nano4M-Audio: Adding an Audio Modality to a 4M Multimodal Model

MELLAL Ziyad (379965), FARHAT Marc (325811), BADDOUR Hassan (378836)
COM-304 Final Project Report

***Abstract*—We ask whether a single shared 4M-style transformer can absorb *audio* as a fifth modality at academic scale. We extend nano4M (∼96M parameters) with EnCodec audio tokens and one training-side change to the 4M recipe—contiguous *span* masking for the temporal audio stream—alongside RGB, depth, surface normals, and a class caption. We self-build a 9,192-clip audio–visual dataset over 11 animal classes, cleaned by a three-stage PANNs/CLIP/Silero-VAD oracle. The pipeline is clean and the gains are concrete: depth and normal cross-entropy converge strongly, audio CE falls below its empirical marginal entropy ($5.2 < 6.2$ nats $= \sim$1 nat of conditional structure per token), and the framework cleanly predicts structural targets from RGB (RGB→depth/normal/caption). High-entropy image synthesis is the limit: caption→RGB collapses to 0% ResNet top-5—exactly like audio→RGB—and cross-modal audio↔vision generation mode-collapses with retrieval faint (R@5 up to 4.5%). We isolate a sharp *asymmetry*—discriminative audio→class structure emerges while generation collapses—and trace it to three specific causes. The diagnostic, not the generation, is the contribution: a controlled study identifying the exact levers needed to make this work at scale.**

1. Introduction

Foundation models such as 4M [1] and 4M-21 [2] unify many modalities—RGB, depth, surface normals, semantic maps, captions—by tokenizing each into a discrete sequence and training a single encoder–decoder transformer with random masking. Every modality in that family is *visual*, derived from rendered or curated imagery. A natural omission stands out: *audio*, a temporal real-world signal, is absent. We ask whether the recipe extends to it at the small-data, modest-compute budget where most academic projects operate ($\sim 10^4$ clips, $\sim 10^8$ parameters):

> **H1.** Can a 4M-style model trained with random Dirichlet masking learn *bidirectional* cross-modal alignment between a temporal, real-world modality (audio) and visual modalities, from a small paired dataset?

Audio is a deliberate stress test for the recipe. It is *temporal* rather than spatial, so the Dirichlet masking that 4M tailors to spatial token grids lets a decoder copy local context instead of learning long-range structure. It is high-frequency, sourced from noisy web video, and a strong neural codec is itself lossy, so raw codec tokens are *semantically opaque*: the model must learn class-level meaning on top of acoustic codes. We restrict the problem to a tractable, evaluable subset—**animal sounds**—where each class has a distinctive acoustic signature and a visually identifiable subject, giving naturally paired clips and a clean classification oracle.

Our central finding is an *asymmetry*. In the *encoding* direction, audio-only inputs are class-discriminative (`pig` and `sheep` recovered at $5.1\times$ chance, `dog` at $2.4\times$), and audio carries as much class signal as the native RGB stream. In the *decoding* direction, generation collapses in either direction. This asymmetry, together with a train≈test memorization probe, localizes the failure to the generation procedure and the data/tokenizer regime rather than the learned representation—and isolating it precisely is the point of this work.

We make four contributions: (1) an end-to-end pipeline integrating audio (EnCodec tokens, a MusicGen-style delay/flatten pattern, and span masking) into the 4M framework with no architectural change (Fig. 1); (2) a self-built audio–visual dataset over 11 animal classes with multi-stage PANNs/CLIP/VAD cleaning and a leakage-free clip-level split; (3) an empirical study that shows strong learning on structural modalities (depth, normal), localizes the generation limit to high-entropy image synthesis rather than to audio specifically, and quantifies what audio does and does not learn; and (4) a three-cause diagnostic—train/inference masking mismatch, a lossy acoustic tokenizer, and limited scale with acoustic-confusion clusters—each mapped to a concrete lever for making the approach work at scale.

2. Related Work

The 4M family. 4M [1] tokenizes every modality into a discrete sequence and trains one encoder–decoder transformer with random masking; 4M-21 [2] scales this to twenty-one modalities. Both operate on *spatial* streams at large scale. We follow the course re-implementation, nano4M, and ask whether the recipe transfers to a *temporal* modality with only contiguous span masking added and no architectural change.

Audio tokenization and generation. EnCodec [3] provides the discrete residual-VQ audio tokens we model, flattened following the MusicGen delay pattern [4]. AudioLM [5] shows that high-quality audio generation typically layers *semantic* tokens on top of such purely acoustic codes—a gap that helps explain our generation collapse—motivating semantically grounded tokenizers, SpeechTokenizer [6] and MERT [7], as future work.

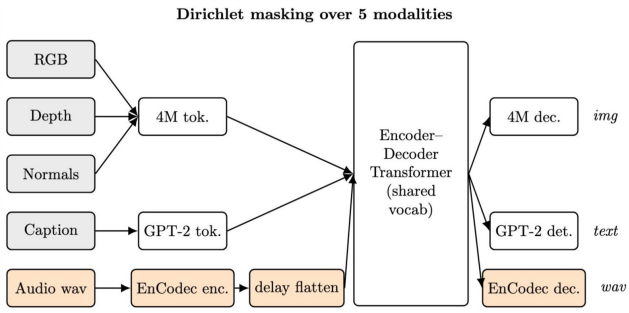

Figure 1. **The nano4M-Audio pipeline.** A single d6-6w512 encoder–decoder (∼96M parameters) over five tokenized, aligned modality streams—RGB, audio, depth, surface normals, and a class caption—sharing one unified token vocabulary with additive modality and position embeddings. Each clip is tokenized by a dedicated per-modality tokenizer (4M DiVAEs for the spatial modalities, EnCodec for audio, GPT-2 BPE for the caption); Dirichlet masking allocates input/target tokens at random across modalities, with the sole exception of *span* masking on the temporal audio stream. Per-modality loss averaging prevents the 512-token audio sequence from dominating the short caption.

Cross-modal audio–visual learning. AudioCLIP [8], CLAP [9], ImageBind [10], and Wav2CLIP [11] align audio with a CLIP [12] space via *contrastive* objectives on 10^5–10^6 pairs. We instead study whether audio–vision alignment emerges *implicitly* from a shared generative masked objective. For generation we use the iterative MaskGIT [13] decoding procedure.

Our work differs in three respects: it is *generative* rather than contrastive; audio is *added to an existing* multimodal framework rather than designed audio-first; and we operate at $\sim 10^4$ clips rather than 10^5–10^6, a deliberate study of the small-scale academic regime. Clips are drawn from AudioSet [14] and VGGSound [15] and curated with PANNs [16], CLIP [12], and Silero VAD [17] oracles, with depth and normal pseudo-labels from Depth-Anything-V2 [18] and DSINE [19].

3. Method

3.1. Architecture

We use **nano4M**, the course re-implementation of the 4M recipe [1], [2]: a d6-6w512 encoder–decoder transformer (∼95.8M parameters, head dimension 64), architecturally *identical* to the visual baseline—we add a fifth modality without touching the network. All modalities share a single *unified/overlap* token vocabulary (∼50,304 ids); modality and position embeddings are added to the token embeddings. Input/target token allocation across modalities follows standard 4M *Dirichlet* random masking, and the per-modality cross-entropy is averaged (length-normalized) so that the 512-token audio sequence does not dominate the ∼5-token caption in the loss.

TABLE 1. THE FIVE TOKENIZED MODALITIES.

Modality	Tokenizer	Shape	Vocab
`tok_rgb@196`	4M-16k DiVAE	$[10, 196]$	16384
`tok_audio@512`	EnCodec 24k, $K{=}2$	$[10, 512]$	2048
`tok_depth@196`	DAv2 → 4M-8k	$[10, 196]$	8192
`tok_normal@196`	DSINE → 4M-8k	$[10, 196]$	8192
`scene_desc`	GPT-2 BPE	list	50304

3.2. Audio tokenization

Audio is tokenized with EnCodec [3] at 24 kHz and 1.5 kbps, using $K{=}2$ RVQ codebooks at 75 Hz. A ∼3.413 s clip yields 256 frames over two codebooks, which we flatten to a length-512 sequence by interleaving the codebooks $(c_1[0], c_2[0], c_1[1], \ldots,$ a MusicGen-style delay/flatten pattern [4]) and offsetting codebook 2 by $+1024$ so that all 2048 audio ids are distinct in the shared vocabulary (Table 1). We deliberately allocate 512 tokens to audio versus 196 to RGB, prioritizing acoustic context—the project's central focus—over visual fidelity; we defer the full reflection to Sec. 5. This is an *acknowledged limitation that foreshadows our diagnostic*: EnCodec is an *acoustic* codec optimized for waveform reconstruction, and its tokens carry no class-level semantic structure. The model must therefore *learn semantics on top of acoustic tokens*—a hypothesis we test empirically in Sec. 4.

3.3. Dataset construction

We query AudioSet [14] and VGGSound [15] for 11 animal classes (`pig`, `sheep`, `dog`, `cat`, `horse`, `cow`, `chicken`, `duck`, `pigeon`, `coyote`, `lion`), a scope where each class has a distinctive acoustic signature and a visually identifiable subject. Clips pass *three oracle filters*: a PANNs [16] audio score ≥ 0.30 on class-specific indices, a CLIP [12] image–text cosine ≥ 0.25 for the best of 10 frames, and a Silero VAD [17] voiced-fraction $= 0$ (no human speech). After deduplication across the two sources we obtain **9,192 clips**, split at the *clip* level and stratified by class×source into 7,347 / 907 / 938 train/val/test to prevent leakage. Each clip provides $K{=}10$ augmentation slots / keyframes, decoupling visual diversity from audio uniqueness. The two visual targets are pseudo-labeled—depth by Depth-Anything-V2-Small [18] and surface normals by DSINE [19] (called per-frame)—and quantized by 4M-8k DiVAEs.

3.4. Training

We train for **18,311 steps** (batch 64, ∼600M tokens total, ∼1h10 on a single H100). The final run is *fp32*: bf16 caused NaNs in the unified 50k-vocab softmax, and fp32 fixed it. We use a cosine schedule ($10^{-4} \rightarrow 10^{-6}$, 916 warmup steps) with AdamW $(0.9, 0.95)$, weight decay 0.05, and gradient clipping 1.0. We fix the random seed and release the deterministic split with the code.

3.5. Span masking for audio

The *single* training-side change to the 4M recipe is the masking of the audio modality. Standard 4M masks tokens at random; because audio is *temporal*, random masking lets the decoder copy adjacent context and never learn long-range structure. We therefore mask audio with *contiguous spans* aligned to codebook pairs (stride 2) for both input and target tokens. The other four modalities keep random masking. The motivation is temporal contiguity—forcing the decoder to predict acoustic structure that cannot be copied from a neighbouring frame.

4. Experiments

4.1. Setup

We evaluate on the held-out 938-clip test set, drawn from the clip-level stratified split (Section 3) so that no clip, frame, or augmentation slot leaks between train and test. At evaluation time we draw a *single random frame per stem*, avoiding the $10\times$ correlation inflation that scoring all $K{=}10$ keyframes would introduce. Our suite spans six families of metrics: per-modality token cross-entropy (CE); audio CE against the codebook's empirical *marginal* entropy; cross-modal retrieval; audio-only classification; conditional generation; and external-classifier validation of generated samples. Every metric is reported against an explicit random baseline: chance $= 1/11 = 9.1\%$ for classification, $k/200$ for retrieval over 200 candidates, $\log(\text{vocab})$ for token CE, and $\sim$0.5–5% top-5 for an ImageNet classifier.

4.2. Training dynamics

Figure 2 shows the per-modality training and validation CE curves; Appendix Fig. **??** summarizes the total CE drop per modality. Three observations structure the rest of this section. First, train and validation track tightly for all five modalities, with validation never exceeding train: the clip-level split removed the leakage we observed in early runs. Second, the two *structural* modalities learn strongly: depth drops $\sim$1.2 nats ($\sim$6.3 $\rightarrow\sim$5.1, against random $\log(8192) = 9.01$) and normals drop $\sim$1.2 nats ($\sim$4.7 $\rightarrow\sim$3.5). Third, audio drops $\sim$1.55 nats (6.75 $\rightarrow$ 5.2); crucially, its final CE of 5.2 lies *below* the empirical marginal entropy of the EnCodec codebook ($\sim$6.2 nats), so the model captures $\sim$1.0 nat of conditional structure per audio token beyond a marginal predictor (Sec. 4.5). The caption saturates near 0.05–0.13, trivially predictable from any other modality. RGB is the hardest modality: it plateaus $\sim$0.45–0.5 nats below random ($\log 16384 = 9.70$), reflecting YouTube-sourced noise, a dense 16k vocabulary, and a small dataset.

4.3. Structural modality reconstruction

We begin with depth and surface normals—the modalities that learned strongly—to demonstrate that the pipeline

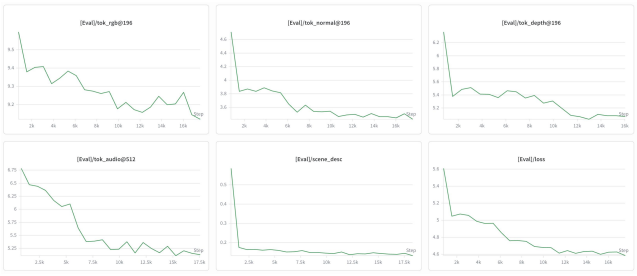

Figure 2. **Per-modality training dynamics.** Training (solid) and validation (dashed) cross-entropy for all five modalities plus the total loss. Train and validation track tightly throughout and validation never exceeds train, confirming the clip-level split prevents leakage.

functions end-to-end. For each test sample we encode the RGB input, predict the depth and normal tokens, and detokenize the predictions back to pixel space through the corresponding 4M-8k DiVAE. Figure 3 shows the result: animal silhouettes are clearly visible in the reconstructed depth maps and the predicted normals capture coherent surface orientation, confirming that the standard 4M tokenizers and decoder integrate cleanly even in our small-data, five-modality regime. Quantitatively, RGB$\rightarrow$depth and RGB$\rightarrow$normal token prediction reach 11.1% and 18.0% token-level top-1 accuracy on the test set respectively, roughly $900\times$ and $1500\times$ over the $1/8192$ random-token baseline.

Visual reconstructions through the 4M-16k DiVAE tokenizer on our YouTube-sourced imagery exhibit noticeable degradation (PSNR averaged $\sim$21 dB across test samples), likely reflecting a domain shift between the tokenizer's CC12M training distribution and our web-scraped dataset. This affects the visual fidelity of decoded RGB samples in qualitative figures but does *not* affect our quantitative cross-modal results, which are measured directly in token space (token-level cross-entropy, classification accuracy on predicted token IDs, structural similarity between predicted and ground-truth token grids). The token-space results—the 11–18% top-1 accuracies above—demonstrate that the model has learned meaningful cross-modal correspondences independent of pixel-level reconstruction fidelity.

4.4. Framework validation: cross-modal directions known to work in 4M

Before attributing observed failures to audio specifically, we exercise the generation infrastructure in four directions the original 4M framework handles canonically: caption$\rightarrow$RGB, RGB$\rightarrow$depth, RGB$\rightarrow$normal, and RGB$\rightarrow$caption. Spatial targets are produced with 8-iteration MaskGIT [13] decoding and the caption by single-step prediction; each direction is scored with a metric tailored to its target (RGB$\rightarrow$depth and RGB$\rightarrow$normal by token-level top-1 accuracy, caption$\rightarrow$RGB by whether an ImageNet ResNet-50 [20], [21] places the target class in its top-5, RGB$\rightarrow$caption by exact class match). A clear split emerges (Fig. 4). The framework predicts *structural* targets from RGB well above chance: RGB$\rightarrow$depth and

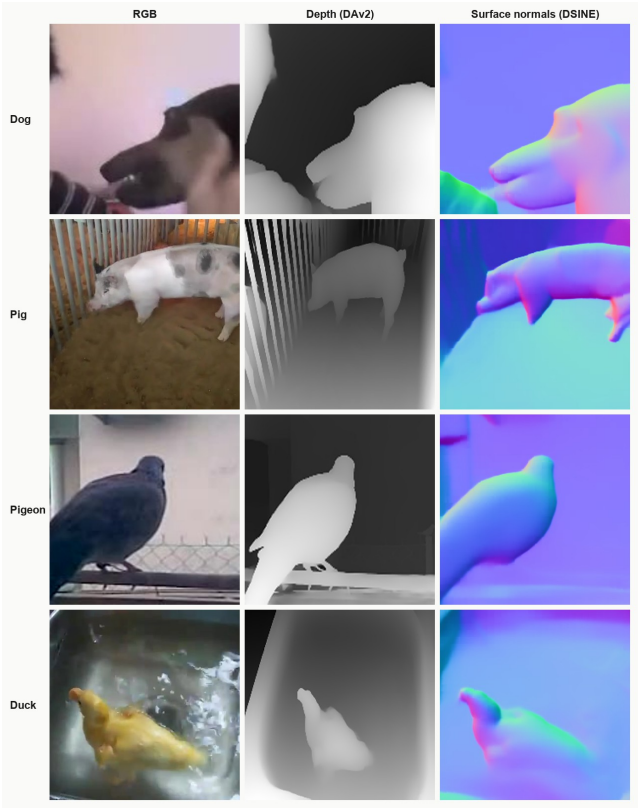

Figure 3. **Structural modality reconstruction.** Four test samples (rows). Left column: RGB input. Middle column: depth predicted from RGB and detokenized via the 4M-8k DiVAE. Right column: surface normals predicted from RGB and detokenized. Animal silhouettes are visible in depth and surface orientation is captured in the normals, demonstrating the pipeline functions end-to-end on the structural modalities.

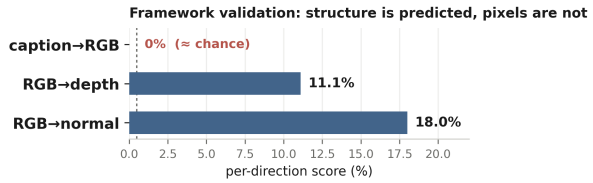

Figure 4. **Framework validation: structure is predicted, pixels are not.** Per-direction score against the relevant random baseline. RGB→depth and RGB→normal reach 11.1% and 18.0% token-level top-1 (≈900× and ≈1500× the 1/8192 random-token baseline), and RGB→caption recovers the class. The single high-entropy synthesis direction, caption→RGB (conditioned on "a photo of a <class>"), scores 0% ImageNet ResNet-50 top-5 (chance ∼0.5%)—the same score as audio→RGB (Sec. 4.5), showing the synthesis limit is general rather than audio-specific.

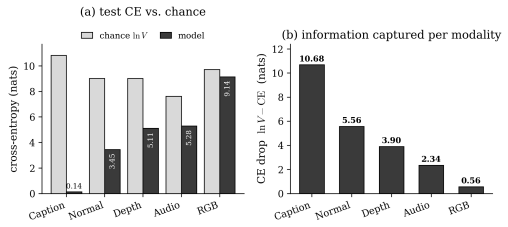

Figure 5. **Audio acquires conditional structure.** Empirical marginal entropy of the EnCodec codebook (6.2 nats) vs. the final audio validation cross-entropy (5.2 nats). The ≈ 1.0-nat gap is conditional structure the model has learned per audio token; it is not, however, enough to drive usable cross-modal generation (Sec. 4.6).

RGB→normal reach 11.1% and 18.0% token-level top-1 (≈900× and ≈1500× their 1/8192 random-token baseline), and RGB→caption recovers the class—so the decoding machinery is sound for low-entropy targets. High-entropy *image synthesis* is the exception: caption→RGB, conditioned on the same "a photo of a <class>" string used at training, scores 0% ResNet-50 top-5 (chance ∼0.5%)—identical to the audio→RGB result in Section 4.5. The inability to synthesize an RGB image is therefore *not* specific to audio conditioning; it is a property of high-entropy pixel generation at our scale. We accordingly read the audio modality through directly measurable token-level quantities—cross-entropy, classification, and retrieval—rather than through generated pixels.

4.5. Audio modality analysis

Audio is the modality the project set out to add, and the one where the framework meets its limit. At the token level it does learn: the final audio validation CE settles at 5.2 nats, below the empirical marginal entropy of the EnCodec codebook (≈ 6.2 nats, itself below $\log 2048 = 7.62$ since the token distribution is non-uniform), i.e. ≈ 1.0 nat of *conditional* structure per audio token beyond a marginal predictor (Fig. 5)—non-trivial learning, not memorization.

Encoding partially works. Ranking classes by the model logit on each class token given audio only, three acoustically distinct classes are recovered well above the $1/11 = 9.1\%$ chance baseline: pig 46.8% (5.1×), sheep 46.6% (5.1×), and dog 21.8% (2.4×). Globally, audio-only reaches top-1 10.4% (logit-ranking; 10.7% under naive sequence-decoding), top-3 28.6%, top-5 48.4%. An input-modality ablation is the key control: audio alone is *comparable* to RGB alone (8.8/26.7/44.8) and to all modalities combined (10.2/27.9/43.5), so the audio stream carries as much class signal as the native visual stream (Appendix Tables **??**, **??**). Acoustically ambiguous classes collapse systematically onto neighbours—the bird cluster (chicken/duck/pigeon) and the canid cluster (coyote/lion→dog)—rather than at random.

Retrieval is faint but real. On mean-pooled encoder embeddings over n=200 candidates (chance R@5 = 2.5%), retrieval is weakly above chance and bidirectional, peaking at depth→audio R@5 = 4.5% (∼ 1.8× chance); the full breakdown is in the Appendix.

Generation collapses. Despite the conditional structure above, generation in either direction mode-collapses. From audio, an ImageNet ResNet-50 [20], [21] recognizes the target class in 0% of 80 audio→RGB images (chance ∼ 5%); the outputs are near-empty and class-independent (Appendix Fig. **??**). From a class label, audio varies in

energy but not content: the per-class RMS-energy spread is 75% of ground truth, yet adding *more* conditioning (RGB+depth+normal+caption) *reduces* the spread to 58%—the decoder does not exploit the extra context, localizing the bottleneck to the generation procedure rather than the input (Appendix Fig. **??**). A memorization probe is decisive: on 80–20 audio-suffix completion, train accuracy (2.9%) $\approx$ test accuracy (4.1%), so the model does not even memorize the training set and the failure is one of scale and train/inference alignment, not over-fitting. **The gap between "audio CE learns conditional structure" and "cross-modal audio generation fails" is our central observation,** and motivates the diagnostic below.

4.6. Failure mode diagnostic

We isolate three causes, each with direct evidence. **(1) Train/inference mismatch.** Random Dirichlet masking almost never presents the "predict an entire modality from another modality alone" condition, so single-source MaskGIT [13] decoding is out of distribution—which is exactly why adding context *hurts* the energy spread rather than helping. **(2) Lossy acoustic tokenizer.** EnCodec [3] at 1.5 kbps is an *acoustic* codec with no class-level semantics: small token errors decode to similar textures, and the model learns timbre and energy rather than categories. This reconciles the 1.0-nat conditional-CE gain with the near-chance global classification—it captures acoustic regularities, not semantic ones. **(3) Scale and acoustic clusters.** At $9{,}192$ clips we sit $\sim 1000\times$ below 4M's data, and several classes are inseparable from their acoustic neighbours. The CE curves saturate cleanly with val $\leq$ train throughout, so additional compute on this data would not help: the binding limits are data *quantity* and tokenizer *quality*, both addressed by the levers in Sec. 5.

4.7. Ablations

Table 2 summarizes the five engineering decisions that shaped the final run, each reported as a self-baseline against the simpler alternative on the metric it most affects. Migrating the RGB tokenizer from Cosmos-64k to the 4M-16k DiVAE lowered the RGB CE plateau by ~ 1.4 nats; doubling the audio window from 256 to 512 tokens dropped audio CE from 6.5 to 5.2 nats; expanding from 1 to 10 keyframes per clip multiplied the visual training signal $10\times$ while keeping audio uniqueness fixed; tightening the PANNs+CLIP oracle thresholds roughly doubled dataset purity at half the size; and a model-size sweep placed d6-w512 as the sweet spot, with d8-w640 over-fitting and d5-w384 under-fitting. The audio-duration row is the one tweak that targets audio directly, and it yields the largest single audio-CE improvement—consistent with the diagnostic that audio learning is data- and tokenizer-bound rather than architecture-bound.

TABLE 2. ENGINEERING ABLATIONS, EACH A SELF-BASELINE AGAINST THE SIMPLER ALTERNATIVE. "IMPACT" IS REPORTED ON THE MOST AFFECTED METRIC.

Decision	Change	Impact
RGB tokenizer	Cosmos-64k $\rightarrow$ 4M-16k	RGB CE $10.6 \rightarrow 9.2$
Audio window	$256 \rightarrow 512$ tokens	Audio CE $6.5 \rightarrow 5.2$
Keyframes/clip	$1 \rightarrow 10$	$10\times$ visual signal
Oracle thresh.	PANN+CLIP tightened	Purity $\times 2$, size $\div 2$
Model size	d5-w384 / **d6-w512** / d8-w640	d6-w512 sweet spot

5. Conclusion and Limitations

What we built. We integrated a temporal, real-world modality into a 4M-style model end-to-end: a clean five-modality pipeline (RGB, depth, normals, audio, caption) with aligned tokenizers, a single training-side change (span masking for audio), and a fully reproducible training run. The structural modalities learned strongly—depth and normal CE each drop $\sim$1.2 nats, their tokens detokenize into recognizable animal silhouettes and surface orientations, and RGB$\rightarrow$depth/normal/caption are predicted well above chance. High-entropy image synthesis is the lone exception (caption$\rightarrow$RGB at 0%, like audio$\rightarrow$RGB), localizing that limit to pixel generation rather than to audio. The 4M recipe absorbs a new modality cleanly *when that modality is structural*—well-tokenized and spatially aligned.

What we observed (the contribution). Audio behaves differently. At the token level it acquires $\sim$1.0 nat of conditional structure per token (CE $5.2 <$ the 6.2-nat marginal), and discriminative audio$\leftrightarrow$class structure emerges for acoustically distinct classes (`pig`, `sheep` at $5.1\times$ chance). But this does *not* lift to usable cross-modal alignment or generation at our scale—a clear *asymmetry* between encoding and decoding. The decisive evidence is the train$\approx$test memorization probe (2.9% vs 4.1% on audio-suffix completion): the model does not even memorize the training set, localizing the failure to the generation procedure and the data/tokenizer regime rather than the learned representation. We treat this systematic invalidation of the generative half of H1 as our central contribution.

Diagnostic levers. Each failure cause maps to a concrete, validatable next experiment. **(L1)** A *semantic* audio tokenizer (SpeechTokenizer [6], MERT [7], or AudioLM-style [5] semantic tokens) on top of the acoustic code. **(L2)** *Asymmetric / curriculum / modality-dropout* masking that explicitly trains single-source-to-full-modality prediction, removing the train/inference mismatch that puts single-source MaskGIT decoding out of distribution. **(L3)** $10\times$+ more data (full VGGSound, $\sim$150k clips) and a larger model, with a 10k $\rightarrow$ 30k $\rightarrow$ 100k scaling study to locate the emergence threshold. **(L4)** A quantitative generation metric, Fréchet Audio Distance [22] via a CLAP [9] embedder, to replace informal listening.

References

[1] D. Mizrahi, R. Bachmann, O. F. Kar, T. Yeo, M. Gao, A. Dehghan, and A. Zamir, "4M: Massively multimodal masked modeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[2] R. Bachmann, O. F. Kar, D. Mizrahi, A. Garjani, M. Gao, D. Griffiths, J. Hu, A. Dehghan, and A. Zamir, "4M-21: An any-to-any vision model for tens of tasks and modalities," *arXiv preprint arXiv:2406.09406*, 2024.

[3] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research (TMLR)*, 2023, arXiv:2210.13438.

[4] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[6] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "SpeechTokenizer: Unified speech tokenizer for speech language models," *arXiv preprint arXiv:2308.16692*, 2023.

[7] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[8] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 976–980.

[9] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 180–15 190.

[11] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from CLIP," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4563–4567.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[13] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "MaskGIT: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 315–11 325.

[14] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[15] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[16] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[17] Silero Team, "Silero VAD: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[18] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything V2," *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[19] G. Bae and A. J. Davison, "Rethinking inductive biases for surface normal estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9535–9545.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[22] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," in *Proc. Interspeech*, 2019, pp. 2350–2354.