation is deliberately generator-specific, and the specificity is informative rather than incidental. Because exemplar anchoring acts by initializing a continuous latent trajectory, it applies to latent-diffusion and rectified-flow generators but not to autoregressive models, which expose no such trajectory. The MusicGen negative control makes this concrete: supplying the identical exemplar raises 10-way identification only marginally, from 0.18 to 0.20, whereas the same exemplar restores

AudioLDM and TangoFlux to retrieval-level accuracy (0.43 and 0.40). Trajectory initialization, not the mere availability of a retrieved clip, is therefore the operative mechanism.

We are precise about what RAG-Audio does and does not achieve. It does not surpass plain retrieval on FAD, and it cannot: retrieval returns real training audio, which is FAD-optimal by construction (0.89). RAG-Audio’s contribution is instead to match retrieval-level identification while emitting a newly generated sample rather than replaying a stored clip, cutting FAD by roughly an order of magnitude relative to direct generation (e.g., AudioLDM $13.5 \rightarrow 1.25$). Its niche is exactly the setting that retrieval cannot serve: where novel audio is required and a verbatim training clip is unacceptable, yet faithfulness to the heard stimulus must still be preserved.

8 Limitations

Several limitations bound our claims. The study covers a single dataset and a single modality: music decoded from fMRI so transfer to speech or environmental sound remains untested. The memory bank is bounded by the training stimuli; RAG-Audio anchors to, and therefore cannot reconstruct content outside, that support. Exemplar anchoring requires an initializable latent trajectory and thus does not extend to autoregressive generators. Our N -way identification measures semantic agreement in CLAP space rather than sample-accurate waveform recovery, so the spectrograms should not be read as evidence of exact reconstruction. Finally, the anchoring strength s is a hyperparameter: we report a working range ($s \approx 0.25\text{--}0.40$) rather than a single optimum, and the best operating point may differ across generators and datasets.

9 Conclusion

We identified *prior domination* as a general failure mode of the decode-then-generate recipe for brain-to-audio reconstruction: a decoded embedding that is itself identifiable (0.43 at 10-way, against 0.10 chance) can nonetheless collapse to near-chance reconstructions (0.14–0.18) once a strong generator is applied. We introduced exemplar anchoring, initializing the generator’s sampling trajectory from a retrieved real-audio clip, via SEdit-style noising for diffusion models and the analogous rectified-flow interpolation for flow models and showed that it restores retrieval-level identification while keeping generation genuinely generative and reducing FAD by roughly an order of magnitude. An autoregressive negative control localizes the effect to latent-trajectory initialization rather than retrieval alone. Beyond the method, our single-harness comparison turns a failure mode previously reported only qualitatively in brain-to-image work into a quantitative, generator-general characterization one likely to sharpen as pretrained generators continue to improve.

Acknowledgements

This work was supported by the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101205348 (CASPER). We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium, through the EuroHPC AI Factories "AI for Science and Collaborative EU Projects" Access call (proposal No. EHPC-AIF-2026SC01-041). We further acknowledge the CINECA award under the ISCRA initiative (Class C project IsCd5_CASPER-A), for the availability of high performance computing resources and support.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

- Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.
- Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.
- Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.
- Ludovic Bellier, Anaïs Llorens, Déborah Marciano, Aysegul Gunduz, Gerwin Schalk, Peter Brunner, and Robert T Knight. Music can be reconstructed from human auditory cortex activity using nonlinear decoding models. *PLoS biology*, 21(8):e3002176, 2023.
- Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *arXiv preprint arXiv:2204.11824*, 2022.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.
- Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22710–22720. IEEE Computer Society, 2023.
- Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. arXiv:2108.02938.
- Matteo Ciferri, Matteo Ferrante, and Nicola Toschi. Reconstructing music perception from brain activity using a prior guided diffusion model. *Scientific Reports*, 15(1):42108, 2025.
- Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023.
- Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, 2023.

- Timo I Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. Brain2music: Reconstructing music from human brain activity. *arXiv e-prints*, pages arXiv–2307, 2023.
- Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.
- Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- Matteo Ferrante, Matteo Ciferri, and Nicola Toschi. R&b–rhythm and brain: Cross-subject decoding of music from human brain activity. *arXiv preprint arXiv:2406.15537*, 2024.
- Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022.
- Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.
- Jingyang Huo, Yikai Wang, Xuelin Qian, Yun Wang, Chong Li, Jianfeng Feng, and Yanwei Fu. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. *arXiv preprint arXiv:2403.18211*, 2024.
- Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.
- Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.
- K Kilgour et al. Frechet audio distance: A reference-free metric for evaluating music enhancem. *InterSpeech*, 2019.
- Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

- Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*.
- Che Liu, Changde Du, Xiaoyu Chen, and Huiguang He. Reverse the auditory processing pathway: Coarse-to-fine audio reconstruction from fmri. *arXiv preprint arXiv:2405.18726*, 2024.
- 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 *International Conference on Machine Learning*, pages 21450–21474. PMLR, 2023a.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023b.
- Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.
- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.
- Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.
- Tomoya Nakai, Naoko Koide-Majima, and Shinji Nishimoto. Music genre neuroimaging dataset. *Data in Brief*, 40:107675, 2022.
- Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.
- Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, 88(6):1281–1296, 2015.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Jong-Yun Park, Mitsuaki Tsukamoto, Misato Tanaka, and Yukiyasu Kamitani. Sound reconstruction from human brain activity via a generative model with brain-like auditory features. *arXiv preprint arXiv:2306.11629*, 2023.

- Emilian Postolache, Natalia Polouliakh, Hiroaki Kitano, Akima Connelly, Emanuele Rodolà, Luca Cosmo, and Taketo Akama. Naturalistic music decoding from eeg data via latent diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024.
- Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.
- Paul S Scotti, Mihir Tripathy, Cesare Kadir Torricco Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, pages 44038–44059, 2024.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463, 2023.
- Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, 2023.
- Greta Tuckute, Jenelle Feather, Dana Boebinger, and Josh H McDermott. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, 21(12):e3002366, 2023.
- George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Retrieval-augmented text-to-audio generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585. IEEE, 2024.

A Code and Reproducibility

For reproducibility, we release the full codebase, including preprocessing, the fMRI-to-CLAP decoder, retrieval, exemplar anchoring, and evaluation, in a public repository:

[TBA](#)

The repository documents the environment, checkpoints, and configuration files needed to reproduce every table and figure in this document.

B Use of Large Language Models

Large language models were used only to improve grammar and clarity in author-written text.

C Implementation Details

Decoder. The contrastive decoder f_θ is a multilayer perceptron mapping the 2000 selected voxels to the 512-dimensional CLAP space. It is trained for 400 epochs with the InfoNCE objective (Oord et al., 2018) at temperature $\tau = 0.07$, using the Adam optimizer with weight decay; full layer widths, learning rate, and batch size are given in our released configuration. Voxels are selected by the relevance score of Eq. 5 ($K = 2000$), with a haemodynamic lag of $\ell = 3$ TRs and a temporal window of $w = 8$ TRs. Motion correction is rigid-body, performed in native space with ANTsPy (Avants et al., 2011).

Generators. All three generators are frozen and used at their public checkpoints: AudioLDM (cvssp/audioldm-s-full v2) (Liu et al., 2023a), TangoFlux (declare-lab/TangoFlux) (Hung et al., 2024), and MusicGen (facebook/musicgen-melody) (Copet et al., 2023). For anchored generation we sweep the strength grid $s \in \{0.15, 0.25, 0.35, 0.5, 0.65, 0.8, 1.0\}$; generation step counts and guidance scales are listed in the released configuration. Identification uses the LAION CLAP checkpoint laion/clap-htsat-unfused (Wu et al., 2023), which is distinct from the CLAP variant used as the decoder target.

D Decoder Baselines

The contrastive decoder is compared against ridge, linear, and MLP regressors trained to predict the CLAP target directly. As summarized in Figure 8, the contrastive decoder attains the best 10-way identification despite the *lowest* dimension-wise correlation with the target embedding (Pearson $r = 0.20$, versus 0.68–0.71 for the regression decoders). This apparent paradox reflects the training objective: InfoNCE optimizes the relative cosine geometry that N -way identification actually measures, whereas ridge/linear/MLP minimize dimension-wise error, which need not preserve nearest-neighbor rankings. Because retrieval and identification both operate on cosine similarity, the contrastive criterion is the better match to the downstream task.

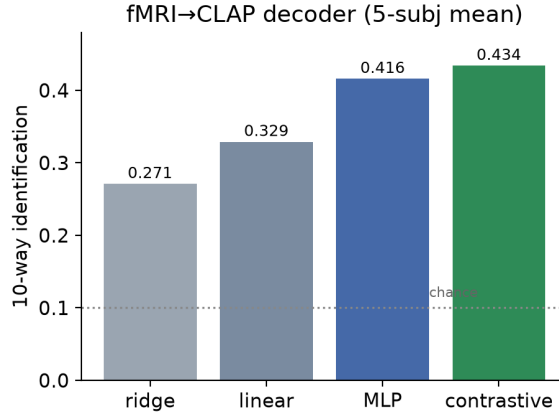


Figure 8: Decoder 10-way identification (5-subj mean).

Table 2: Identification at $N=2, 5, 10$ for every generator $\times$ arm (direct, RAG, retrieval; $n=300$, chance $1/N$). RAG matches the retrieval bound for the latent-trajectory generators but not MusicGen. †: only 10-way was aggregated for AudioLDM (Table 1).

Generator	Arm	2-way	5-way	10-way
AudioLDM	direct	†	†	0.14
	RAG	†	†	0.43
	retrieval	†	†	0.40
TangoFlux	direct	0.515	0.249	0.140
	RAG	0.785	0.555	0.403
	retrieval	0.792	0.554	0.406
MusicGen	direct	0.590	0.310	0.180
	RAG	0.615	0.336	0.201
	retrieval	0.796	0.553	0.402

E Per-Subject Results

The decoder is reliable across all five subjects. Table 3 reports per-subject identification for the contrastive decoder (60 test clips each): every subject exceeds chance at all N , with a mean 10-way accuracy of 0.434 (chance 0.10) and a best subject (sub-003) at 0.532. Table 4 reports the peak voxel relevance (max $|\text{corr}|$ with the CLAP target) per subject, ranging from 0.425 (sub-001) to 0.599 (sub-003); the ordering tracks identification accuracy, indicating that subjects with stronger auditory-cortex coupling are decoded more reliably. Bilateral superior-temporal-gyrus localization is consistent across subjects (Figure 3), and the single-subject map for sub-001 is shown in the main text.

Table 3: Per-subject identification of the contrastive decoder (60 test clips each; chance $1/N$). All five subjects exceed chance at every N , confirming the decoder is reliable across subjects.

Subject	2-way	5-way	10-way	50-way	Pearson
sub-001	0.782	0.555	0.393	0.121	0.177
sub-002	0.820	0.591	0.421	0.145	0.189
sub-003	0.836	0.651	0.532	0.206	0.234
sub-004	0.790	0.560	0.437	0.173	0.174
sub-005	0.794	0.546	0.388	0.154	0.233
mean	0.804	0.581	0.434	0.159	0.201

Table 4: Per-subject peak voxel relevance (max $|\text{corr}|$ with the CLAP target), indexing decoder strength in auditory cortex for each subject.

Subject	sub-001	sub-002	sub-003	sub-004	sub-005
peak $ \text{corr} $	0.425	0.496	0.599	0.454	0.485

Table 5: Audio realism ($n=300$; lower is better): FAD (VGGish) with FD and KL (PANNs Cnn14). RAG reduces all three for AudioLDM and TangoFlux toward the real-audio retrieval reference, while MusicGen changes little.

Generator	Arm	FAD	FD	KL
AudioLDM	direct	13.49	79.40	3.08
	RAG	1.25	17.23	1.07
	retrieval	0.89	16.40	1.05
TangoFlux	direct	7.89	59.11	2.24
	RAG	2.36	21.49	1.01
	retrieval	0.89	16.40	1.05
MusicGen	direct	8.06	52.99	1.85
	RAG	6.21	59.32	1.82
	retrieval	0.89	17.55	1.08

Table 6: Novelty ($1 - \cos$ to the nearest training clip) and pairwise diversity for the TangoFlux arms. RAG is far more novel than verbatim retrieval (≈ 0) while staying below unconstrained direct generation.

Arm	novelty	diversity
retrieval	0.06	0.42
RAG (ours)	0.18	0.39
direct (prior)	0.33	0.48

F Memory-Bank Size Ablation

G Genre Confusion

To characterize the residual errors, each generated clip is labelled by the genre of its nearest ground-truth neighbour in CLAP space (leave-one-out). Figure 9 shows the resulting confusion matrix for the RAG arm (TangoFlux): top-1 genre accuracy is 0.33, or $3.3\times$ the 0.10 chance level. Well-separated genres are recovered most often (classical 0.63, jazz 0.53, hip hop/pop 0.43), while the errors concentrate

among the acoustically overlapping GTZAN (Tzanetakis and Cook, 2002) genres (metal, rock, blues) that are confusable even for audio-only classifiers. The confusions are musically coherent (e.g., disco↔pop↔reggae, rock↔disco), indicating that reconstruction errors preserve broad timbral and rhythmic structure rather than being random.

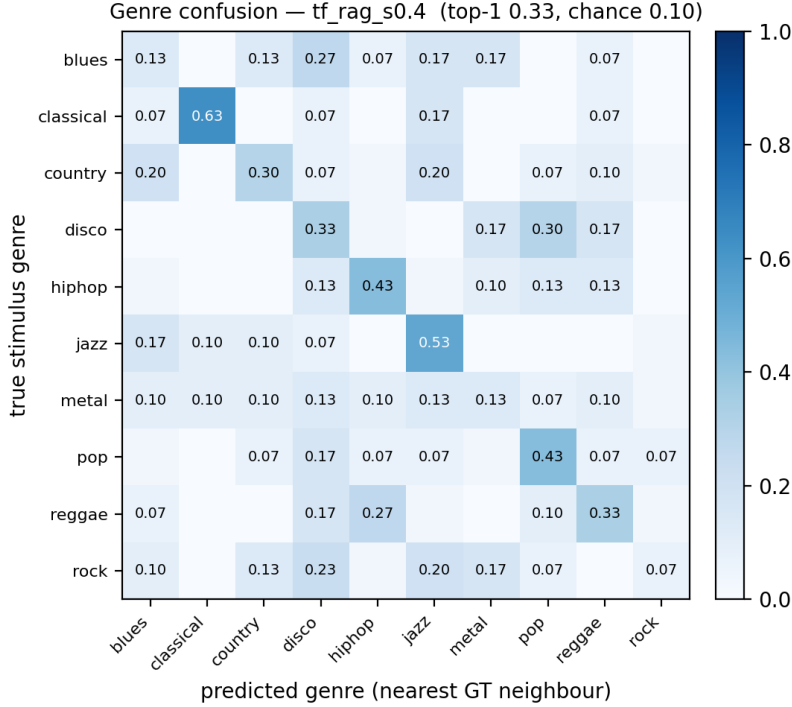


Figure 9: Genre confusion for the RAG arm (TangoFlux).

H SDEdit Strength Sweep

Table 7 details the dependence of 10-way identification on the anchoring strength s that underlies Figure 4. For AudioLDM, identification is highest at low strength (0.49 at $s = 0.2$) and decreases monotonically toward the 0.10 chance level as $s \rightarrow 1$, crossing the retrieval bound within the working range. TangoFlux was evaluated at its operating point $s = 0.4$, where it reaches 0.40. Realism is stable across the anchored range: AudioLDM FAD is 1.14, 1.12, and 1.25 at $s = 0.15$, 0.2, and 0.25 respectively, so faithfulness can be tuned over this interval without degrading audio quality. We therefore adopt $s \approx 0.25$ –0.40 as the default operating region.

Table 7: Ten-way identification versus anchoring strength s (AudioLDM sweep; TangoFlux at its operating point). Accuracy is highest at low s and decays to the 0.10 chance level as $s \rightarrow 1$; *retr.* = retrieval bound, *direct* = prior-only.

Generator	$s=0.2$	0.25	0.3	0.4	0.5	retr.	direct
AudioLDM	0.49	0.43	0.39	0.30	0.14	0.40	0.14
TangoFlux	—	—	—	0.40	—	0.40	0.14

We tested whether RAG-Audio’s advantage grows as the retrieval memory shrinks, hypothesizing

graceful degradation. The data did not support this. Table 8 report AudioLDM RAG ($s = 0.2$) against retrieval FAD at bank sizes 30, 120, and 480. At reduced bank sizes the estimates are noisy ($n = 60$ pairs versus 300 at the full bank), and RAG tracks retrieval within that noise rather than separating from it as the bank shrinks. We report this as an honest negative result: the benefit of anchoring does not increase under retrieval scarcity in this dataset.

Table 8: Memory-bank size ablation: AudioLDM RAG ($s=0.2$) vs. retrieval FAD at bank sizes 30/120/480 ($n=60$ pairs, noisy). RAG tracks retrieval within noise and does not degrade gracefully as the bank shrinks.

	bank 30	bank 120	full (480)
RAG FAD	1.43	1.88	1.12
retrieval FAD	1.90	1.28	0.89

I Dataset and Reproducibility

We use the Brain2Music fMRI dataset (Denk et al., 2023), released as OpenNeuro ds003720 (Nakai et al., 2022), obtained via DataLad (`datalad get` on the `*_bold.nii` files); the auditory stimuli are drawn from the ten GTZAN genres (Tzanetakis and Cook, 2002). We follow the official subject-wise split of 480 training and 60 test clips per subject. Preprocessing applies ANTsPy rigid motion correction in native space (Avants et al., 2011), a lag of 3 TRs and window of 8 TRs, and top-2000 voxel selection by Eq. 5. Auditory-cortex overlap is computed against the Harvard-Oxford atlas (Desikan et al., 2006). All runs use a single GPU per job; we release seeds, code, and configurations.