# MONOPLEX: PARALLEL SPEECH AND TEXT FROM A SINGLE LATENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A person walking down a street can point, describe the scene aloud and take a note, all at once and from one thought. Machines built on language models say aloud what they write, so their modalities render one content stream rather than working together. We build the parallel alternative, *monoplex*. In monoplex a frozen vision-language model emits one latent, read at once by its own text head and by a speech head we train. Speech carries a summary, text the detail. With a causal completion stage the speech head reaches $85\%$ of a captioning cascade's content in one pass, first audio at $0.12$ s against a streaming cascade's $0.32$ s. Trained on triplets that speak the VLM's own descriptions, the speech head learns directly on a speech-distilled codec, and an acoustic codec a third as much on a clean cache. Ranking eight completed speech samples at inference by the text stream and their mutual agreement reaches $98\%$ of the cascade's content on an embedding measure. A coherence judge in that ranking makes two thirds of transcripts coherent, against all of the cascade's, at $96\%$ of its content. On human captions neither stream wrote, the two streams cover more than text alone, and ranking widens the margin.

## 1 INTRODUCTION

A person speaks, writes and acts from one thought, and none of the three actions waits for another to finish; machines built on language models chain them. Speech generation from images (Hasegawa-Johnson et al., 2017; Wang et al., 2021) is dominated by the cascade: a vision-language model (VLM) writes a caption (Wang et al., 2024), and a text-to-speech (TTS) model reads it aloud. Reusing pretrained components this way makes text a representational bottleneck, since prosody and acoustic context have no text tokens (Défossez et al., 2024), and a serialization barrier, since no audio exists until its text does. Parallel-stream systems are previewed at frontier scale (Thinking Machines Lab, 2026) and ship in the open (Xu et al., 2025a), and text sits on the path wherever their streams differ in content (Chiang et al., 2026): no open system renders one vision latent into two content-divergent streams with no text between them.

We propose *monoplex*, the architecture that analogy suggests: one latent source feeding several surface renderings. Monoplex decodes speech and text in parallel from one frozen VLM's latent: the VLM's own text head writes the detail, a 430M-parameter speech head we train speaks a summary, and no text passes between them. The two streams are deliberately content-divergent, as speaking and writing are.

Our corpus is small for discrete autoregressive speech generation, and the speech stream trains on it because of its target's structure, not our objective. Under the same procedure, the head learns the alignment on Mimi's first codebook, distilled from WavLM (Chen et al., 2022). At 173M it collapses to the random-token floor on acoustic EnCodec (Défossez et al., 2023) on the original cache; on a clean cache it reaches a third of Mimi's share. From that first codebook, a 41M completion stage fills the acoustic codebooks without touching the head, as Moshi's coarse-to-fine decoding does (Défossez et al., 2024).

Two heads that never exchange a token cannot coordinate what they say; we measure what that costs, and coordinate them at inference. One draw of the speech head falls well short of the cascade on content; the best of eight draws from the same latent nearly matches it, so most of that distance is already inside the model. Ranked against the text stream that monoplex emits anyway, those draws recover most of that distance with nothing trained: *cross-stream coupling*.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

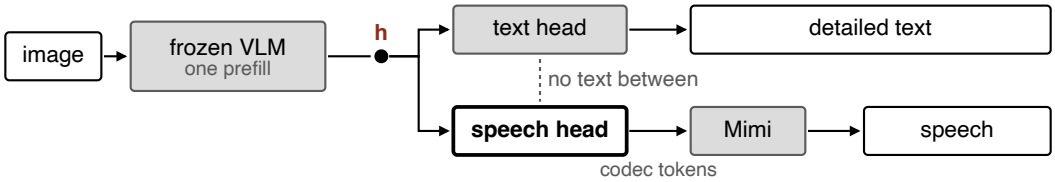



Figure 1: **The system.** One prefill of a frozen Qwen2-VL-2B produces a latent $h$ that the VLM's text head and our speech head read in parallel. The heavy-ruled speech head is trained; grey boxes are frozen.

Our three contributions answer what that parallel arrangement buys and costs:

- **The structure.** *Monoplex*: one frozen-VLM latent read by parallel heads that exchange no text, built and measured end to end against a matched cascade.
- **Why the speech stream is trainable.** The speech head trains because the target codec's first codebook is distilled from a speech model: its alignment is inherited from that pretraining rather than produced by our objective. We show which instruments see the alignment and which do not.
- **Cross-stream coupling.** Two heads with no path between them factorize, and we price what that costs on a target neither stream wrote. Ranked against the sibling text stream with no parameters updated, speech candidates recover most of the eight-draw oracle's gap. An unrelated ASR and a second encoder confirm the gain is the sibling's rather than the draws' agreement.

## 2 Related Work

**Speech codec language models and their targets.** This recipe descends from codec-token language models (Borsos et al., 2023; Wang et al., 2023a; Copet et al., 2023; Kreuk et al., 2023), mostly over acoustic codecs (Défossez et al., 2023). That an acoustic target costs a codec LM its content and a semantic one restores it is found for unit choice (Lakhotia et al., 2021), codec design (Ye et al., 2025a) and across tokenizers (Mousavi et al., 2026); SpeechTokenizer (Zhang et al., 2024) and Mimi (Défossez et al., 2024) distill the first codebook from HuBERT (Hsu et al., 2021a) and WavLM (Chen et al., 2022) for that reason.

**Direct image-to-speech.** Text-free image-to-speech runs from Image2Speech (Hasegawa-Johnson et al., 2017) to Show and Speak (Wang et al., 2021), Im2Wav (Sheffer & Adi, 2023) targets sound, and SVLA (Huynh et al., 2025) compares that direct path with a text-mediated one. Closest are Hsu et al. (2021b), over units learned on paired image–speech data, and Kim et al. (2024), a HuBERT-unit decoder over a frozen vision-language encoder on COCO, whose targets are the captions read aloud, so its scale and ours do not meet. Our codebook is vision-agnostic, so the alignment cannot be inherited from it, and our conditioner's own text head runs beside the speech head.

**Parallel generation, and what it leaves in place.** Interaction Models (Thinking Machines Lab, 2026) and GPT-Live (Uberti & Malkani, 2026) established the parallel behavior we build on. PSLM (Mitsui et al., 2024), Mini-Omni2 (Xie & Wu, 2024b) and DrVoice (Tan et al., 2026) generate text and speech tokens in parallel from one LLM, and Qwen3-Omni's Talker reads text tokens and multimodal features (Xu et al., 2025b). Speech-Omni-Lite (Tao et al., 2026) trains a speech-token generator over the hidden states of a frozen Qwen3-VL (Bai et al., 2025a), conditioned on speech, as VLAS (Zhao et al., 2025) is for manipulation, and its two outputs render one content; ours read an image and diverge in level of detail. Parallel streams that differ by role are published text-only (Su et al., 2026), leaving open the cross-modal case: different content from one vision latent. Speech-capable LLMs (Xie & Wu, 2024a; Fang et al., 2025; Défossez et al., 2024; Nguyen et al., 2025; Wang et al., 2025) and omni models with vision (Zhang et al., 2025; Wu et al., 2024) emit speech that renders their text. MoshiVis (Royer et al., 2025) adapts a frozen Moshi to images with its inner monologue on the path, and reasoning-while-speaking models feed the speech their reasoning as text (Chiang et al., 2026): faster, one content stream at a time. Modality experts (Lou et al., 2026) are the alternative to a separate head; we test one (§5.4).

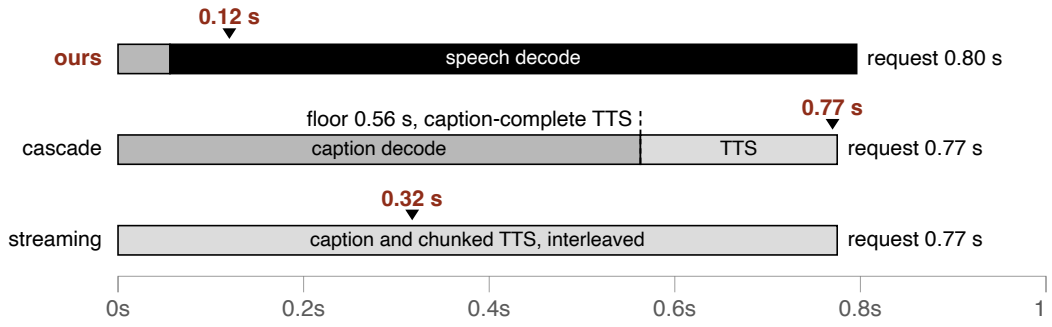



Figure 2: **What its shape buys.** The same system on one clock, one idle GPU for every row; the streaming row hands six-word chunks to a GPU synthesizer (Appendix B.4); the coupled arm of §5.5 is a second operating point, priced on its own device in Appendix B.4, and is not drawn.

**Selection.** Keeping the candidate a utility scores highest is minimum Bayes-risk decoding (Kumar & Byrne, 2004; Freitag et al., 2022; Eikema & Aziz, 2020; Bertsch et al., 2023), whose consensus, cross-modal, speech and shared-encoder forms are self-consistency (Wang et al., 2023b), CLIP reranking (Ramesh et al., 2021), verifier-ranked candidates (Betker, 2023; Ye et al., 2025b) and two-pass ASR (Sainath et al., 2019). In a monoplex the utility is a stream already emitted. FLAIR keeps a continuous latent within one stream (Wu et al., 2026); CoDi couples parallel generators by cross-attention (Tang et al., 2023), a path a monoplex does not have.

# 3 Method

## 3.1 Shared-latent parallel dual-stream generation

A monoplex maps one input to a shared latent that several decoders read in parallel, each in its own surface form and level of detail (speech and text, in this paper). A frozen VLM gives hidden states $h(x) \in \mathbb{R}^{L\times d}$, $d$=1536; Mimi (Défossez et al., 2024) maps the waveform to discrete tokens $\tau_{1:T}$ over its first codebook, cb0, a codebook of $V$=2048. Two heads read that latent:

$$p(\tau, w \mid h) \;=\; \underbrace{p_\psi(\tau \mid h)}_{\text{speech: summary}}\;\underbrace{p_{\text{txt}}(w \mid h)}_{\text{text: detail}}, \qquad p_\psi(\tau \mid h) \;=\; \prod_{t=1}^{T} p_\psi\big(\tau_t \mid \bar{h},\, \tau_{<t}\big), \;\; \bar{h} = \tfrac{1}{L}\textstyle\sum_\ell h_\ell. \tag{1}$$

We leave the detail to the VLM's frozen text head $p_{\text{txt}}$, which attends over all of $h$. The speech head $p_\psi$ we train reads the latent's mean over the prefill, $\bar{h}$. From that mean it speaks the summary (§3.3; Appendix A gives the full notation). The product in Eq. 1 is the whole design, drawn in Figure 1. Both streams condition on $h$ and on nothing else. So neither waits for the other, and neither can read what the other said. We chose that independence for this instantiation; the architecture does not require it. §5.5 measures what it costs.

**Objective.** The speech head minimizes autoregressive cross-entropy on the codec tokens, $\mathcal{L}_\psi = -\sum_t \log p_\psi(\tau_t \mid \bar{h}, \tau_{<t})$, weighted across codebooks as in §3.4, plus one auxiliary term at weight 0.1 predicting the utterance's length from $\bar{h}$: a stream with no text beside it has nothing that tells it when to stop. Nothing in $\mathcal{L}_\psi$ mentions $w$; the text head is untouched.

## 3.2 Cross-stream coupling

The speech head never sees the sibling's text, so nothing in training prefers the speech that agrees with it; the dependence returns at inference, from streams the architecture already emits. We draw $K$ speech candidates $\tau^{(k)} \sim p_\psi(\cdot \mid h(x))$ independently. The text stream $w$, decoded greedily from $p_{\text{txt}}(\cdot \mid h(x))$, is the one the VLM's own head generates anyway. We keep the candidate that agrees

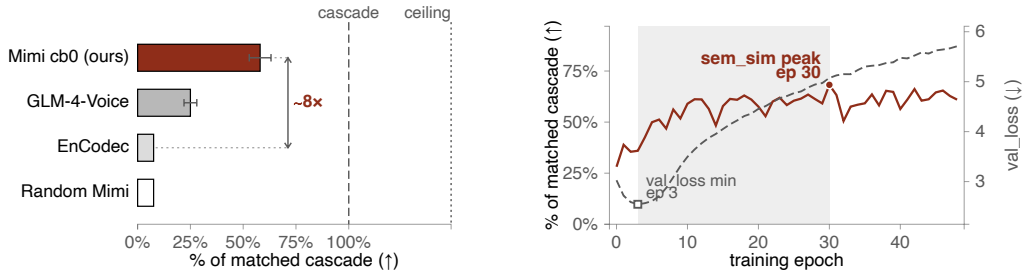



Figure 3: **The codec decides whether the stream trains, and the loss cannot see it.** Left: codec swap under matched protocol on the original 10K cache (test split, $n$=200 per seed, whisper-turbo); rules mark the protocol's matched cascade and the transcription ceiling, error bars the seed standard deviation. Right: on one run, val_loss bottoms at epoch 3 while sem_sim peaks at epoch 30; the shaded span is where they move against each other.

most with it:

$$\tau^\star \;=\; \underset{k\in\{1,\dots,K\}}{\arg\max}\; s\big(\tau^{(k)},\, w\big). \tag{2}$$

The score $s$ measures agreement in a shared semantic space. §5.5 instantiates it as the sentence-embedding similarity between a candidate's transcript and $w$. Those are the metric's own instruments, and Appendix C controls what they share. Appendix C.1 reads the codec tokens instead. Nothing is trained and nothing about the target is used, yet the emitted speech now depends on the text, $p(\tau^\star \mid w, h) \neq p_\psi(\tau \mid h)$. The factorization of Eq. 1 is undone at inference, on the streams the system already emits. The price is drawing $K$ candidates before the pick.

### 3.3 Architecture

Everything trained sits outside the VLM (Figure 1). The backbone is Qwen2-VL-2B (Wang et al., 2024); we extract its last layer's hidden states for a fixed prompt ("Describe this image.") once and train on the cache (§3.4).

**Two sizes, one objective.** We use a causal transformer (Vaswani et al., 2017) over Mimi codebook positions as the speech head. The 10K experiments and the 100K 173M row use it at 173M (12 layers, $d_{\text{model}}$=1024); the 100K headline uses it at 430M (14 layers, $d_{\text{model}}$=1536). The same hidden states reach both sizes, mean-pooled over the prefill and projected into every decoder input, under one objective. Mimi (Défossez et al., 2024) produces 8 codebooks at 12.5 Hz, cb0 distilled from WavLM and cb1-7 acoustic. We score every row of Table 1 above the completion stage on the cb0 decode (Appendix C.3; the head predicts all eight at 100K, but cb1-7 train at weight 0.02 and learn little). We train a 41M transformer $p(\text{cb}_{1:7} \mid \text{cb}_0)$ on the corpus's Mimi tokens with the head frozen (Appendix C.4): this *acoustic completion* stage, AudioLM's and Moshi's coarse-to-fine move (Borsos et al., 2023), fills cb1-7 from whatever cb0 stream the speech head emits.

### 3.4 Training

The objective is cross-entropy on Mimi tokens. Semantic weight is 100, as in Moshi (Défossez et al., 2024), against acoustic weight 0.02, as in PersonaPlex (Roy et al., 2026).

## 4 Experiments

**Dataset.** We train and test on two disjoint corpora with different synthesizers, Bark (Suno AI, 2023) at 10K and edge-TTS (rany2, 2025) at 100K. Both are MS-COCO (Lin et al., 2014) image-text-speech triplets, a Qwen2-VL (Wang et al., 2024) long description per image summarized to a short spoken summary text (SST) by Qwen2.5-1.5B-Instruct (Qwen et al., 2025), then synthesized as audio and tokenized with Mimi (Défossez et al., 2024). Appendix A audits the targets: a quarter of them end mid-clause at the generator's length cap, and the content instrument does not move with the cut.

Table 1: Main results on the 100K fixed test split, whisper-large-v3, $n$=199 (smaller cells in Appendix B.1). % of cascade: sem_sim as a share of the matched cascade's, ± across training seeds. Latency: medians on one GPU. Retrieval and coherent: seed 43. Shaded: this paper's system. † Kokoro latency with edge-TTS scores; ‡ non-streaming generation; ¶ as spoken (Appendix B.4).

| | Latency (s) | | | Retrieval @1 | | |
|---|---|---|---|---|---|---|
| **System** | **first audio** | **request** | **% of cascade** | **transcript** | **audio** | **coherent** |
| ***Text-serialized systems*** | | | | | | |
| cascade, Qwen2-VL-2B → TTS | 0.56† | 0.77† | 100 | 1.000 | 1.000 | 100 |
| streaming, six-word chunks → Kokoro | 0.32 | 0.77 | 100 | | | |
| Qwen2.5-Omni-3B, Thinker → Talker | | 14.2‡ | 89¶ | | 1.000 | |
| ***Monoplex: one latent, no text between the heads*** | | | | | | |
| monoplex, 173M head | | | 69±3 | | | |
| monoplex, 430M head | 0.12 | 0.80 | 73±2 | 0.930 | 0.759 | 27 |
| + acoustic completion, 41M, head frozen | | | 85.6±0.7 | 0.975 | 0.965 | 33 |
| made causal at 90M, streams | 0.12 | 0.80 | 85 | | | |
| control, target summary as its input, retrained (§B.2) | | | 87 | | | |
| + acoustic completion, the same 41M stage | | | 104 | | | |
| ***Cross-stream coupling over $K$=8 draws, nothing trained*** | | | | | | |
| rank the draws by the sibling text | | | 90 | | | |
| + acoustic completion (draw 2: 91 →) | | | 94 | 0.990 | 0.985 | |
| rank the completed draws instead, both seeds | | | **98** | | | 37 |
| + a fluency judge in the vote | | | 96 | | | 66 |
| ***Ceiling and floor*** | | | | | | |
| ground-truth summary, rendered | | | 148 | 1.000 | 1.000 | 85 |
| through Mimi cb0, our decode path ($n$=60) | | | 115 | | | |
| permutation null ($n$=50), chance | | | 19 | 0.250 | 0.250 | |

The original 10K cache's prompt carried the summary itself, a leak we document and test for in Appendix A (the 100K cache and a clean 10K cache do not have it). Rows marked *clean cache* use that clean 10K cache, rows 10,000–19,999 of the 100K edge-TTS corpus cached with the fixed prompt. We measure the 100K rows and the VLM block on it under a leakage-controlled protocol: fixed test split, whisper-large-v3 (Radford et al., 2023), 2 seeds per 100K head, checkpoints selected on validation.

**Metrics.** **Semantic similarity** (sem_sim) scores what was said and not which words: we generate speech tokens, decode them via Mimi to waveform, and transcribe via Whisper (Radford et al., 2023). Take the sentence-embedding cosine similarity (all-MiniLM-L6-v2 (Reimers & Gurevych, 2019)) between transcript and the ground-truth SST. **Token accuracy** is a diagnostic only (§5.2). The **permutation null** (each transcript against another image's SST; Ernst, 2004) is Table 1's chance line; the reference summary, rendered and transcribed, is its ceiling. Every % of cascade, tables included, is a ratio of means: the arm's mean score over the test items against that of the **cascade** that prompts the same VLM for one short sentence, the prompt every other comparator answers (Appendix B.4 reads the same split against a cascade prompted in the summary's own register). Appendix B.1 drops the reference and scores the transcript against its own image and three distractors, chance 0.250, on CLIP ViT-L/14 (Radford et al., 2021) embeddings.

# 5 Results

## 5.1 The Speech Stream

We train the speech stream on the 10K MS-COCO (Lin et al., 2014) triplet corpus of §4, and the target codec decides whether it does: our 173M speech head reaches a 3-seed mean of $58 \pm 5\%$ of the matched cascade on the 10K test split against the 8% random-token floor the same 173M head reaches on EnCodec (Défossez et al., 2023), and $48 \pm 3\%$ against 16% on a clean cache carrying no leaking prompt, in Figure 3, left, Tables 2 and 9. At 100K we reach 73% with the 430M head. With the 41M completion stage we lift the same draws to 0.577, 86% of the cascade on semantic similarity, and the second training seed's draws to 0.570, in sentences a judge reads as coherent a third of the time (Table 1, last column). Table 6 reads the same stream, fact by fact, at 0.20 of the reference's units against the cascade's 0.30. DNSMOS (Reddy et al., 2022) rises from 1.07 to 3.27



**122%** best of 58

TARGET A curious black cat sits inside an open white suitcase on a beige bed surrounded by various objects and fabrics, creating a cozy and inviting.

OURS A **black** and **white cat sits** calmly **inside** an old and **black suitcase** with its furniture on top. Your hand.



**117%** second of 58

TARGET A cozy living room with a wooden entertainment center, framed pictures, a wooden side table with a lamp, an armchair near a TV, a side table with.

OURS A **cozy living room** featuring a **wooden** coffee **table,** a vase wall, and a dark **wooden** chair, arranged in various items like mirrors, books, papers, and pens.



**114%** third of 58

TARGET A messy office setup with a black desk adorned by a blue exercise ball, cluttered with a computer monitor, keyboard, mouse, lamp, various books, and.

OURS A **cluttered desk** in an **office** space featuring a laptop displaying documents with multiple **books,** CD and Reds, and the **books** all set against A.

Figure 4: The three of the sixty held-out images of Appendix E that our audio scores highest on; the appendix shows all sixty. Whisper transcripts of our audio against the head's training target, with a word set dark where it also occurs in the target, computed rather than chosen.

for these draws, where the reference through all eight Mimi (Défossez et al., 2024) codebooks reads 3.40 (it is a reference-free predictor of a listener's quality rating on a 1-to-5 scale). On the forced choice of Appendix B.1, completion lifts the raw head's 0.759 to 0.965 (Table 1, audio column).

**Latency, and time to first audio.** Per request the monoplex and cascade paths are even on one idle GPU (0.80 s ours, 0.77 s the cascade on Kokoro (hexgrad, 2025); Figure 2), so the gain is in first audio and not in the request. A listener waits for the first audio, which arrives at 0.12 s. The Kokoro cascade's full caption takes 0.56 s and a streaming cascade of six-word chunks 0.32 s, and that streaming cascade still renders one content stream (Figure 10, right, and Appendix B.4). Qwen2.5-Omni-3B (Xu et al., 2025a), answering the same question in speech, reaches 0.598 at 14.2 s per request in its non-streaming generate (its text completes at 1.2 s). Its speech and its own text carry the same content, at cosine 0.98 on $n$=199, so its two streams differ in delivery and not in what they say (Appendix B.4).

### 5.2 Token vs Semantic Decoupling

Figure 4 shows examples and Table 11 the rates of token-level and semantic-level metrics decoupling, Mimi reaching 58% of the cascade at 4.9% token accuracy (while EnCodec sits at the 8% content floor with the higher 7.4%), so token accuracy reports the tokens wrong (where the embedding metric reports the content right). We put the split in numbers with a judge that ignores content (asked whether a transcript reads as coherent English, an instruction-tuned model says yes for 0.33 of the completed stream's items against 0.85 for the reference summary through the same TTS and ASR; Table 5). We check the split against a legible voice: with the 41M completion stage the speech head's WER stays at 0.91 while the same path returns the reference's words at 0.05, so the string metric is reading the paraphrase, not the audio. See Table 12. Unlike the position-sensitive metrics, word overlap does not split from the embedding metric. We read the same picks within a point of sem_sim on METEOR (Banerjee & Lavie, 2005), and lower on ROUGE-L (Lin, 2004) (Appendix B.4).

Table 2: The 10K ablation arms: one factor varied per block, same head, data and schedule unless the block names otherwise (Appendix B.5). Filled rows are this paper's system.

| Arm | % of cascade |
|---|---|
| *Codec target* | |
| Mimi cb0, 3 seeds | 58±5 |
| GLM-4-Voice, 3 seeds | 25±3 |
| EnCodec, acoustic | 8 |
| Mimi cb0, clean cache, 3 seeds | 48±3 |
| SpeechTokenizer cb0, clean cache | 60±5 |
| EnCodec, clean cache | 16 |
| *Head architecture* | |
| causal AR, 173M, 3 seeds | 58±5 |
| cross-attention + depth, 268M | 14 |
| causal AR, clean cache, 3 seeds | 48±3 |
| cross-attention, clean cache | 9 |

Table 3: Cross-stream selection at $K=8$, decomposed. Every arm is scored on *identical* draws inside one run. Fixed test split, 100K 430M checkpoint, seed 43, whisper-large-v3 (Radford et al., 2023), $n=199$; intervals are 95% on the paired per-image difference against no selection. The last two columns are a selector's gain as a share of the oracle's, and its paired record against one draw, where a tie is an image on which the two arms returned the same draw.

| Selector | % of cascade | gain [95% CI] | % of oracle | win / tie / loss vs. one draw |
|---|---|---|---|---|
| none (one draw) | 73 | | | |
| self-consensus (no text) | 85 | +12 [+9, +15] | 54% | 130/32/37 |
| **sibling text stream** | **90** | **+17** [+14, +20] | **76%** | **140/29/30** |
| oracle (target-aware) | 95 | +22 [+20, +25] | 100% | 171/28/0 |

*Sibling text over self-consensus:* +5 [+3, +7], 86/65/48, so the sibling wins 86 of the 134 images where the two selectors differ. *The same, scored against five human captions of the image instead of the target* ($n$=1993): +0.015 [+0.011, +0.018] on that scale, 813/576/604, so higher on 813 of the 1,417 where they differ.

### 5.3 The Target Codec

The codec sets §5.1's random-token floor, not the architecture. The share of the matched cascade falls when the target is swapped (Mimi's 58% to EnCodec's 8% on the original cache, 48% to 16% on the clean one, protocol held; Table 2, codec block; Figure 13). The same TTS on ground-truth text reads 148% (Figure 3, left). We swap in a second semantic codec, SpeechTokenizer (Zhang et al., 2024), whose first codebook is distilled from HuBERT (Hsu et al., 2021a) at 50 Hz, and it reaches $60 \pm 5\%$ over two seeds on the same clean cache with nothing else changed, at four times Mimi's frame rate. Mimi stays under the latency rows because of that frame rate (Appendix B.4).

### 5.4 The Head's Architecture

The plain causal head, a causal transformer (Vaswani et al., 2017), is the arm to beat, and nothing more elaborate beats it. We hold training to matched cached hidden states and decode every arm through the same cb0-only path, and the plain causal head dominates: cross-attention (Vaswani et al., 2017) over the VLM states loses to it (while adding parameters), on the clean cache as on the original (Table 2, head block). A speech expert inside the VLM in the style of MoST (Lou et al., 2026) collapses in both variants (in the live-forward regime its own control scopes; Appendix B.5). Every VLM in Table 8 sustains the head's effect.

Data moves more than architecture and less than the target codec (read against its own matched cascade): the 173M head goes from 58% at 10K to 69% at 100K and the 430M head to 73% (Table 10, Appendix B.5). Those corpora differ in synthesizer and in the 7.5× training items (the transfer control of Appendix B.5 bounds that difference).

### 5.5 Cross-Stream Coupling

**The model already draws a near-cascade sample.** The two heads factorize (§3.1), so part of the distance to the cascade is coordination they never had (we measure that part here and remove most of it at inference, training nothing). With $K=8$ samples drawn per image, the best of the 8 scores 95% of the matched cascade against 73% for one draw (Table 3, $n=199$).

**Agreement among the draws recovers half of it.** Self-consensus, minimum Bayes-risk decoding over the same draws (Kumar & Byrne, 2004; Eikema & Aziz, 2020), reaches 0.569 and recovers 54% of the oracle gap under the metric's own ASR and encoder (Table 3). When the draws are rescored end to end over 1,323 images by an ASR sharing no lineage with those and a second encoder, wav2vec2-CTC (Baevski et al., 2020; Graves et al., 2006) and MPNet (Song et al., 2020), agreement gains +0.021 on a bootstrap interval (Efron, 1979) of $[+0.014, +0.028]$. The sibling stream gains +0.141 $[+0.134, +0.148]$ on the same images. Appendix C.1 gives both.

**The sibling stream adds what agreement cannot.** Ranking the same draws against the sibling text stream by Eq. 2 reaches sem_sim 0.603, 90% of the matched cascade against 73% for one draw. On the second training seed it reaches 91% against 72%. The gain over self-consensus is +5 points,

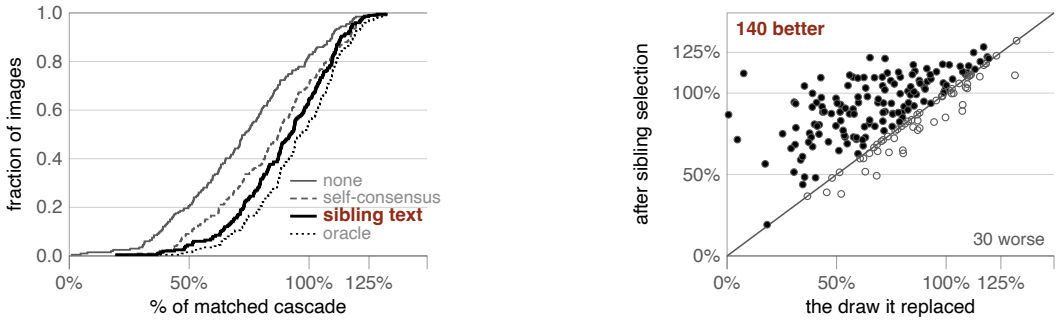



Figure 5: **Selection moves the body of the distribution, not a tail.** Same run as Table 3, $n$=199, identical draws across arms. Left: empirical CDF per arm. Right: each image's selected draw against its first; 30 images fall below the diagonal.

higher on 86 of 199 images, and $+5$ $[+3, +7]$ on the second training seed. Figure 5 shows it moving the whole distribution rather than a tail. Ranking after the completion stage rather than before it, on what a listener actually hears, reaches $0.647$, and adding the agreement among the completed draws to the sibling score reaches 0.654 on both training seeds, which is 98% of the matched cascade (Appendix C.4). On fact units the ranked stream reads $74\%$ of the cascade against the completed draw's $65\%$ (Table 6). That instrument separates the arms with and without the completion stage, whose intervals are clear of each other, and separates what ranking adds only on the paired difference, $+0.026$ $[+0.002, +0.051]$. The eight draws buy content and little fluency: that arm reads as coherent English in $0.37$ of items against the completed draw's $0.33$ (Table 5). A judge that reads only the transcript can join the sibling score and the agreement in the vote. With its coherence margin as a third signal, the pick reads as coherent English in two thirds of items, against $0.37$ and $0.48$ without it on the two training seeds, at a content cost under $0.01$ (Table 5).

## 6 Discussion

**What the structure buys.** A monoplex is feasible. With a small completion stage, monoplex reaches most of a text-serialized cascade's content on an embedding measure, and with a fluency vote over monoplex's own draws, two thirds of its transcripts read as coherent English at $96\%$ of that content (Tables 1 and 5). Monoplex's audio is out before the cascade has a caption. Scored against five human COCO (Lin et al., 2014) captions of the image, a target neither stream wrote, the pair carries more than the written stream alone, $+0.050$ on bootstrap intervals (Efron, 1979) of $[+0.042, +0.058]$ on the share of captions matched and $+0.037$ $[+0.022, +0.053]$ with the word budget held fixed. A head retrained to read the target summary itself reaches $104\%$ under the same completion stage, $18$ points above the parallel arm (Table 1). A finished text is worth that much content, and it is what parallelism gives up to start before any text exists.

**What the coupling result does and does not show.** We read the second-ASR rescoring of §5.5 as the second stream contributing content the draws do not share among themselves. The cascade does not lack selection: ranking eight of its own captions returns it to its greedy score (Appendix C.1). Shared authorship alone does not explain it either: the sibling selector still gains over both baselines when the target is five human captions ($n$=1993, Appendix C). The arm prices coordination rather than deploying it: on one GPU it takes $6.5$ s, $16\times$ the streaming cascade's request, and three quarters of that is the eight transcriptions (Appendices B.4 and C.1).

**Why the speech stream is trainable.** We find the alignment inherited from the target codec's pretraining, not produced by the recipe (Appendix C.2). The decode path is not what fails (Appendix B.5): the speech head is handed an intelligible EnCodec (Défossez et al., 2023) target it learns little from on this corpus, and a Mimi target (Défossez et al., 2024), organized by WavLM (Chen et al., 2022) at a sixth of EnCodec's frame rate, that it learns.

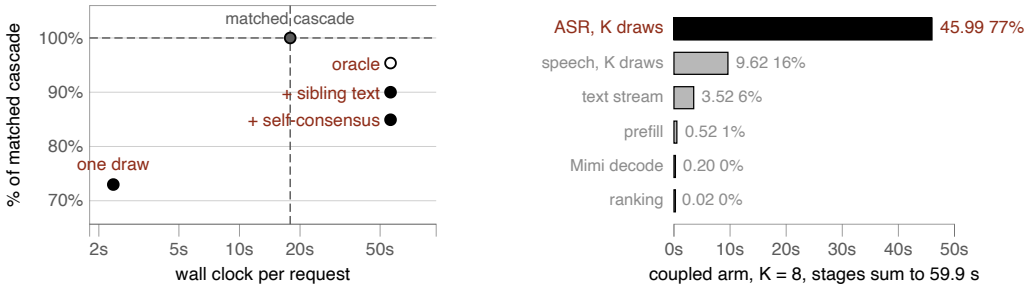



Figure 6: **What cross-stream selection costs.** Left: every arm on one GPU against wall clock per request; the three selection arms share one cost because each transcribes the $K$ candidates. Right: where the time goes. The text stream runs alongside the draws, so the stages sum past the $56\,\mathrm{s}$ request. The cascade point carries the Qwen2-VL-to-Bark cascade's $17.87\,\mathrm{s}$ and edge-TTS's quality; Appendix B.4 times the Kokoro cascade against the arm on one card.

## 7 Limitations

An ASR mediates every semantic number. Table 1 therefore also asks the image directly (Appendix B.1), as does Appendix B.2 on NoCaps (Agrawal et al., 2019). The headline rows are 199-item cells, and the one-draw cell holds on the whole test split (Appendix B.5). Every target descends from one corpus family and one VLM's descriptions. More broadly, the evidence covers one task, describing English photographs aloud, with synthesized voices as the speech targets and automatic instruments as the judges. Whether one latent serves open conversation, other languages and human listeners as well is untested.

## 8 Conclusion

For describing what it sees, a monoplex lets a machine do what a person can do, with one latent, several renderings and none waiting for another. We show it on one task, describing English photographs aloud. Trained on a small corpus of triplets that speak the VLM's own descriptions, its speech head learns directly when the codec's first tokens already carry a speech model's semantics. The speech stream is measurable only by instruments that read content rather than words. Coordination the factorization never gave the two heads is most of what separates monoplex from the cascade. At inference, ranking speech samples by the sibling text stream recovers most of that. Training that coordination, with both streams kept parallel and neither serialized through the other, is the next step.

## AI Use Statement

Generative AI was used to aid and polish the writing of the paper; every training and test triplet comes from the model-generated pipeline of §4. The authors take full responsibility for the content, including any text written or polished with the aid of generative AI.

## References

Harsh Agrawal, Karan Desai, Yufei Wang, et al. nocaps: Novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956. IEEE, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, et al. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report, 2025b.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Adrien Bardes, Quentin Garrido, Jean Ponce, et al. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Amanda Bertsch, Alex Xie, Graham Neubig, and Matthew Gormley. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In Yanai Elazar, Allyson Ettinger, Nora Kassner, Sebastian Ruder, and Noah A. Smith (eds.), *Proceedings of the Big Picture Workshop*, pp. 108–122, Singapore, December 2023. Association for Computational Linguistics.

James Betker. Better speech synthesis through scaling, 2023.

Kevin Black, Noah Brown, Danny Driess, et al. $\pi_0$: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025.

Zalán Borsos, Raphaël Marinier, Damien Vincent, et al. AudioLM: a language modeling approach to audio generation, 2023.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, et al. STITCH: Simultaneous thinking and talking with chunked reasoning for spoken language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jade Copet, Felix Kreuk, Itai Gat, et al. Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Featured Certification, Reproducibility Certification.

Avihu Dekel, Slava Shechtman, Raul Fernandez, et al. Speak while you think: Streaming speech synthesis during text generation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11931–11935. IEEE, 2024.

Zhihao Du, Yuxuan Wang, Qian Chen, et al. CosyVoice 2: Scalable streaming speech synthesis with large language models, 2024.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, et al. Moshi: a speech-text foundation model for real-time dialogue, 2024.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

Michael D. Ernst. Permutation Methods: A Basis for Exact Inference. *Statistical Science*, 19(4): 676–685, 2004.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Qingkai Fang, Shoutao Guo, Yan Zhou, et al. LLaMA-Omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 369–376, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832.

Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image2speech: Automatically generating audio descriptions of images. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, Casablanca, Morocco, 2017.

hexgrad. Kokoro-82M: An open-weight text-to-speech model. Hugging Face model repository, 2025. URL `https://huggingface.co/hexgrad/Kokoro-82M`. Version 1.0.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021a.

Wei-Ning Hsu, David Harwath, Tyler Miller, Christopher Song, and James Glass. Text-free image-to-speech synthesis using learned segmental units. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5284–5300, Online, August 2021b. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Imran Razzak, Hakim Hacid, and Sunil Aryal. SVLA: A unified speech-vision-language assistant with multimodal reasoning and speech generation, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B, 2023.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137. IEEE, 2015.

Hyeongju Kim, Jinhyeok Yang, Yechan Yu, et al. SupertonicTTS: Towards highly efficient and streamlined text-to-speech system. *arXiv preprint arXiv:2503.23108*, 2025.

Minsu Kim, Jeongsoo Choi, Soumi Maiti, et al. Towards practical and efficient image-to-speech captioning with vision-language pre-training and multi-modal tokens. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7970–7974. IEEE, 2024.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, et al. AudioGen: Textually guided audio generation, 2023.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2004. Association for Computational Linguistics.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yuxuan Lou, Kai Yang, and Yang You. MoST: Mixing speech and text with modality-aware mixture of experts, 2026.

Andrés Marafioti, Orr Zohar, Miquel Farré, et al. SmolVLM: Redefining small and efficient multimodal models, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. PSLM: Parallel generation of text and speech with LLMs for low-latency spoken dialogue systems. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2692–2700, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Pooneh Mousavi, Jarod Duret, Darius Petermann, et al. DASB - discrete audio and speech benchmark. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

Karl Pearson. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, December 1895. ISSN 0370-1662.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Alec Radford, Jong Wook Kim, Tao Xu, et al. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, et al. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021.

rany2. edge-tts: Use Microsoft Edge's online text-to-speech service from Python. GitHub repository, 2025. URL `https://github.com/rany2/edge-tts`. Version 7.2.7.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2021.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890. IEEE, 2022.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, et al. PersonaPlex: Voice and role control for full duplex conversational speech models, 2026.

Amélie Royer, Moritz Böhle, Gabriel de Marmiesse, et al. Vision-speech models: Teaching speech models to converse about images, 2025.

Takaaki Saeki, Detai Xin, Wataru Nakata, et al. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Proceedings of Interspeech 2022*, pp. 4521–4525. ISCA, 2022.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. In *Proceedings of Interspeech 2024*, pp. 4943–4947. ISCA, 2024.

Tara N. Sainath, Ruoming Pang, David Rybach, et al. Two-pass end-to-end speech recognition. In *Proceedings of Interspeech 2019*, pp. 2773–2777, 2019.

Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, et al. SmolVLA: A vision-language-action model for affordable and efficient robotics, 2025.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556.

Guinan Su, Yanwu Yang, Xueyan Li, and Jonas Geiping. Multi-stream LLMs: Unblocking language models with parallel streams of thoughts, inputs and outputs, 2026.

Suno AI. Bark: Text-prompted generative audio model. GitHub repository, 2023. URL https://github.com/suno-ai/bark.

Chao-Hong Tan, Qian Chen, Wen Wang, et al. DrVoice: Parallel speech-text voice conversation model via dual-resolution speech representations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16083–16099. Curran Associates, Inc., 2023.

Dehua Tao, Xuan Luo, Daxin Tan, et al. Speech-Omni-Lite: Portable speech interfaces for vision-language models, 2026.

Thinking Machines Lab. Interaction models: A scalable approach to human-AI collaboration. *Thinking Machines Lab: Connectionism*, May 2026.

Justin Uberti and Zahan Malkani. How we built a realtime system for responsive voice AI in six months. OpenAI Blog, August 2026. URL https://openai.com/index/continuous-voice-interaction-with-gpt-live/.

Nik Vaessen. JiWER: Evaluate your speech-to-text system with similarity measures such as word error rate. Python package, 2018. URL https://github.com/jitsi/jiwer.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Chengyi Wang, Sanyuan Chen, Yu Wu, et al. Neural codec language models are zero-shot text to speech synthesizers, 2023a.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024.

Wenhui Wang, Furu Wei, Li Dong, et al. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020.

Xinsheng Wang, Siyuan Feng, Jihua Zhu, Mark Hasegawa-Johnson, and Odette Scharenborg. Show and speak: Directly synthesize spoken description of images. In *2021 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2021 - Proceedings*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pp. 4190–4194. Institute of Electrical and Electronics Engineers Inc., 2021.

Xiong Wang, Yangze Li, Chaoyou Fu, et al. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b.

Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, June 1989. ISSN 0899-7667.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, June 1927. ISSN 1537-274X.

Donghang Wu, Tianyu Zhang, Yuxin Li, et al. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning*, 2024.

Yusong Wu, Ke Chen, Tianyu Zhang, et al. 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)*, pp. 1–5. IEEE, 2023.

Zhifei Xie and Changqiao Wu. Mini-Omni: Language models can hear, talk while thinking in streaming, 2024a.

Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards open-source GPT-4o with vision, speech and duplex capabilities, 2024b.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-Omni technical report, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report, 2025b.

Zhen Ye, Peiwen Sun, Jiahe Lei, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025a.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, et al. Llasa: Scaling train-time and inference-time compute for Llama-based speech synthesis, 2025b.

Taehyung Yu and Seongjae Kang. Best-of-$n$ TTS evaluation is confounded by ASR family alignment, 2026.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, et al. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot, 2024.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. Stream-Omni: Simultaneous multimodal interactions with large language-vision-speech model, 2025.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. SpeechTokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Wei Zhao, Pengxiang Ding, Min Zhang, et al. VLAS: Vision-language-action model with speech instructions for customized robot manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

## A METHOD AND TRAINING DETAILS

Figure 7 maps which corpus produced which result.

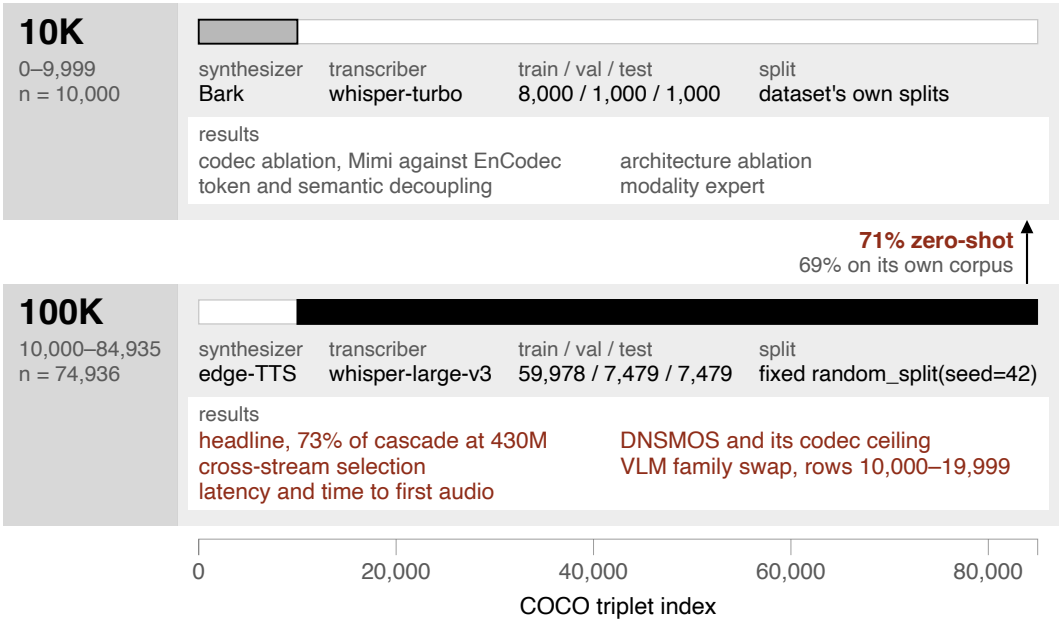



Figure 7: Which corpus produced which result. The two corpora are disjoint; the arrow is the one measured relation, the 430M head trained on edge-TTS audio scored on the Bark corpus it never saw ($n$=200, whisper-turbo).

**Notation.** Let $x$ be an image and $y$ its associated speech waveform. A frozen VLM encoder produces a sequence of hidden states $h(x) = f_\theta(x) \in \mathbb{R}^{L\times d}$ with $d = 1536$. A pretrained neural audio codec maps the waveform to a discrete sequence

$$g_\phi(y) \;=\; \big(\tau_1, \tau_2, \ldots, \tau_T\big), \quad \tau_t \in \{1, \ldots, V\},$$

indexing a codebook $E \in \mathbb{R}^{V\times d_e}$ ($V = 2048$, $d_e = 256$ for Mimi cb0 (Défossez et al., 2024)). The latent $h(x)$ feeds two parallel heads in our dual-stream setup:

$$\begin{aligned} \text{text head:}\;& p_{\text{txt}}(w \mid h(x)), \\ \text{speech head:}\;& p_\psi(\tau_t \mid h(x), \tau_{<t}), \end{aligned}$$

where $w$ is the VLM's native text and $p_{\text{txt}}$ is the VLM's frozen LM head. We train $p_\psi$ only (the VLM is frozen).

**Training details.** To reduce exposure bias we use scheduled sampling (Bengio et al., 2015) with a linear warmup, $\text{ss_ratio}(\text{epoch}) = \min(\text{ss_max}, \text{ss_max} \cdot \text{epoch}/\text{warmup})$, and find $\text{ss_max} = 0.5$ with a 5-epoch warmup optimal. We optimize with AdamW (Loshchilov & Hutter, 2019) at learning rate $1.5 \times 10^{-4}$, weight decay 0.01, batch size 8, with early stopping on validation AR token accuracy at patience 20, and 40 at 100K; Table 8's arms and the 100K runs of seeds 44 and 45 ran at patience 10, and Table 8's select on in-loop sem_sim. At inference, speech candidates are sampled at temperature 1.0 with top-$k$ 50 (Fan et al., 2018), and the text stream is decoded greedily. AR token accuracy is a better proxy for sem_sim peaks than val loss (Appendix C.3).

Each image passes through the frozen VLM once and its last-layer hidden states are cached to disk, which training reads from. Hidden states are deterministic under the frozen VLM and chat template. So the hidden-state cache is equivalent to a live forward pass while removing the VLM cost (∼1.4 s/sample for Qwen2-VL-2B (Wang et al., 2024)) from the training loop. Running the VLM live in the loop, with and without LoRA (Hu et al., 2022), underperformed the cached setup at

10K scale (Table 4). Part of that gap belongs to the 10K cache: we built it with the target summary inside the conditioning prompt (Appendix A), so a head trained on it is off distribution under any live pass that does not carry the summary. The rest is the live-forward regime itself, which lands at the floor whichever head sits on it (Appendix B.5). We built the 100K cache with `"Describe this image."`, and a live forward pass under that prompt reproduces it exactly: cosine $1.000$ between the tensors and $0.4626$ against $0.4626$ sem_sim over 100 test images (173M head). Table 4 keeps the live rows because the failure is silent, lands at the random-token floor, and is the reason to audit a cache before training a head on it.

**Which prompt built the cache, recovered without the images.** A quarter of the spoken targets end mid-clause where the generator hit its length cap, $1{,}970$ of the $7{,}463$ test targets against $5{,}493$ that close cleanly. The speech head reads the two alike (the difference is $+0.0009$ $[-0.0083, +0.0101]$, a bootstrap interval (Efron, 1979) tight enough to be a null rather than a shortage of items). A terminal period is still written after the cut, so a check that asks whether a target ends in punctuation flags only $133$ of the $7{,}463$. Training reads hidden states we cached once per image, so their content is part of the method and not an implementation detail. A cached state is one row per sequence position, and that length is vision tokens plus whatever text the prompt carried. Vision tokens under Qwen2-VL follow the image's pixel dimensions rather than its content, so under a fixed prompt the cached length cannot know anything about that image's summary. Regress length on the summary's own token count and the two hypotheses separate cleanly: a fixed prompt requires slope $0$, and a prompt carrying the summary requires slope $1$, one position per token.

| corpus | $n$ | slope | 95% CI | from 0 | from 1 |
|---|---|---|---|---|---|
| 10K | 10 000 | $+0.911$ | $[+0.554, +1.271]$ | $5.1\sigma$ | $0.5\sigma$ |
| 100K | 74 782 | $+0.118$ | $[+0.025, +0.209]$ | $2.5\sigma$ | $18.8\sigma$ |

The 10K cache carries its own target. The means say it a second way: 10K sequence length averages $417.2$ against the 100K's $376.1$, a shift of $41.1$. The mean 10K summary is $41.1$ tokens; subtract the summary and the 10K distribution lands on the 100K's ($376.1$ against $376.1$, sd $67.6$ against $67.1$). The 100K residual slope of $+0.118$ is one eighth of a position per token and is the signature of a mild content correlation, busier scenes drawing longer summaries (it is $18.8\sigma$ from the $1$ a leak requires). The 100K cache is separately verified against a live forward pass at cosine $1.000$ with identical generated speech.

Two consequences follow. First, we read every 10K row as a matched comparison and claim no absolute 10K value outside B.4's comparison with prior work: the codec swap and the architecture variants each hold the cache fixed across their arms, and the modality expert its live forward, so the conditioning was equally available to both sides and only one side could use it. That is the direction those rows are cited for. The headline, the selection decomposition, the cross-synthesizer transfer and the latency numbers are all 100K and unaffected. Second, the 10K live-inference collapse is not sensitivity to a one-token prompt change: the speech head was trained on states that carried the summary. A live forward pass under `"Describe this image."` does not carry it.

The cause is the prompt, not an accident. Given a summary, our feature extraction builds the prompt as "Read aloud:" followed by the summary, on the reasoning that conditioning on the summary yields representations aligned with the audio target. It produces exactly that alignment, and the alignment is the defect, because the speech head is then asked to speak a summary its conditioning already contains. Every 10K row was extracted with its summary, while the 100K cache was built through the constant prompt. That prompt predicts one position per summary token and we measure $+0.911$, so the regression and the prompt agree on the same mechanism.

Pricing the effect on the original 10K cache needs a head retrained on the clean cache and every 10K arm re-run against it. Those arms are cited only as matched comparisons, which is the one reading that the leak leaves intact, and arms are compared inside one cache and never across the two, which differ in synthesizer as well as in the leak. The audit above needs neither the images nor a GPU. That is why we run it. We would press the check on anyone training a decoder against a cache whose prompt can vary with the target: this failure is hard to see because the reasoning for it reads as a benefit. The length it leaves behind is the only trace it leaves (Figure 8).

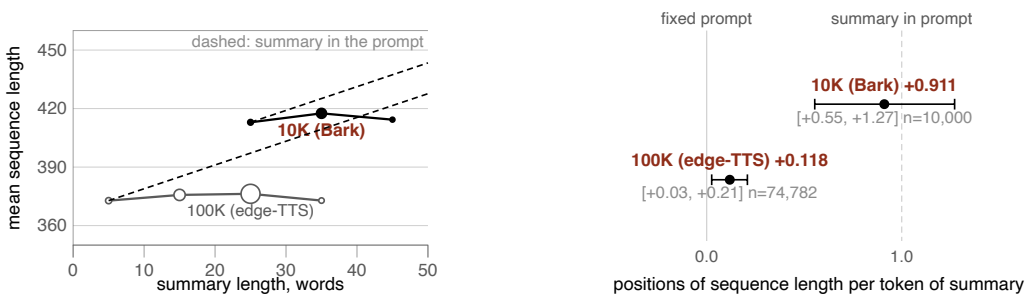



Figure 8: **The 100K conditioning latent carries no text about its own target.** A summary in the prompt would cost one sequence position per summary token, about 1.2 per word, which is the dashed counterfactual. Left: mean sequence length by summary length; marker area is rows per bucket. Right: the fitted slope, in positions per token, against both hypotheses; the 100K cache excludes the contaminated counterfactual and the 10K cache settles the other way.

Table 4: Pre-computed hidden states against a live VLM forward at 10K scale; validation split, $n$=200, whisper-turbo, on the original 10K cache whose prompt carried the summary (Appendix A).

| Mode | % of pre-computed ↑ |
|---|---|
| Pre-computed offline (frozen VLM) | 100 |
| Live forward (frozen VLM) | 11 |
| Live forward (LoRA r=16) | 11 |

# B Evaluation and Ablations

## B.1 Evaluation protocol

Table 1's cells cover the 199 fixed test items except Qwen2.5-Omni's audio and the control (198), the ranked audio (197), and the rendered reference and the coherent column (60); its matched cascade reads sem_sim 0.670.

**Asking the image instead of the reference.** Semantic similarity against a written reference pays for matching one particular wording, and on this corpus that scale is closer to exhausted than it appears. The ground-truth summary and the ground-truth description of the same image, one generated from the other, score 0.700 against each other, against a mismatched floor of 0.190, so a faithful paraphrase tops out near 0.700. On that band the matched cascade sits at 94%, our 430M head at 58%, its completed audio at 76% and the ranked completed draws at 91%, the same order as the percentage of cascade. The instrument also moves under the decoder: the seed-42 checkpoint swings by 0.029 between whisper-large-v3 (Radford et al., 2023) and turbo, wider than the two-seed spread we report as uncertainty.

**The instrument.** We therefore add a reference-free measurement that asks the image directly: each utterance is transcribed once, embedded with CLIP ViT-L/14 (Radford et al., 2021) and matched against its own image and three distractors from the same split (cosine below 0.30, never reused), so chance is 0.250. Our speech identifies its own image in 185 of 199 test items, 0.930 $[0.885, 0.958]$ on a Wilson interval (Wilson, 1927), and most misses are rank two. With the completion stage one completed draw reaches 0.975 and the completed sibling pick 0.990, and the ground-truth summary scores 1.000, a ceiling this instrument actually reaches.

Seed 43, the checkpoint the retrieval rows use, reads 59.8% of the band on semantic similarity and 90.6% on retrieval, and we report both: forced choice is the coarser question, since an utterance that captures one salient object can win it while missing most of the sentence.

A control fixes the reading in both directions. Unconditioned audio scores 0.23 against chance 0.25 ($n$=13), so the instrument returns nothing when there is nothing to find, and the matched cascade scores 1.000 on the same 199 items, so the cascade sits at this instrument's ceiling.

**The same forced choice with no transcript anywhere.** Every arm above transcribes first, so each carries the ASR's error. We therefore repeat the identical forced choice with an omni model, Qwen2.5-Omni-3B, whose Whisper-initialized encoder reads the waveform rather than a transcript, scoring a yes-or-no question as the margin between the two token logits. The instrument passes its own controls: random-token audio sits at $0.283$ against chance $0.250$, and the ground-truth summary rendered to speech reaches $1.000$.

Our speech identifies its own image in $0.759$ of $199$ items $[0.695, 0.813]$, against the matched cascade's $1.000$ $[0.981, 1.000]$. With the completion stage of §5.1 one draw reaches $0.965$ and the completed sibling pick $0.985$, inside the cascade's interval, so the raw first codebook loses acoustics the completion stage restores and not content. Removing the ASR costs us rather than helps us, $66.4\%$ of this instrument's band against $90.6\%$ transcribed, since transcription corrects degraded speech into fluent English that CLIP then reads cleanly.

**The cost of hearing rather than reading.** Giving the same model each item's transcript in place of the waveform moves us from $0.759$ to $0.859$ and leaves the cascade at $1.000$, so hearing rather than reading costs our speech $0.100$ and the cascade's nothing.

### B.2 Controls on the selector and the metric

**The selector's pick, judged by an instrument sharing none of its parts.** Table 3 ranks candidates by a Whisper (Radford et al., 2023) transcript's agreement with the sibling description and then scores the winner the same way, so selector and metric share a checkpoint and an encoder. Letting the image judge the audio removes the ASR and the sentence encoder together; the judge's audio encoder is Whisper-initialized (Appendix B.1), so the Whisper family stays. On a separate $K{=}8$ run of $53$ images, the sibling selector's pick scores $0.868$ against an average draw's $0.778$, a gain of $+0.090$ $[+0.000, +0.175]$ on a bootstrap interval (Efron, 1979), while self-consensus gains $+0.014$, so ranking by the other stream wins here too.

**Images the system was never built from.** The triplets are built from COCO (Lin et al., 2014), so a retrieval score there could be exploiting a scene distribution the speech head was fit to. On NoCaps validation (Agrawal et al., 2019), built around object classes COCO does not cover, the frozen system's speech identifies its image in $0.734$ $[0.668, 0.790]$ against chance $0.250$ ($n{=}199$), and on the reference scale it holds $47\%$ of a NoCaps cascade scored the same way, against $73\%$ in distribution.

The in-distribution $0.930$ is therefore not scene-distribution memorization.

**The strongest input the speech head could be handed.** If the speech head were starved of what the text stream knows, handing it the sentence it is to say would close most of the gap. We train the speech head on the target summary itself as its conditioning input, a copying oracle no deployment could supply, and evaluate it at $n{=}199$. It reaches $87\%$ of the cascade, covering roughly half the distance from our $73\%$ and stopping there.

**How far that gets.** Telling the speech head the very sentence it is to say moves it only halfway, so the remaining distance is not the speech head being told too little about the image.

**What the embedding metric certifies, and what it does not.** Every null above is across items and asks whether the metric sees the image (the within-item null asks whether it sees the sentence). Shuffling the word order of each generated transcript and rescoring it against its own target moves sem_sim from $0.3056$ to $0.2853$, $0.020$ of the $0.181$ that separates matched from mismatched ($n{=}200$, clean 10K arm). Sentence-embedding similarity reports which content is present far more than how it is composed (the embedding metric certifies content, WER certifies composition, and a head at WER $0.91$ that still separates from the mismatched null by $0.439$ is delivering the first without the second).

**A reference-aware metric with no transcript.** SpeechBERTScore (Saeki et al., 2024) matches WavLM-large (Chen et al., 2022) features of the generated and reference audio greedily, and we report the F1 of the matched cosines, with no ASR in the chain. On the test items ($n{=}199$) our 430M audio scores $0.520$ (sd $0.019$), the edge-TTS cascade $0.804$ and the reference summary rendered by the

same synthesizer $0.998$. Its floor, each utterance against another item's reference, is not one number ($0.510$, $0.761$ and $0.761$ respectively). Read as matched minus mismatched, the margin the score assigns to content is $+0.237$ for the rendering, $+0.044$ for the cascade and $+0.010$ for ours, while the transcript-based instruments place the same audio at $74\%$ of the cascade on meaning and at $0.930$ on retrieval. The ordering across arms is carried by voice and channel, which WavLM's features encode and cb0-only decoding does not reproduce, with the content signal compressed underneath it (that is why the paper's content instruments either transcribe or ask the image).

**An off-the-shelf ASR-free metric measures the wrong thing.** Transcription is the noisiest link in the chain, so we tested whether a joint audio-text encoder could remove it. Scored on faithful renderings of the target sentence, CLAP (Wu et al., 2023) gives $0.0399$ on matched pairs against $0.0201$ on deliberately mismatched ones ($d$=$0.18$ and a win rate of $0.50$, which is a coin flip exactly). The encoder registers that a voice is present and not what it says, the expected behavior of an audio tower trained on environmental sound; we report the arm as dropped.

**Token accuracy does not predict meaning at the sample level.** The same decoupling is measurable within one checkpoint. Teacher-forced accuracy on the seed-43 430M head is $0.528$ on cb0 against a unigram baseline of $0.058$, and $0.153$ on cb1–7, as the semantic weight of $100$ against the acoustic $0.02$ predicts. Row by row ($n$=$199$), codebook accuracy correlates with semantic similarity at $r$=$+0.104$ $[-0.056, +0.273]$ for cb0 and $r$=$-0.099$ $[-0.245, +0.083]$ for the acoustic codebooks (Pearson, 1895), both intervals containing zero, so how many tokens a row recovers says nothing about whether it meant the right thing.

### B.3 Fluency and fact units

Table 5: Fluency of the transcripts, content ignored. Coherent: the percentage of items an instruction-tuned judge (Qwen2.5-7B-Instruct; Qwen et al., 2025) answers yes to "is this coherent, grammatical English that a listener would follow as a sentence?", as a share of the cascade's, which it calls coherent on all 199; Wilson 95% interval. log-ppl: mean GPT-2 log-perplexity, bootstrap 95% interval. The 199 fixed test items (60 for the rendered reference), whisper-large-v3 transcripts, the same arms as Table 1.

| Text | coherent, % of cascade ↑ | log-ppl ↓ | words |
|---|---|---|---|
| ***Ceiling: text a person or a cascade wrote*** | | | |
| cascade transcript (caption → TTS → ASR) | 100 | 3.58 [3.52, 3.65] | 17.3 |
| reference summary, written | 84 [79, 89] | 4.17 [4.07, 4.29] | 22.6 |
| reference summary, rendered → ASR ($n$=60) | 85 [74, 92] | 4.17 [3.96, 4.44] | 22.2 |
| ***Monoplex, and what selection does to fluency*** | | | |
| monoplex, 430M head, one draw | 27 [21, 33] | 5.04 [4.94, 5.13] | 23.5 |
| + acoustic completion | 33 [27, 40] | 5.09 [4.97, 5.21] | 22.9 |
| + ranked by sibling text and agreement | 37 [31, 44] | 4.85 [4.75, 4.96] | 22.9 |
| + a fluency judge in the vote (content 96%; seed 42: 69, 96%) | 66 [59, 72] | 4.65 | |
| ranked by the judge alone (content 89%; seed 42: 86, 88%) | 84 [78, 88] | 4.57 | |

**What the semantic metric does not see.** Semantic similarity credits the nouns that arrive whether or not the sentence around them holds, so two instruments that ignore content score the same transcripts in Table 5: the share an instruction-tuned judge (Zheng et al., 2023) calls coherent English, and GPT-2 (Radford et al., 2019) perplexity.

**The judge's ceiling.** The judge is strict on the reference's own register, a long noun-phrase sentence of twenty-two words, which it accepts $0.84$ of the time written and $0.85$ rendered and transcribed, so that pair is the ceiling of the scale and the cascade's shorter captions sit above it. The completed stream is judged coherent in $0.33$ of items and the ranked stream in $0.37$, intervals overlapping, and perplexity separates the same pair, $5.09$ against $4.85$. The raw first codebook reads $0.27$, so the completion stage recovers a little intelligibility through the ASR, while the ranking by content agreement recovers little. Read beside Table 1, that is the split the semantic instrument was chosen to make visible: the stream carries the image's content to $86\%$ of the cascade and identifies the image in $0.97$ of items, in sentences that hold together two fifths as often as the reference's.

**Fluency as a third vote.** The eight draws already contain fluent sentences, and a judge that reads only the transcript finds them: its yes-over-no margin as a third signal in the vote, standardized and summed with the sibling and agreement scores, lifts the pick's coherence from $0.37$ to $0.66$ on seed 43 and from $0.48$ to $0.69$ on seed 42, at a content cost of $0.009$ and $0.008$ ($0.654$ to $0.645$ and $0.646$), while the judge alone reaches the reference's own rate at a content cost of $0.06$: over the same candidates content and fluency are a front, not a trade.

**A second judge.** A judge from another family, Mistral-7B-Instruct (Jiang et al., 2023), reads the same 199 transcripts per arm more leniently: $0.51$ on the raw head, $0.60$ on the completed stream and $0.76$ on the ranked stream, against $0.98$ for the rendered reference. Unlike the first judge, which reads completion and ranking as tied ($0.33$ and $0.37$), it separates the ranked stream from the raw head on Wilson intervals (Wilson, 1927). The second judge also grades the two picks the first judge made, so the model that selects is not the model that scores: over the eight candidates of seed 43 it reads the vote's pick at $0.86$ $[0.80, 0.90]$ and the judge-alone pick at $0.92$ $[0.87, 0.95]$, against $0.76$ $[0.69, 0.81]$ for the sibling-and-agreement pick on those same candidates, so the ordering the selecting judge reports is not an artifact of its own grading.

Table 6: Fact-unit recall. Each reference summary is split into checkable units by Qwen2.5-7B-Instruct ($6.3$ per item), and the same judge reads every arm's transcript against each unit; recall, the share of units the transcript states, is given as a percentage of the cascade's, with paired bootstrap 95% intervals over items. 198 of the 199 fixed test items (one target yields no unit), whisper-large-v3 transcripts, the arms of Table 5.

| Transcript | recall, % of cascade ↑ |
|---|---|
| cascade (caption → TTS → ASR) | 100 |
| monoplex, 430M head, one draw | 47 [40, 54] |
| + acoustic completion | 65 [58, 74] |
| + ranked by sibling text and agreement | 74 [66, 83] |
| + a fluency judge in the vote | 74 [65, 83] |

**Fact by fact.** Semantic similarity credits content as a whole, so Table 6 breaks it down one unit at a time. The judge of Table 5 splits each reference summary into checkable units (Nenkova & Passonneau, 2004; Min et al., 2023), $6.3$ per item, and reads every arm's transcript against each unit, asking whether the transcript states it. The order of Table 1 holds except between the ranked stream and the fluency vote's pick (the raw head states $0.14$ of the units, the completed draw $0.20$, the ranked stream $0.22$ and the fluency vote's pick $0.22$, with the cascade's transcript at $0.30$). The instrument is stricter than similarity, and stricter for every arm at once (the cascade's caption states two of the summary's six units). On a paired bootstrap over items (Efron, 1979), the whole pipeline states $+0.08$ $[+0.06, +0.11]$ more of the units than the raw head does, and the completion stage alone $+0.06$ $[+0.03, +0.08]$ (the fluency vote's pick gives none of that back, $0.22$ against the ranked stream's $0.22$, so the judge in the vote is choosing sentences that say as much of the target at the similarity cost of $0.01$ that Table 5 shows).

### B.4 The cascade and latency

The cascade's caption prompt, shared with the streaming row, is "Describe this image in one short sentence."; our own conditioning prompt is "Describe this image." (§3.3). The cascade's speech therefore renders a one-sentence caption and not the long description the summary was written from. The cascade row's content scores are edge-TTS's; the same captions streamed through Kokoro, timed and scored end to end, read $0.667$ against edge-TTS's $0.670$ ($n$=199), so the two synthesizers agree on content to $0.003$ and every percentage of the cascade holds against either.

**The cascade prompted for the summary's form.** Prompting the cascade instead for a spoken sentence of about twenty words, the length of the summary target, moves it from $0.670$ to $0.697$ ($n$=199); the completed and ranked arms of §5.5 read $83\%$ and $94\%$ of that row, so the split does not turn on the cascade's prompt. Every other comparator, the streaming cascade, the cascade given eight

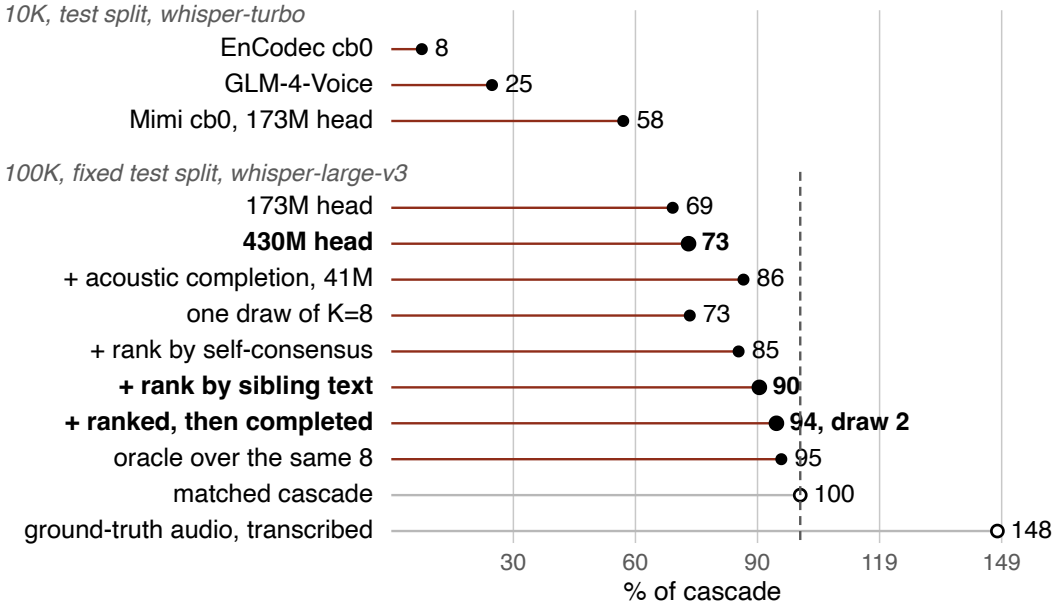



Figure 9: **Every one-draw and ranked sem_sim in the paper on one axis.** Blocks differ in transcriber and split, so compare within a block; percentages are of that block's cascade. Ranking arms select among one checkpoint's $K{=}8$ draws; completion arms use the completion stage (Appendix C.4). Table 1's $98\%$, which completes the draws before ranking them, is not drawn.

draws and the omni model, answers the one-short-sentence prompt, so the percentages in the body are taken against it.

**What the omni comparator's two streams carry.** Qwen2.5-Omni-3B (Xu et al., 2025a) is the closest deployed arrangement to ours at our VLM's scale; Qwen3-Omni's Talker reads text tokens and multimodal features, and its released backbone is a 30B mixture. Its speech and its own emitted text agree at cosine $0.98$ ($n{=}199$, $88\%$ of items above $0.95$): the Talker renders the Thinker's sentence, which is the text-on-the-path case of §2. Its delivery is conversational, and $98\%$ of its utterances close with a question the reference never contains. With that sentence removed it reaches $0.668$ against the cascade's $0.670$, so the gap between the rows is speaking style rather than content. We report the number as spoken, since that is what a listener receives.

On one idle GPU our request totals 0.80 s (5-run medians): VLM prefill 0.06 s, speech-head autoregressive decode 0.73 s (110 frames at 12.5 Hz, the dataset's median utterance, no KV cache) and Mimi (Défossez et al., 2024) decode 0.01 s. The cascade's caption decode alone is 0.56 s (median over the 199 test items, greedy, the one-sentence prompt), and with the same GPU synthesizer its full request is the streaming run's 0.77 s, the same synthesis stages in a different order. The asymmetry is where the first audio comes from: that cascade decodes its whole caption before synthesis can start, but monoplex needs only a prefill, so the VLM term drops from 0.56 s to 0.06 s and the two stages stop being serial.

Time to first audio was swept over the chunk size at which the first codec frames are handed to Mimi, the only free parameter on our side: one frame (0.08 s of audio) gives 0.07 s, thirteen (1.04 s, the row we report) 0.12 s, twenty-five (2.00 s) 0.17 s. The caption-complete cascade has no such parameter, since its caption must finish before any audio exists. That is why 0.56 s is a floor for that form rather than a measurement of any synthesizer.

**What a faster synthesizer would not remove.** A faster TTS backend such as SupertonicTTS (Kim et al., 2025) would cut synthesis cost, but no backend removes the structural one: synthesis cannot begin until its caption is complete, so a caption-complete cascade's latency sums two sequential stages. Latent-conditioned generation removes that barrier, since the speech head emits codec tokens




Figure 10: **Time to first audio is a design choice on the direct path and a floor on the cascade.** Left: buffering more audio before our first chunk moves the wait, with the model identical in each of our rows, beside the cascade's streaming and caption-complete forms. Right: on the same GPU no synthesizer after a caption decode can start before $0.56$ s; the streaming one starts at $0.32$ s.

from the forward pass the text head reads, and it reopens the non-verbal channels text serialization discards. Both figures are model-side. Deployed voice systems report transport and session setup contributing comparably to perceived responsiveness (Uberti & Malkani, 2026), so what transfers across deployments is the removed barrier rather than the absolute seconds (Figure 10).

**The cascade given a streaming synthesizer.** We give the cascade the strongest streaming form that runs locally: the same Qwen2-VL-2B streams its caption token by token, every six words (or a clause boundary after three) go to Kokoro-82M (hexgrad, 2025) on the GPU as they arrive, and the chunks' audio is concatenated and scored under the protocol of Table 1. Over the $199$ test items the median first audio is $0.32$ s and the median request $0.77$ s (Table 1), at sem_sim $0.667$ against $0.670$ for the non-streaming cascade on the same items. So streaming buys the cascade $0.25$ s to first audio; the caption-complete cascade's request was not timed. What remains is structural: our first audio needs no text tokens and no second model in front of it, and a cascade closes to within a factor of $2.6$ only by adding a streaming synthesizer (Dekel et al., 2024; Du et al., 2024), still rendering one content stream.

**The coupled arm and the cascade on one device.** Figure 6 prices the coupled arm on one GPU and Table 1 times the cascade on another, so the two were never on one clock. On a third GPU that runs both, one draw of the 430M head takes $0.47$ s, the streaming cascade of Appendix B.4 with Kokoro on the same GPU takes $0.40$ s per request and $0.18$ s to first audio, and the $K{=}8$ ranking arm takes $7.7$ s with every stage serial and $6.5$ s with its text and speech streams overlapped (5-run medians, first image discarded as warmup, $n{=}9$ and 10). The draw decomposes into prefill $0.04$, decode $0.42$ and Mimi $0.01$; the coupled arm into prefill $0.04$, the text stream $1.43$, the eight draws batched $1.23$ and their eight transcriptions $4.96$. The frame rate of the target is priced the same way, since it is the reason Mimi is kept over the SpeechTokenizer arm that reads higher on content. Holding the head, the prompt and the device fixed and varying only the number of frames decoded, $110$ frames cost $2.732$ s of autoregressive decode and $440$ cost $27.201$ s, against prefill of $1.34$ to $1.38$ s either way (5-run medians, 233M head). Four times the frames is therefore about ten times the decode rather than four, because the decode carries no KV cache and recomputes its prefix at every step, so a $50$ Hz target pays quadratically for the same seconds of audio. The arm is therefore $14\times$ the head's own single draw and $16\times$ the cascade's request on the same GPU, against $24\times$ on the GPU of Figure 6, and two thirds of the serial $7.7$ s is the transcription the selector reads, which is the cost a transcript-free selector would remove.

**Comparison to prior direct image-to-speech work.** The closest prior system, Hsu et al. (2021b), trains on SpokenCOCO (human read-alouds of the COCO 5-caption text) and produces caption-like speech. For evaluation, they ASR-transcribe their model's speech and compute BLEU-4, METEOR, and ROUGE-L against the original written COCO (Lin et al., 2014) 5-caption set (their Table 4 caption: "ASR is used to transcribe the spoken captions generated by the proposed VQ3 model into text for evaluation"); their reported numbers are $0.233$ / $0.212$ / $0.478$. A fair comparison needs two adjustments. First, the relevant reference set is the written COCO 5-captions (the Karpathy split has the full mapping per image), not the SpokenCOCO audio: SpokenCOCO is the training corpus, the

COCO captions are the evaluation reference. Second, our model targets a different output distribution: Hsu et al. regress to verbatim COCO captions; we regress to Qwen2.5-summarized SST text, so word-level overlap with COCO captions is not what our outputs aim to produce.

Under the same evaluation pipeline, with Whisper transcribing our outputs in place of their ASR and scoring against multi-reference COCO 5-caption text from the Karpathy split (Karpathy & Fei-Fei, 2015) ($n$=1000, seed-44 checkpoint), we score the first 80 characters of each transcript at BLEU-4 $0.008$ / METEOR $0.154$ / ROUGE-L $0.196$. These are not a matched comparison to either prior system, and this submission contains none; the matched comparator throughout is the cascade of Table 1 on the same corpus and target. The same prefixes scored against the first 80 characters of our training target (single-reference SST text) give BLEU-4 $0.035$ / METEOR $0.167$ / ROUGE-L $0.222$. Kim et al. (2024) report caption-target numbers under this protocol from a text-pretrained decoder over HuBERT units; ours is a summary-target system, so the drop on every metric when switching from SST-reference to COCO-reference is a measurement of the task-target distance, not of model quality: a matched-target replication (training a Hsu-style pipeline on our SST data, or training our pipeline on SpokenCOCO captions) is future work.

**Within-task cascade baseline.** To isolate task-target distance from model quality, we run the same evaluation on a matched cascade (Qwen2-VL-2B caption → Bark TTS → whisper-turbo ASR), scored against single-reference SST text. The cascade gives BLEU-4 (Papineni et al., 2002) $0.081$ / METEOR (Banerjee & Lavie, 2005) $\mathbf{0.278}$ / ROUGE-L (Lin, 2004) $0.346$ ($n$=200, test split). The comparison supports three readings. (i) On those prefixes our direct model reaches 60% of the cascade's METEOR on whole texts ($0.167/0.278$), with first audio $4.7\times$ earlier ($0.12$ s against the $0.56$ s caption floor, Figure 10, right). (ii) Hsu et al.'s METEOR ($0.212$) comes from a different system, evaluation sample and reference set (COCO 5-cap), so we report it for context rather than as a matched comparison. (iii) The 40% METEOR gap from cascade is a latency-quality tradeoff that the paper's main figure already characterizes via sem_sim; reporting the same gap on string metrics makes the tradeoff visible at the level of n-gram overlap. The within-system metric on the same Whisper transcripts and the same training target is sem_sim ($64\%$ of the matched cascade for the 10K seed-44 run at $n$=1000, the best single seed of Table 10's 10K block).

**The string instruments on the 100K arms.** The paragraph above is the 10K protocol, whose cascade runs through Bark and whisper-turbo, so its $0.278$ is not the denominator for the 100K arms and its $60\%$ is the raw cb0 arm rather than either arm the paper leads with. We transcribed the matched cascade of Table 1 over the same $199$ items under the same protocol and read METEOR $0.263$ and ROUGE-L $0.367$ from it; that run's sem_sim is $0.6696$ against the $0.6696$ Table 1 already cites, and every item pairs to a selection record by target text, so the two are the same item set. Selection arms are scored on the draw each rule keeps, recovered from the stored candidates; re-embedding those transcripts returns the recorded sem_sim to $10^{-6}$ on both seeds, which is what says the stored text is the text those scores were taken on. Against that denominator, over the two training seeds: the completed draw reads $84.5$ and $87.7\%$ of the cascade on METEOR, the ranked stream $94.1$ and $100.9\%$, and the eight-draw oracle $102.2$ and $109.1\%$. On ROUGE-L the same arms read $78.3$ and $79.7$, $84.7$ and $87.3$, and $90.3$ and $93.4\%$. The METEOR readings sit within a point of the sem_sim readings on the same picks, $85.7$ for the completed draw and $97.7$ for the ranked stream, so an $n$-gram measure with stemming and a fragmentation penalty returns what the embedding measure returns on the arm the paper leads with.

### B.5 Architecture, backbone and data

**Architecture ablation.** Codec choice is necessary for the alignment. We next check whether it is also the dominant lever, by asking whether architecture changes move sem_sim by anything close to the same margin.

Causal AR with shifted teacher forcing (Williams & Zipser, 1989) dominates every variant trained on the cache. Adding cross-attention to VLM hidden states does not improve over the pooled-context causal head and increases parameter count without benefit (Figure 11). The four heads differ in architecture and not only in size, so that axis ranks variants rather than tracing a capacity ladder. The open points, the modality-expert runs, are compared against the cached rows under a confound the next paragraph scopes.

Table 7: Architecture ablation: $n$=200, whisper-turbo, 10K Mimi triplets; 3 seeds for the causal head, one for the rest. The cached rows are the test split under one matched protocol; the two expert rows run a live VLM forward and are read beside their own control in that regime (Appendix B.5); the last column reruns the arm on the clean 10K subset of Appendix A under the external protocol.

| Speech head | Params | % of cascade ↑ | clean cache |
|---|---|---|---|
| **Causal AR (ours)** | 173M | $\mathbf{58 \pm 5}$ | $\mathbf{48 \pm 3}$ |
| Cross-attention + depth | 268M | 14 | 9 |
| Dense expert, Qwen3-VL-2B | 187M | 9 | |
| MoE expert, Qwen3-VL-30B-A3B | 304M | 9 | |

Rerun on the clean cache of Appendix A under the external protocol, cross-attention holds 9% of the matched cascade against the causal head's $48 \pm 3$% over three seeds, so the ordering is not a property of the original cache. That arm never fit its objective: its best validation loss is 6.05 against 7.62 for random guessing over the 2048-entry codebook, where causal arms on the same cache sit at 2.56 and 3.29. The EnCodec (Défossez et al., 2023) arm fit its objective at 3.29 and still reads only 16%. Token accuracy separates none of them, reading between 6.7 and 8.0% for the cross-attention arm of Table 2 and for the Mimi and EnCodec causal arms of Table 9 alike. These are clean-cache arms, and the decoupling of §5.2 holds on both caches.

Retrained with uniform attention, across a 33× learning-rate span, and until it memorized its own training items, the arm stays at the floor, 7 to 10% of the matched cascade against 8% for random-token audio, on test and on the items it trained on alike.

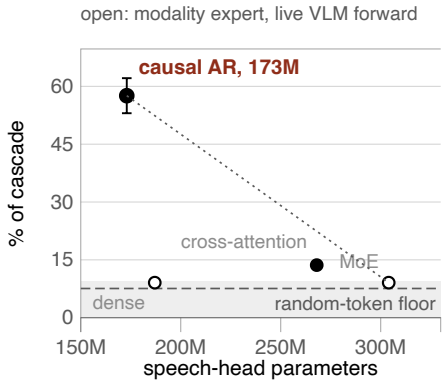



Figure 11: **The plain causal head beats every more elaborate variant trained on the cache.** The rows of Table 7; intervals are the seed standard deviation, open points ran a live VLM forward.

**Modality expert.** Cross-attention added parameters without benefit. A more capacity-heavy alternative is to put the speech expert inside the VLM itself, which we tested next: a modality-expert architecture inspired by MoST (Lou et al., 2026), with dedicated speech experts inside the VLM and attention shared between text and speech tokens.

**Both routing variants.** Dense (Qwen3-VL-2B with token-type routing) and mixture-of-experts (Qwen3-VL-30B-A3B with extended router) variants both reach $\sim$ 9% of the matched cascade. These variants require a live VLM forward pass. Control runs show the collapse tracks that training regime, not the routing architecture. Under the same live-VLM (LoRA; Hu et al., 2022) regime a plain causal head also reaches 9% and a cross-attention head 8% on validation. We did not carry these runs to test evaluation, since validation was already at the random floor. The comparison between the in-VLM and separate-head designs is therefore regime-confounded, and the modality-expert collapse is scoped to that training regime. The cached-regime ablation (Table 7) supports the claim that added attention capacity does not help, since cross-attention (14%) loses to the pooled causal head (58%) under matched training, consistent with the converged separate-decoder design across VLA ($\pi_0$, SmolVLA (Black et al., 2025; Shukor et al., 2025)) and speech generation (Moshi's Depth Transformer (Défossez et al., 2024), LLaMA-Omni (Fang et al., 2025)).

**Is the effect specific to one VLM family?** Every separate-head result so far conditions on one VLM, so the effect could be a property of Qwen2-VL and not of the architecture. We therefore swap the conditioning VLM and nothing else (Table 8): the same head on the same 10,000 triplets under the same fixed split, differing only in which VLM produced the cached hidden states. Three arms are consecutive Qwen-VL generations (Wang et al., 2024; Bai et al., 2025b;a) and one, SmolVLM2-2.2B (Marafioti et al., 2025), shares neither the vision encoder nor the language

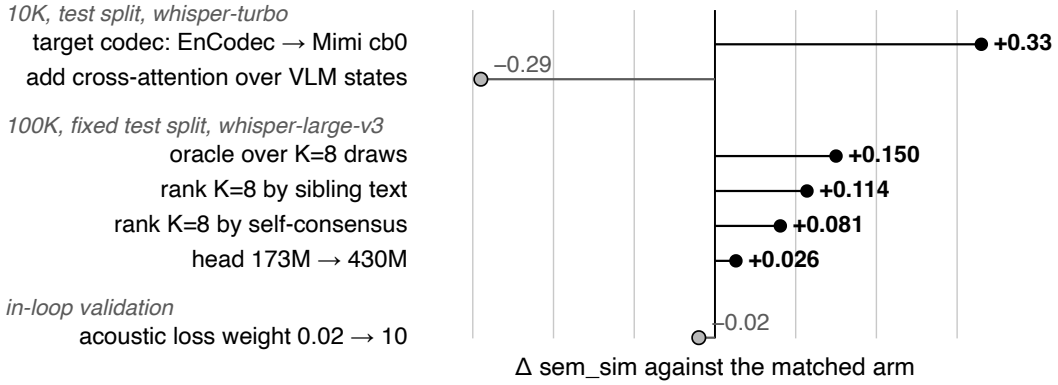



Figure 12: **Every matched comparison, signed.** Each row is a pair of arms differing in the named axis and in nothing else (the capacity row also retunes the optimizer, and one acoustic arm stopped early); rows are grouped by protocol because the paper does not difference across protocols. Within its protocol the target codec moves sem_sim further than any other axis, while the two levers one would reach for first, more parameters and more attention, are the smallest and the most negative.

Table 8: Conditioning VLM swapped, everything else held fixed: the same speech head ($d_{\text{model}}$=1024, 12 layers, 16 heads, cb=8), the same 10,000 triplets, the same Mimi cb0 targets, the same fixed seed-42 random split, the same external protocol (test split, $n$=200, whisper-large-v3, all-MiniLM-L6-v2), one seed per arm, patience 10. Conditioning tokens are the median cached sequence length over 200 items; sem_sim is measured against the control's.

| Conditioning VLM | Vision + language stack | $d$ | Cond. tokens | Δ sem_sim |
|---|---|---|---|---|
| SmolVLM2-2.2B | SigLIP + SmolLM2 | 2048 | 126 | +0.039 |
| Qwen3-VL-4B | Qwen, 3rd generation | 2560 | 274 | +0.028 |
| Qwen2.5-VL-3B | Qwen, 2.5 generation | 2048 | 370 | +0.013 |
| Qwen2-VL-2B (control) | Qwen, 2nd generation | 1536 | 370 | — |

backbone. The Qwen2-VL-2B arm is the control, on the same triplet set and tokenization as the others.

Under the external protocol (fixed test split, $n$=200, whisper-large-v3, identical across arms) every conditioner sustains the effect, the one furthest from the original setup included; every arm scores at or above the control, two of the four inside the control's own three-seed spread of $\pm 0.02$, Qwen3-VL-4B and SmolVLM2 beyond it, so the column is a band and not a ranking. The effect is not confined to the Qwen family.

All arms stopped on the same patience, so each is a lower bound on its own peak, and none is read against the 10K row of Table 10, which uses a different triplet set.

The clean cache behind this block is a 10K subset of the 100K corpus cached with the fixed prompt of §3.3; it overlaps the 100K test split, so every arm here is scored on the subset's own test split. The SpeechTokenizer (Zhang et al., 2024) arm on that cache tokenizes the same audio with the released 8-codebook, 1024-way model at 50 Hz and 16 kHz, keeps 800 frames (16 s) as EnCodec's arm keeps 1200, and changes nothing else in the recipe.

**Every comparison on one axis.** Figure 12 collects every matched comparison in the paper on one signed axis.

Swapping Mimi for GLM-4-Voice (Zeng et al., 2024) (single-codebook, 16384-vocab, Whisper-large-v3-derived) under a matched 3-seed training protocol gives $25 \pm 3\%$ of the matched cascade ($-57\%$ vs. Mimi mean; per-seed 26, 28, 21). That is $\sim 3.1\times$ above the EnCodec random-token floor of $8\%$ but $\sim 2.3\times$ below Mimi cb0. Its evaluation is not matched to the Mimi arms: its evaluation takes a positional last-10% slice of the triplets rather than the fixed seed-42 random partition

Table 9: Codec ablation: the same causal head, data and hyperparameters, only the codec target differing. Test split, $n$=200 per seed, whisper-turbo; GLM-4-Voice's evaluation slice overlaps its training data, so its number is an upper bound (Appendix B.5). Last column: the clean 10K subset (Appendix A), fixed test split, $n$=200, whisper-large-v3.

| Codec | first codebook | vocab | % of cascade ↑ | clean cache |
|---|---|---|---|---|
| Mimi (3 seeds) | WavLM-distilled | 2048 | $\mathbf{58 \pm 5}$ | $\mathbf{48 \pm 3}$ |
| Mimi (best, seed 44) | | 2048 | 63 | |
| SpeechTokenizer (2 seeds) | HuBERT-distilled | 1024 | | $60 \pm 5$ |
| GLM-4-Voice (3 seeds) | Whisper-derived | 16384 | $25 \pm 3$ | |
| EnCodec | acoustic | 1024 | 8 | 16 |

training used, so the GLM test set overlaps its own training data and 25% is an upper bound on what that arm holds out. The direction favors GLM, which is the arm that loses.

The comparison is confound-bound in the codec as well, which is why Figure 13 draws it open: GLM-4-Voice differs from Mimi cb0 in vocabulary size (16384 vs. 2048), codebook count (1 vs. 8), and first-codebook source (Whisper vs. WavLM) simultaneously, and a regularized re-run (lr 1e-4, dropout 0.2, label smoothing 0.05) produces the same number, so the result is not an optimizer artifact. The causal claim rests on the Mimi–EnCodec contrast; the GLM row is a confound-bounded observation that swapping in a newer single-codebook semantic codec at this scale does not improve over Mimi cb0; its seed range ($[21, 28]\%$) is well separated from both the Mimi range ($[53, 63]\%$) and the EnCodec floor (Table 9).

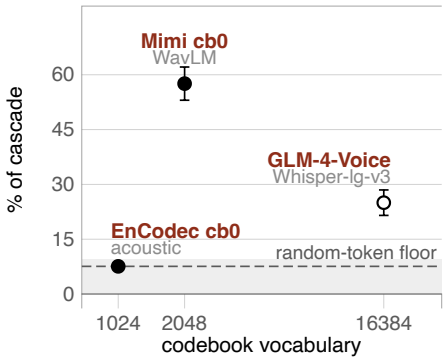



Figure 13: The original-cache rows of Table 9: what the first codebook was distilled or derived from, not how many codes it has. GLM-4-Voice is drawn open.

**The decode path, separated from the head.** A floor on an acoustic target could be EnCodec's first codebook carrying too little to transcribe, in which case no head could do better. We round-trip the reference audio of the 100K test split through each codec at a fixed number of codebooks and score the result under the paper's protocol (whisper-large-v3, $n$=60): EnCodec's first codebook alone transcribes at sem_sim 0.922 (sd 0.133), its first two at 0.990 and all eight at 0.994, so a head that reproduced EnCodec's first codebook would sit near the ceiling, and the 8% of Table 9 is the head failing to learn that target rather than the target failing to carry the sentence. Mimi's first codebook alone round-trips the same items at 0.769 (sd 0.205) and all eight at 0.991, so both decode paths carry the sentence and the acoustic one carries it better; the distance between 8% and 58% is what the head can learn from the target, not what the target can say. EnCodec runs at 75 Hz against Mimi's 12.5 Hz, so its first codebook is six times the sequence length under the same head. The SpeechTokenizer arm of §5.3 separates length from organization on one cache: at 50 Hz it carries four times Mimi's sequence length, within $1.5\times$ of EnCodec's, and reads $60 \pm 5\%$ of the matched cascade where Mimi reads $48 \pm 3\%$ and EnCodec reads 16%. The longer sequence is the one that reads higher, so length is not what holds EnCodec at 16%. What the round-trip control settles is separate: the decode path is not what holds it there either. The same contrast on the clean cache of Appendix A orders the two arms the same way under the external protocol (Table 9, last column), so the gap is not a property of the original 10K cache either.

**What Table 10 holds fixed.** The 10K rows and the 100K 173M row share one 173M head (12 layers, $d_{\text{model}} = 1024$, Mimi vocab 2048), decoded at cb0 above the completion stage. The 100K head predicts eight codebooks, but cb1-7 train at weight 0.02 and decoding the same tokens at one, two or eight returns the same audio (Appendix C.3), so the rows differ in data and protocol but not architecture. Protocols differ across scales (transcriber and split machinery), so rows are comparable

Table 10: Data and head-size scaling on Mimi. 10K rows: Bark, whisper-turbo, the dataset's own splits. 100K rows: edge-TTS, fixed seed-42 random test split, whisper-large-v3, $n$=199, 2 seeds per condition, checkpoints selected on validation AR token accuracy.

| Data ($\rightarrow$ TTS) | cb | head | split | % of cascade | note |
|---|---|---|---|---|---|
| 10K Bark | 8 | 173M | val | 68 | ss=0.5, ep30 peak |
| 10K Bark | 8 | 173M | val | $71 \pm 5$ | 3-seed mean |
| 10K Bark | 8 | 173M | test | 53 / **63** | seed-42 / best (seed-44) |
| 10K Bark | 8 | 173M | test | $58 \pm 5$ | 3-seed mean |
| 10K Bark | n/a | n/a | test | 100 | matched cascade (Qwen2-VL-2B $\rightarrow$ Bark) |
| 100K edge-TTS | 8 | 173M | fixed test | $69 \pm 3$ | 2 seeds |
| 100K edge-TTS | 8 | 430M | fixed test | $\mathbf{73 \pm 2}$ | 2 seeds |
| 100K edge-TTS | n/a | n/a | fixed test | 100 | matched cascade (same split) |
| 100K edge-TTS | n/a | n/a | fixed test | 148 | GT-audio transcription ceiling |

within a scale and not differenced across scales (Figure 9). The two 100K rows train *identical* codebook settings (8 codebooks, semantic weight 100, acoustic weight 0.02), are decoded identically and use the same batch size, so the 69% to 73% gap is not a codebook-count effect. That step also lowered the learning rate to $1 \times 10^{-4}$ and doubled warmup to 1000 steps, so it is read as a recipe change and not as a capacity effect. Checkpoints are selected on validation AR token accuracy rather than validation loss, which Appendix C.3 shows selects the wrong epoch for this target.

**Checkpoint selection at 100K.** Checkpoints are the validation-selected checkpoint for every cell of Table 10. The final-epoch checkpoint scores higher on test in three of the four cells (up to $+0.039$; best single run 0.513), and we do not report those numbers, since choosing between checkpoints on the test split is selection on the evaluation set. Read within its own protocol, $7.5\times$ training data moves the 173M head from 58% to 69% of its matched cascade.

**The synthesizer changes with the volume, and the speech head does not notice.** The 10K and 100K corpora differ in synthesizer as well as size, so the comparison carries that confound. We bound it by transfer: the 430M checkpoint trained on edge-TTS audio alone reaches 71% of the matched cascade on the 10K Bark corpus, against 69% on its own test split at the same checkpoint and transcriber ($n$=200, whisper-turbo). The two corpora hold disjoint triplet indices, so this is held out by construction rather than by a split, and it scores marginally higher on the corpus it never saw. One direction, one checkpoint: this bounds the synthesizer's share of the scaling delta rather than eliminating it. Every 100K number in the paper is under the fixed split.

**The headline cell against the whole test split.** Every headline cell is 199 items of a 7,479-item test split, so the reported number could be a draw rather than a level. We ran the seed-43 checkpoint over the entire split under the same protocol, whisper-large-v3 and the same generation settings, and it reads 72.5% of the matched cascade $[71.9, 73.1]$ over the 7,463 items scored, against the seed-43 cell's 74 $[70.2, 77.7]$. The full-split value sits inside the cell's interval, so the two agree at $38\times$ the sample. The spread a 199-item cell can take is worth stating beside it: resampling 199 items from the split gives $[68.7, 76.3]$%, and the reported cell sits near the middle of that rather than at an edge. Over the full split 5.5% of transcripts end without terminal punctuation.

# C Analysis

## C.1 What $K$ buys, and what selection costs

**Controls on the cross-stream term.** The metric's own permutation null (Ernst, 2004) rules out a scorer that cannot discriminate at all: scoring each clip against another image's target collapses sem_sim by 0.309 to the null's 0.130 and 96% of images rank their own target first ($n$=50, Figure 14, left). Two confounds could still manufacture the cross-stream term specifically: target and selector share a Qwen ancestor, and selector and evaluator share an ASR (Yu & Kang, 2026) and a sentence encoder. Against five human COCO (Lin et al., 2014) captions, over all 1,993 images, the whole decomposition holds: one draw reads 0.563, agreement among the draws 0.615 ($+0.052$ [$+0.047$,

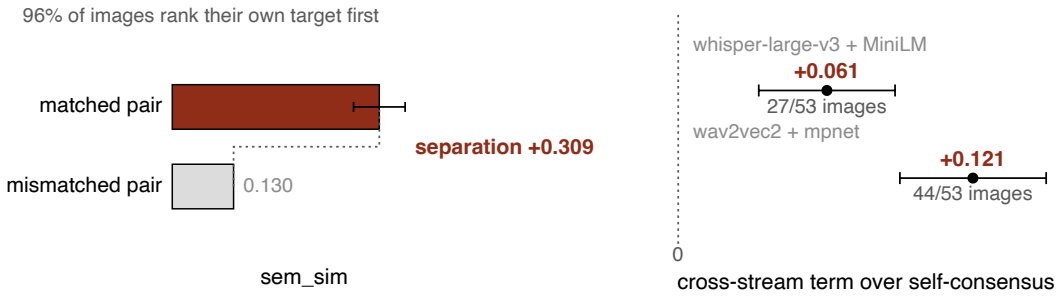



Figure 14: **The two controls the metric depends on.** Left: the permutation null ($n$=50); scoring each clip against another image's target collapses sem_sim to the null's 0.130, and 96% of images rank their own target first. Right: the lineage control; rescoring identical draws through an unrelated ASR and a second encoder leaves the cross-stream term clear of zero. Intervals are on paired differences.

+0.057]) and the sibling text 0.630 (+0.067 [+0.062, +0.072]) against an oracle's 0.674. Hence the cross-stream term over agreement alone is +0.015 [+0.011, +0.018], higher on 813 of 1993 images ($n$=1993). The ancestor inflates that term rather than creating it: it reads +0.034 against the descended target and +0.015 against a human-written one. Rescoring identical draws through wav2vec2-CTC (Baevski et al., 2020; Graves et al., 2006), which shares no lineage with the selecting pair, and a second encoder, MPNet (Song et al., 2020), gives +0.121 [+0.091, +0.152] on 44 of 53 against +0.061 there, on paired differences rather than a shared scale (Figure 14); repeated over 1,323 images it reads +0.120 against +0.043 (Appendix C.1). The pipeline ceiling (GT-SST → edge-TTS → Whisper) bounds the setup at 0.993.

**What $K$ buys, and why 8.** The selection run stored every candidate's similarity to its target, so the best-of-$K$ curve for any $K$ up to 8 needs no new experiment: we enumerate every $K$-subset of an image's 8 draws and take the best in each. At $K$=8 it returns the 95% of Table 3, and at $K$=1 the mean over all eight draws, 72.7%, against the 73.0% of the first draw alone.

| $K$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| best of $K$, % of cascade | 72.7 | 82.5 | 87.0 | 89.8 | 91.8 | 93.3 | 94.5 | 95.4 |
| gain from draw $K$ | — | +9.8 | +4.5 | +2.8 | +2.0 | +1.5 | +1.2 | +0.9 |
| images ≥ cascade | 17% | 25% | 31% | 35% | 39% | 41% | 43% | 45% |

The first row is where §5.5's "95% of the matched cascade" comes from. The last counts images whose best draw meets or beats the cascade outright, and it travels further, from one image in six to nearly one in two, so most of what $K$ buys is images crossing the line rather than the average creeping toward it.

The gain shrinks with every draw: $K$=4 takes 75% of what $K$=8 takes for half the transcription bill, and the eighth draw buys a tenth of what the second did. $K$=8 therefore sits past the knee, and at this decay a further doubling is a worse trade than the one Appendix B.4 prices, where selection costs 24× the draw it improves.

**Does the cross-stream margin need $K$ to be small?** Both selectors rank candidates rather than score them, so their $K$ curves need every candidate's ranking score and not only the winner's. The run behind Table 3 kept those, so the arms come out of the same run by the same enumeration, each selector picking inside each subset.

| $K$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| self-consensus | 72.7 | 78.9 | 81.4 | 82.7 | 83.5 | 84.2 | 84.6 | 85.0 |
| sibling text stream | 72.7 | 80.7 | 84.3 | 86.4 | 87.8 | 88.8 | 89.5 | 90.1 |
| margin | — | +1.8 | +2.9 | +3.7 | +4.2 | +4.6 | +4.9 | +5.0 |

Self-consensus gains voters as $K$ grows and the text stream does not, so the objection to Table 3 is that its cross-stream term is voter starvation at $K{=}8$ and would close at larger $K$. It widens instead, from $+1.8$ to $+5.0$ points, with an interval clear of zero at every $K$. The comparison is generous to consensus rather than to us: the stored consensus score is a mean over all seven siblings, so that arm ranks with eight voters at every $K$ rather than the $K$ it is entitled to.

Rebuilding consensus inside each subset, so that its pool grows with $K$ as it should, needs the candidate transcripts, which survive for 30 of the images. Consensus does climb there, $69\%$ at $K{=}2$ to $81\%$ at $K{=}8$, and the text arm climbs with it: the margin is flat at $+6.1$ to $+6.8$ points from $K{=}3$ on rather than closing. That row ranks against the dataset description, the weaker of the two rankers, and its $K{=}8$ column is a single subset per image rather than a mean over many, which is why its interval is the widest there. Nothing above extrapolates past 8, since the draws stop at 8 (Figure 15).

**Why not ask the image?** Both selectors compare one generation with another, and the obvious third is to ask the conditioner itself: the frozen VLM is already in the pipeline, so a candidate can be scored by its log-likelihood given the image, with no parameters and no second model family added. On the same 199 images and the same eight completed draws it reaches $94\%$ per token and $94\%$ unnormalized, against $96\%$ for agreement among the draws and $97\%$ for the sibling text, and adding it to the two the paper uses lowers them, $98\%$ to $97\%$. The conditioner's own opinion of a transcript is a weaker utility here than the stream it emitted beside it, which is the sharper form of §5.5's claim: what helps is not more grounding in the image but a second rendering of it to agree with.

**Is the cross-stream term a property of the scorer?** Ranking transcribes each draw with whisper-large-v3 and embeds it with MiniLM (Wang et al., 2020), and sem_sim reads the winner through the same two checkpoints. A selector that had learned those checkpoints rather than the content would produce the table above, so the arms were rescored on $1{,}323$ images and the same eight draws per image with a different family, wav2vec2-CTC (Graves et al., 2006) in place of Whisper (Radford et al., 2023) and MPNet (Song et al., 2020) in place of MiniLM. Self-consensus, which the shared scorer credits with $+0.083$, gains $+0.021$ there, $[+0.014, +0.028]$. Ranking against the sibling text gains $+0.141$, $[+0.134, +0.148]$, on $1{,}069$ of the $1{,}323$ images, on a scale where one draw reads $28\%$ of what it reads under the shared pair. Against that instrument's oracle over the same eight draws, $+0.156$, the sibling stream recovers $90\%$ of the gap and agreement $13\%$. The cross-stream term is $+0.120$ $[+0.114, +0.126]$ under the independent scorer against $+0.043$ $[+0.037, +0.049]$ under the shared one, the latter an independent rerun of the gain in §5.5 and both consistent with the 53-image check of Appendix C. What only the sibling stream can supply survives the change of instrument nearly whole, and what consensus holds survives at about a seventh of the sibling's gain: at $n{=}199$ its interval crossed zero and at $1{,}323$ it does not, so the separation is one of size rather than of presence.

**The same eight draws, given to the cascade.** Selection is available to any system that can sample, so the cascade was given it: eight captions per image at the speech head's temperature and top-$k$, each synthesized and transcribed through the same stage, on the same 199 images. One sampled caption reads $94\%$ against the greedy $100\%$ the tables report. Ranking the eight by agreement among themselves, the selector a single-stream system has, returns $100\%$: the cascade ends where it started. Its draws are not degenerate, since an oracle over the same eight reaches $111\%$; agreement simply does not find the draw the oracle picks, which is the pattern §5.5 reports on our own draws and the reason a second stream is worth having. The image-grounded selector this paper runs as an instrument does no better. Scoring the eight captions against the image with CLIP ViT-L/14 (Radford et al., 2021) and keeping the highest reads $96\%$ $[94, 99]$, a paired $-3.5$ points $[-5.5, -1.6]$ against greedy, and reading the transcripts rather than the captions reads $97\%$; both land below agreement rather than above it. Selecting a caption for how well it matches the image is not selecting it for how well it matches a spoken summary of a longer description, so the CLIP instrument that grounds our content claims is the wrong utility for this particular choice. Giving a system with one content stream a second look at that stream buys it nothing, by agreement or by image grounding, so the $90\%$ in §5.5 is measured against a cascade that has already been given both selectors available to it.

**Why the trained ASR-free scorer does not select.** A small transformer over Mimi (Défossez et al., 2024) cb0 tokens aligned to the sentence-embedding space retrieves the correct caption out of 256 at 94% on ground-truth codec tokens against a 0.4% chance rate. Yet as a selector it captures

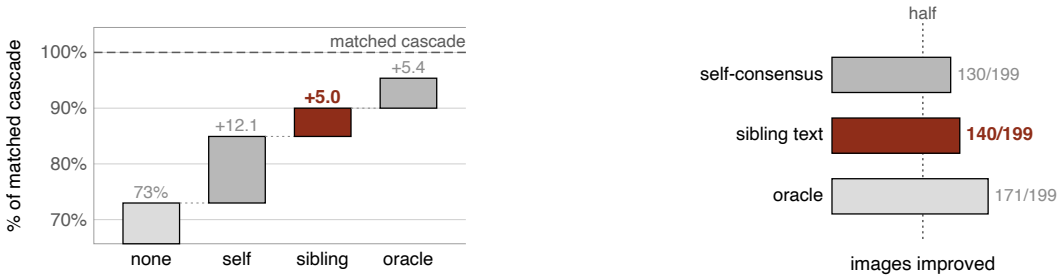



Figure 15: **Where the oracle's 22 points come from.** Left: all four arms on identical draws inside one run, so the differences are paired; self-consensus, which reads no text at all, covers 54% of the distance to the oracle and the sibling stream 76% (Appendix C.1 rescores both with an independent ASR and encoder). Right: per-image win counts over the same $n$=199.

19% of the oracle gap and ranks at chance; retraining it on the speech head's own generations gives $+0.025$ against $+0.031$.

Two effects separate cleanly. Scored on generated rather than ground-truth tokens over identical rows, the same scorer falls from 0.94 to 0.39, a large shift that is not the explanation, since 0.39 remains 101 times chance while the ranking stays at 50%. The binding constraint is the objective: selection needs the scorer to represent what a candidate *says*, so that its inner product with the text stream reproduces the transcription-based score, and it is trained instead to predict the embedding of the target caption, which all $K$ candidates share. An optimum of that objective assigns every candidate the same vector and therefore the same rank. Chance selection is what the objective asks for.

That identifies the scorer this mechanism needs, and we built it: regress the embedding of a candidate's own transcript, labeled by one offline ASR pass per training sample. Over 20,000 generations it reaches a Spearman rank correlation (Spearman, 1904) of 0.562 against the target, well clear of the chance ranking the caption-target objective gives, and selection does not move: $+0.040$ against $+0.040$ at $n$=199. So the objective bounded the ranking, not the selection.

### C.2 Why the speech stream works

**Why does the speech stream work at all?** Mimi cb0 quantizes WavLM's semantic space, so codebook neighbors are semantically close and a near miss draws a smaller correction than a distant one, for the reason the gradient below gives. Tokens are wrong at the index level while staying correct at the meaning level (the substitution test below).

**Borrowing the alignment lens from JEPA.** Joint-embedding methods (Assran et al., 2023; Bardes et al., 2024) study how a predictor $p_\psi$ aligns a source representation $f_\theta(x)$ with a target representation $g_\phi(y)$. Their continuous-target form, with sg the stop-gradient,

$$\mathcal{L}_{\text{JEPA}} = \left\| p_\psi(f_\theta(x)) - \text{sg}(g_\phi(y)) \right\|^2$$

needs a guard against collapse, which both take from an exponential moving average of the source encoder as target, since $f_\theta$ could otherwise shrink to a constant. We are generating, not matching embeddings; the fixed external vocabulary $\{1, \ldots, V\}$ makes our setup non-collapsible by construction, so no such guard is needed. What we *do* import is the analytic move of attributing observed alignment to the geometry of the target encoder $g_\phi$ rather than the predictor or the source encoder.

**Why a semantic codebook is load-bearing.**

**The gradient.** The cross-entropy gradient with respect to the logits $z_t$ at step $t$ is

$$\frac{\partial \mathcal{L}_\psi}{\partial z_{t,j}} = \text{softmax}(z_t)_j - \mathbf{1}[j = \tau_t].$$

For the linear projection $W \in \mathbb{R}^{V \times d_{\text{model}}}$ producing $z_t$, rows $W_j$ act as learned proxies for codebook entries $E_j$. When $E$ is structured by semantics (Mimi cb0, distilled from WavLM; Défossez et al.,

2024), proxies that converge to that geometry give graded supervision. The loss value does not read that geometry, since it is $-\log p(\tau_t)$ however the remaining mass is spread; the update does, because it reaches the hidden state as $W^\top(\mathrm{softmax}(z_t) - e_{\tau_t})$, with $e_{\tau_t}$ the one-hot vector of $\tau_t$, so a prediction close in $E$-space leaves a smaller residual to correct than a distant one. Whether the projection actually carries that geometry is a claim about a trained matrix, so we measure it rather than assume it. Over the $2048$ codes Mimi uses, the pairwise cosine structure of the trained cb0 projection and of Mimi's own codebook agree at Spearman $0.14$ and $0.24$ on training seeds 43 and 44; shuffling the projection's code indices, which keeps both geometries and breaks only the pairing, gives $|\rho| \leq 0.013$ over 200 shuffles on each seed. The agreement is partial rather than a copy, and it is the pairing that carries it. When $E$ is structured by acoustics (EnCodec cb0), the same gradient demands phoneme-level token fidelity that 8K samples cannot supply. This makes the enabling-mechanism claim concrete: the gap between Mimi ($58 \pm 5\%$ of the matched cascade, 3-seed mean; best $63\%$ seed-44) and EnCodec ($8\%$) under matched architecture, data, hyperparameters, and loss (both at $n$=200) is a property of the codebook geometry $E$, not of $f_\theta$, $p_\psi$, or the objective; SpeechTokenizer's first codebook, distilled from HuBERT rather than WavLM, reaches $60 \pm 5\%$ over two seeds on the clean cache where EnCodec's reaches $16\%$.

**A test that needs no training.** The claim that Mimi's cb0 quantizes a semantic space predicts something testable with no training: an index error between two nearby codes should cost little meaning, while an index error to an arbitrary code should cost all of it. We test it directly on ground-truth audio. Mimi's cb0 codebook is recovered from the EMA state (van den Oord et al., 2017) as embedding_sum/ $\max(\text{cluster_usage}, 10^{-6})$. Codes with zero usage are excluded, since their centroids are undefined and win every nearest-neighbor query. A fraction of the ground-truth cb0 tokens is then replaced either by each token's nearest surviving neighbor or by a uniformly drawn live code. The result is decoded at one codebook, transcribed and scored against the target. Mean cosine between a code and its nearest neighbor is $0.807$.

| substituted | 0% | 10% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| nearest neighbor | 100.0 | 96.6 | 85.3 | 66.8 | 56.7 | **46.4** |
| random live code | 100.0 | 67.5 | 33.1 | 15.7 | 8.5 | **10.1** |

At full substitution not one original index survives, yet the nearest-neighbor arm retains 46% of the unsubstituted score while the random arm sits at the floor, a $4.6\times$ separation produced entirely by which wrong code is chosen. This is why cross-entropy on cb0 trains a usable speech stream at $4.9\%$ token accuracy: most of the mass the model misplaces lands near the right code rather than anywhere.

### C.3 The measurement trap and perceptual quality

Table 11: Token-level fidelity (token accuracy of the head's sampled generations against the target tokens, position by position) and embedding-level content on the Whisper transcripts of the same heads. Test split, $n$=200, whisper-turbo + all-MiniLM-L6-v2, three Mimi seeds and one EnCodec run; word-level rates sit beside their reference control in Table 12.

| Codec | token_acc ↑ | % of cascade ↑ |
|---|---|---|
| Mimi (3-seed mean) | 4.9% | $\mathbf{58 \pm 5}$ |
| EnCodec | 7.4% | 8 |

**Token vs semantic decoupling, in full.** The main text summarizes this split; the per-axis numbers are here. Table 11 carries the token-accuracy and content columns and is not repeated.

Figure 4 shows the three best-scoring transcripts and Appendix E sixty held-out items in full. The transcripts are topically on target but not fluent: word choices are frequently wrong even where the gist survives. Word Error Rate of the first 80 characters of the Whisper (Radford et al., 2023) transcript against those of the target SST text is $0.99 \pm 0.02$ on the 10K test split (seeds 43 and 44, $n$=200 each, scored with jiwer; Vaessen, 2018), where Table 12 reads $0.96$ and $0.91$ for the 100K arms, and Character Error Rate is $0.77 \pm 0.01$. Token accuracy (4.9%), WER ($\approx 99\%$), and CER ($77\%$) all indicate near-zero string-level fidelity, while the same transcripts hold $60\%$ of the

matched cascade. Three independent string-level metrics agree the model gets the words wrong; one embedding-level metric says the meaning is preserved. Figure 16 scores the same audio and transcripts four ways.

**The trajectory evidence in detail.** Validation loss bottoms at epoch 3 (val_loss $= 2.54$) while Whisper-based semantic similarity keeps rising through epoch 30 (Figure 3, right). Minimum-val_loss selection picks epoch 3 and AR-token-accuracy selection picks its peak at epoch 28, whose in-loop sem_sim is $1.7\times$ epoch 3's. For semantic codecs, val-loss is the wrong selection criterion.

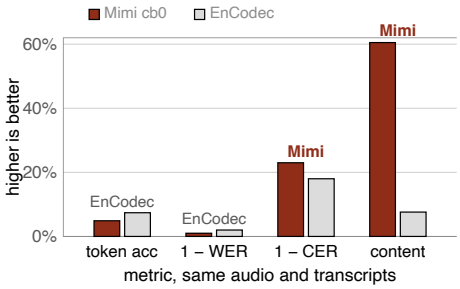



Figure 16: **Two of the three string metrics rank the codecs backwards.** The same audio and transcripts scored four ways, each oriented so higher is better.

**The decoupling extends to perceptual audio quality.** DNSMOS (Reddy et al., 2021; 2022), Microsoft's P.808 and P.835 deep predictors scored on their OVRL / SIG / BAK / P.808 axes, gives a mean overall MOS of $1.08$ on $n$=200 test items for our 10K cb0-only synthesis (seed-44) vs $3.42$ for the Bark ground-truth audio that produced the Mimi (Défossez et al., 2024) tokens. The cb0-only output ($1.08$) lands at the random-noise DNSMOS floor on the perceptual axis (measured: Gaussian noise $1.108$, uniform $[-0.5, 0.5]$ noise $1.067$, $n$=30 each), while the same audio reaches $63\%$ of the matched cascade ($n$=200, seed-44) on the Whisper-transcribed axis.

**Where the floor comes from.** That floor belongs to the decode path and not to the head: passing ground-truth audio of the 100K test split through Mimi's own encoder and straight back out at one codebook scores $1.077$ ($n$=30), so tokens as correct as tokens can be land there too, and a source's score ($3.461$ for that audio) is not reachable through a one-codebook decode by anything. At eight codebooks the same round-trip returns $3.430$ against $3.461$ for the source waveform, so Mimi does carry the missing $2.35$, and it carries all of it in the seven acoustic codebooks we hold at weight $0.02$ rather than in the semantic stream we train. On naturalness the cascade leads.

**The cb progression.** In a controlled cb-progression eval on the 100K 430M 8-codebook checkpoint (decoding the same predicted token sequence with the Mimi decoder set to 1, 2, or 8 codebooks; $n$=30), overall DNSMOS is unchanged ($1.076$, $1.076$, $1.077$; Figure 17). The ceiling for that progression is set by the codec, not by us (encoding the ground-truth audio with Mimi and decoding it straight back gives OVRL $1.077 \pm 0.008$ at one codebook and $3.430 \pm 0.065$ at eight, against $3.461 \pm 0.051$ for the source waveform; $n$=30, 100K test split). A one-codebook decode is therefore on the noise floor with perfect tokens, which is where our $1.076$ sits, and the entire $2.35$ separating the two round-trips lives in the seven acoustic codebooks that carry weight $0.02$ in our objective. The eval was repeated on the two 100K checkpoints, 173M and 430M heads at identical codebook settings, both decoded at cb=8 with $n$=30. It gives the same near-floor numbers, $1.073$ and $1.075$ on a second draw of 30 items. $7.5\times$ training data raises content from $58\%$ of the matched cascade at 10K to $69\%$ at 100K, each under its own protocol, but does not move perceptual MOS.

**Why it follows from the objective.** This follows from the training objective, not from Mimi or data scale (cb1–7 train at weight $0.02$ against cb0's $100$, so the acoustic-codebook predictions get almost no gradient and decoding more codebooks adds nothing). Raising the acoustic weight $500\times$ under otherwise identical training leaves DNSMOS on the floor; that run was stopped at epoch 17, when its in-loop validation sem_sim stood at $95\%$ of the $0.02$ run's best. UTMOS (Saeki et al., 2022) agrees (our audio and the random-token floor both sit near $1.5$, against about $4.5$ for the cascade and for the reference rendered by the same synthesizer, so naturalness is set by how many codebooks are decoded, not by the speech head).

Figure 18 gives latency by stage and the acoustic-supervision control.

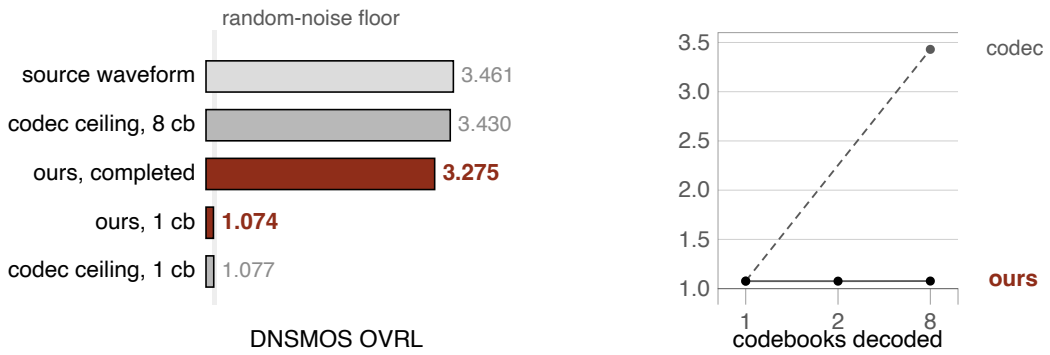



Figure 17: **The perceptual floor is a decode artifact.** DNSMOS OVRL, $n$=30, 100K test split. Left: Mimi's round-trip of ground-truth audio at one codebook scores 1.077, where the head's cb0 sits; completion lifts the head's audio to 3.27 ($n$=199, Appendix C.4). Right: decoding 1, 2 or 8 of the head's codebooks moves nothing, while the round-trip gains 2.35.

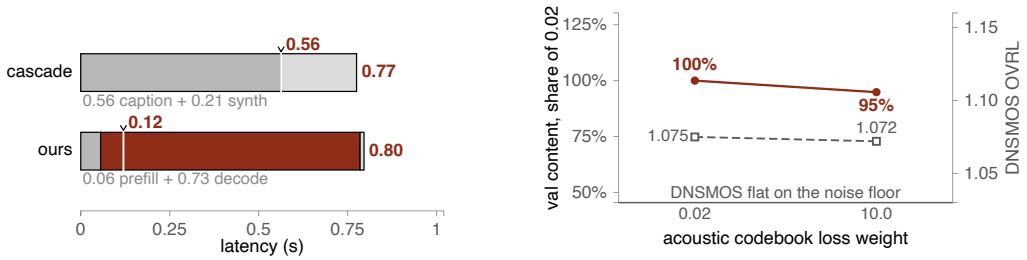



Figure 18: **Latency by stage, and the acoustic-supervision control.** Left: latency on one idle GPU, ours as prefill plus decode, the cascade as caption plus synthesis; the caret marks first audio. Right: peak in-loop validation sem_sim against DNSMOS with the acoustic-codebook loss weight raised from 0.02 to 10; the weight-10 run stopped at epoch 17.

### C.4 Acoustic completion

The head's audio is its cb0 stream decoded alone, and §C.3 shows that this path sits on the perceptual floor with perfect tokens (the seven acoustic codebooks carry naturalness, and weighting them up in the speech head's training leaves it on the floor). The completion stage models $p(\text{cb}_{1:7} \mid \text{cb}_0)$ instead, as AudioLM's coarse-to-fine stage and Moshi's depth transformer do (Borsos et al., 2023; Défossez et al., 2024). The stage is a transformer encoder (Vaswani et al., 2017) over the cb0 sequence (8 layers, $d_{\text{model}}$=512, 41M parameters) and, per frame, a depth-wise chain that predicts codebook $k$ from the encoder state and the embeddings of codebooks $1..k{-}1$, teacher forced. It trains with cross-entropy over codebooks 1-7 on the 100K corpus's own Mimi tokens and sees no hidden state and no text (the speech head is frozen). At inference the speech head's cb0 tokens are completed greedily and all eight codebooks decode through Mimi.

**What completion buys.** Table 12 renders the same 199 items six ways. On the speech head's draws completion lifts content from 72% of the matched cascade to 86% (paired +14 points, bootstrap 95% CI $[+11, +17]$). It lifts DNSMOS (Reddy et al., 2022) from 1.07 to 3.27 (Table 12). The second training seed replicates the gain on the same items, 72% to 85%. The reference's own cb0, completed the same way, reaches 0.960 against 0.989 for its eight native codebooks. The gain on the speech head's draws is the decode path's, not the speech head's (the content was already in cb0, and Whisper reads it better from a natural voice). Completion raises the share of the target's content words the transcript recovers by $+0.044$ $[+0.033, +0.056]$, so the content instrument moves with the words recovered and not with fluency alone. The null column rules out a general decodability effect: against another image's summary every arm sits between 0.130 and 0.139.

Table 12: Acoustic completion on the fixed 100K test split ($n$=199, whisper-large-v3, seed-43 draws from the frozen 430M head). Each arm decodes through the same Mimi decoder; cb0 arms decode one codebook, the others eight. $\pm$ is one bootstrap standard error over items, so it and Table 1's seed spreads measure different things, and the cb0 row sits one point below its two-seed mean there. The null column stays on the raw scale.

| Arm | % of cascade | null | WER | DNSMOS | UTMOS |
|---|---|---|---|---|---|
| head, cb0 alone (Table 1) | 72±2 | 0.130 | 0.96 | 1.07 | 1.43 |
| head, cb0 + completion | **86±2** | 0.138 | 0.91 | 3.27 | 3.31 |
| head, cb0 + causal completion, 90M | 85±2 | 0.139 | 0.95 | 3.26 | 3.08 |
| reference, cb0 alone | 117±2 | 0.130 | 0.28 | 1.07 | |
| reference, cb0 + completion | 143±1 | 0.131 | 0.05 | 3.30 | 3.31 |
| reference, its own 8 codebooks | 148 | 0.130 | 0.01 | 3.40 | 4.39 |

**Capacity.** A 90M model of the same design (12 layers, $d_{\text{model}}$=768) scores the same on the speech head's draws, 86% of the cascade, so completion quality on the generated stream is saturated at 41M. The 90M model, made causal so that frame $t$ reads cb0 up to $t$ only, scores 85% of the cascade on the same draws at DNSMOS 3.26. One pass over a finished stream of 50 to 150 frames takes 15 to 17 ms, and extending by one frame without a cache takes the same 16 ms, inside Mimi's 80 ms frame period, so completion keeps ahead of playback and the completed stream keeps the speech head's first-audio latency.

**Composing with coupling.** Completion composes with coupling. On a second draw of $K$=8 the ranking of §5.5 reaches 91% against 72% for one draw. Completing the ranked stream reaches 94% of the cascade, at DNSMOS 3.28, paired +3 points 95% CI $[+2, +5]$. The gain is smaller on ranked draws because ranking already favors draws whose cb0 transcribes well, so the two mechanisms overlap there and stack the rest.

**Rank what a listener hears, not the stream it came from.** The coupled arm ranks eight draws by their cb0 transcripts, the least legible rendering the monoplex has (at one codebook even the reference transcribes less faithfully than at eight). Completing all eight and ranking those transcripts by the same sibling text differs only in when the completion stage runs, and reaches 0.647 against 0.630 for ranking before completion, on the same 199 images and draws. Agreement among the completed draws, ranked alone, reaches 0.644, and the two signals are complementary: standardized and summed they reach 0.654, above either.

The second training seed replicates every step within 0.005: 0.611 ranked and 0.632 completing the ranked pick, then 0.650 ranking the completed draws, 0.642 by their agreement and 0.654 for the standardized sum, against 0.690 for the oracle ($n$=199). All are above ranking with no completion and above one completed draw, which read 0.611 and 0.570 on this seed and 0.608 and 0.578 on the first. Over all 1,993 images the standardized sum reads 0.654 $[0.648, 0.659]$ against 0.578 $[0.572, 0.585]$ for a completed draw and 0.611 for the cb0 pick, a paired gain of +0.075 $[+0.071, +0.080]$.

The pick is no longer than the draws it is chosen from, 22.9 words against 22.8 for the first draw and for the oracle's pick, so ranking against the longer sibling text does not pull the speech toward its length. Choosing each image's best completed draw against its target, a ceiling rather than a method, reaches 0.686, so the eight draws are not what limits this arm. Completing all eight costs $K$ completions of a 41M-parameter stage rather than one, beside the $K$ transcriptions that already dominate the arm's cost.

# D What the Pair Carries Over One Stream

Two decoders that read one latent with no path between them factorize, so whether their outputs cover each other's gaps has to be measured rather than assumed. The measurement needs a target neither stream wrote: the spoken target of §4 summarizes the written one, and the text stream would recover it by construction. COCO (Lin et al., 2014) already carries five human captions for each image, and they are that target, on the same scale the selection control of Table 3 is scored on. Table 13 adds one thing at a time to monoplex's own text stream and reports the paired difference.

**What the pair carries.** The pair carries more than the text stream alone on both measures. Every bootstrap interval (Efron, 1979) clears zero, including the one that holds the word budget fixed: at the speech stream's budget the pair reads $+0.016$ $[+0.012, +0.021]$ against a text stream given twice those words, and $+0.037$ $[+0.022, +0.053]$ on hit rate. The second stream is neither free of the first nor redundant with it. Selecting a draw by its agreement with the sibling text could have bought content by picking whichever draw renders that text most closely, at the cost of exactly the complementarity a second stream is for. It does not: ranking raises what the pair carries over the written stream alone by $+0.005$ $[+0.004, +0.007]$ on mean best and by $+0.017$ $[+0.010, +0.024]$ on hit rate, both clear of zero, so the ranked pick covers more of a target neither stream wrote than the unranked draw does.

**A second stream, not a second modality.** From a base clipped to the speech stream's word budget, adding the speech stream is worth $+0.062$ and a sampled one-sentence caption from the same VLM $+0.112$; on hit rate the two add $+0.126$ and $+0.270$, the caption about twice the speech stream. The other rows of Table 13 add to the unclipped stream. A reference summary, the generous stand-in for what a serialized system speaks, also adds more than our speech stream does. The instrument is not the limit: assigning the spoken stream exactly the captions the written one covers worst moves hit rate $+0.320$, so a speech stream that covered the written one's gaps would be visible here.

**What is priced.** The price is the pair's as built: the speech head reads a mean pool of $h$, where the text head attends over all of it. Inference-time coupling recovers part of that price in §5.5. Recovering the rest needs a path between the streams that does not serialize one through the other, the direction the Discussion names and this paper does not attempt.

Table 13: What a second stream adds to monoplex's own text stream against five human COCO captions per image: paired differences, bootstrap 95% intervals, $n$=912 of the fixed test split, seed 43. Mean best: each caption's best-matching sentence, averaged; hit rate: captions matched above 0.6. The last block: the 199 items both systems ran on.

| Added to monoplex's text stream | Δ mean best | Δ hit rate |
|---|---|---|
| **its own speech stream** | $\mathbf{+0.014}$ $[+0.012, +0.015]$ | $\mathbf{+0.050}$ $[+0.042, +0.058]$ |
| the same stream before ranking, one draw | $+0.008$ $[+0.007, +0.010]$ | $+0.033$ $[+0.027, +0.039]$ |
| ***What else a second stream of the same size is worth*** | | |
| a reference summary, what a serialized system speaks | $+0.020$ $[+0.018, +0.022]$ | $+0.075$ $[+0.065, +0.085]$ |
| its own speech stream, over the cut stream | $+0.062$ $[+0.057, +0.068]$ | $+0.126$ $[+0.113, +0.141]$ |
| a sampled one-sentence caption, over the cut stream | $+0.112$ $[+0.104, +0.119]$ | $+0.270$ $[+0.252, +0.288]$ |
| an oracle assignment of the captions text covers worst | | $+0.320$ $[+0.303, +0.338]$ |
| ***What a shipped parallel model's second stream adds, on the items both were run on*** | | |
| **monoplex**, $n$=199 | $\mathbf{+0.014}$ $[+0.011, +0.017]$ | $\mathbf{+0.044}$ $[+0.028, +0.061]$ |
| Qwen2.5-Omni-3B, its speech over its own text | $+0.004$ $[+0.003, +0.007]$ | $+0.017$ $[+0.008, +0.028]$ |
| the paired difference | $+0.009$ $[+0.006, +0.013]$ | $+0.027$ $[+0.007, +0.047]$ |

# E CASE STUDY ACROSS HELD-OUT IMAGES

Figures 19 to 22 show the first sixty items of the fixed MS-COCO (Lin et al., 2014) test split, grouped by how the two streams divide the target between them (Figure 4 showed the three that score highest).

**What the panels show.** Appendix D measures that division over the split against a target neither stream wrote, and the words each stream contributes are different words. We render the transcript of the matched cascade in the text-first column rather than our own text stream.

target words recovered by **speech only** ***text-first only*** **both** neither

**The streams split the target between them** 38 of 58



speech **57%** text-first 88%
A ***person*** rides a ***bicycle*** at **night** on a **city street** while holding an ***umbrella*** against the rain.



speech **99%** text-first 120%
A **typical office** ***workstation*** set up includes a ***wooden desk*** facing a white wall with a **computer** setup consisting of a ***monitor, keyboard, mouse,.***



speech **107%** text-first 106%
A **busy urban street** scene in an European ***city*** featuring a ***red double-decker*** **bus** with the route number 29 heading toward the viewer, a ***woman*** **walking.**



speech **92%** text-first 115%
A tranquil **beach** ***scene*** with **clear blue** skies, ***turquoise*** water, white sails on ***anchored*** **boats,** sandy shores, and a gentle slope leading to a **beach.**



speech **40%** text-first 73%
A ***man*** walks away from a **train station platform** adorned with wooden benches and ***luggage,*** hinting at a busy transit **center** where travelers prepare for.



speech **59%** text-first 114%
In a **vibrant** market scene **featuring** ***avocados*** and **oranges,** a ***hanging scale*** provides an important element by measuring the weight of the produce,.



speech **69%** text-first 103%
A peaceful rural scene featuring ***horses grazing*** in a **lush green field,** with electric transmission towers in the **background,** creating a harmonious.



speech **83%** text-first 118%
A **man** is **enjoying** a **large** ***slice*** of ***pizza*** in a casual ***restaurant*** setting, complete with its own metal ***pizza*** tray and various toppings.



speech **78%** text-first 114%
A vibrant **urban street scene** in an unnamed ***city*** featuring a ***yellow taxi,*** multiple skateboarders, and variously dressed pedestrians.



speech **40%** text-first 116%
A ***snowboarder*** **performs** an impressive jump over a snowy terrain, **surrounded** by a diverse ***crowd*** of **spectators** who are eagerly waiting for the action.



speech **78%** text-first 97%
A serene **canal scene featuring** two narrowboats, one with "BOJANGLES ARLICK" signs and the other without, both ***docked*** near a ***wooden*** dock with multiple.



speech **86%** text-first 75%
A ***clear,*** sunny day in a **park features** a ***group*** of ***people*** enjoying activities like kite-flying and strolling, surrounded by lush vegetation and a.

Figure 19: **The two streams do not miss the same words.** The first sixty items of the fixed test split, grouped by whether the pair recovers more of the target than the better single stream does. Each content word of the target is marked by which stream recovered it, by rule and not by hand; *text-first* is the matched cascade. Targets with fewer than three content words are left out. Continued in Figures 20 to 22.

target words recovered by **speech only** ***text-first only*** **both** neither

**The streams split the target between them (continued)** 38 of 58



speech **90%** text-first 134%
A **young** boy **stands** next to a textured, ***graffiti-covered*** **wall, holding** a **tennis racket** and a **tennis** ***ball.***



speech **65%** text-first 138%
A **man stands** on a ***rooftop*** overlooking a ***cityscape,*** wearing a ***white shirt*** and dark ***tie,*** holding a ***cigarette.***



speech **114%** text-first 119%
A messy **office** setup with a black **desk** adorned by a blue exercise ball, **cluttered** with a ***computer monitor, keyboard,*** mouse, lamp, ***various*** **books,** and.



speech **43%** text-first 111%
A sleek **black** ***cat*** walks across a **lush,** manicured ***grassy field*** near a blooming ***tree*** in the ***background,*** creating a tranquil and picturesque scene.



speech **98%** text-first 113%
A **historic** ***steam*** **clock tower stands** prominently in an urban setting, **featuring** a large, ornate **clock face** and black chimneys above it.



speech **97%** text-first 87%
A **busy urban street scene** on a ***hazy*** day, featuring heavy **traffic,** obstructed views through foggy trees, and illuminated by streetlights.



speech **12%** text-first 105%
A ***skier*** in a black jacket and dark green camouflage pants performs a turn on a ***snowy*** mountain ***slope*** during winter, **surrounded** by ***other skiers*** and.



speech **103%** text-first 118%
A classic **parking meter** with a **digital** ***display*** showing remaining **time** in minutes and seconds includes ***buttons*** for adjusting **time,** a coin slot, and a.



speech **46%** text-first 113%
A ***large,*** reddish-brown cow with long, curly hair and distinctive ***horns*** stands in a lush, **grassy field surrounded** by trees and shrubs, with a ***barbed.***



speech **87%** text-first 124%
A well-laid out **dining table** with a light ***blue tablecloth,*** featuring a mix of salads and **sandwiches,** along with proper ***utensils,*** sets the scene for.



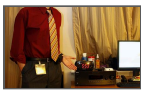
speech **49%** text-first 113%
A **young** ***boy*** performs a ***skateboarding*** trick in a public ***park*** with a beach theme, surrounded by tents and umbrellas for an **outdoor** gathering, with.



speech **95%** text-first 81%
A **rainy** urban scene with two men ***walking down*** the **street,** one in a black coat and the other in a black coat with a green ***tie,*** both dressed hurriedly.



speech **99%** text-first 110%
A **small,** compact **bathroom** features a white vanity with a **sink,** a white medicine cabinet above it, a toilet, and a ***towel rack*** on the wall.



speech **91%** text-first 93%
A **man stands** at an office **desk** with ***various*** office supplies, including a **computer,** keys, and a coffee maker, while his watch shows 559.



speech **111%** text-first 106%
A **giraffe** in a **grassland** setting, with its ***head*** tilted upwards and extending its long neck towards a cluster of **green,** slender branches.

Figure 20: Case study, continued: items 13 to 27 of the grouped list of Figure 19, same marking.

target words recovered by **speech only** ***text-first only*** **both** neither

**The streams split the target between them (continued)** 38 of 58



speech **84%** text-first 120%
***Three*** **sheep,** each with unique markings, stand in a circle within a wooden ***pen*** **surrounded** by a barn-like structure, with straw floors and equipment in.



speech **60%** text-first 97%
A stylish **man stands** outside with an open **umbrella,** dressed in a formal black ***suit*** with a striped pattern, under lush greenery.



speech **4%** text-first 77%
A ***stable*** interior with well-lit metal walls and ceiling a trio of girls grooming and ***riding*** a well-groomed **black** ***horse,*** one ***riding*** while others.



speech **26%** text-first 113%
A tranquil grassland scene **featuring two** ***zebras grazing*** side by side, highlighting their unique black-and-white patterns amidst lush vegetation,.



speech **50%** text-first 100%
A busy urban **street** with a ***double-decker*** **bus** labeled "Evesham 28A" stopping at a **bus** ***stop.***



speech **104%** text-first 104%
A **historic** stone **clock tower** in a quaint small town square, featuring a weathered appearance, stone steps, an arched entrance, and a ***blue*** storefront.



speech **32%** text-first 64%
A **man** holds ***scissors*** while talking on a phone inside an open-plan living space with a **table,** chairs, and a door leading to another ***room.***



speech **74%** text-first 58%
An urban ***street*** **intersection featuring** a black **traffic** light pole with green light, two angled ***street*** signs, a multi-story beige building with dark.



speech **42%** text-first 59%
A Kenwood ***blender*** sits on a wooden surface near a cozy living room **setting** with a **white** sofa and colorful decorations, suggesting an informal and.



speech **65%** text-first 114%
**Two** ***zebras*** **stand** on a sandy ground covered with sparse vegetation, facing slightly to the left of the camera.

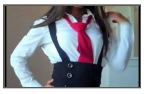


speech **42%** text-first 105%
A skilled ***skateboarder*** performs a high-flying stunt on a **rail** in an outdoor skate park setting, **showcasing** their abilities against a backdrop of.

**One stream carries it and the other adds nothing** 20 of 58



speech **97%** text-first 127%
**Two cats** gaze ***out*** from an old, weathered **window** ***frame*** with ***peeling paint*** and missing pieces, inside a ***wooden*** wall showing signs of age and potential.



speech **43%** text-first 69%
A man in an office setting dressed in a formal attire consisting of a ***white shirt*** with a **red tie, black** pants, and ***suspenders,*** standing near a ***white.***



speech **122%** text-first 120%
A curious **black cat sits inside** an open **white suitcase** on a beige bed surrounded by various objects and fabrics, creating a cozy and inviting.



speech **42%** text-first 111%
A brown ***teddy bear*** with white accents holds a rectangular **wooden** block in its right paw while ***sitting*** on a **wooden surface** next to a **wooden** wall.

Figure 21: Case study, continued: items 28 to 42.

target words recovered by **speech only** ***text-first only*** **both** neither

**One stream carries it and the other adds nothing (continued)** 20 of 58



speech **99%** text-first 117%
A **young girl** holds a **slice** of golden-brown cheese **pizza** in her mouth at a busy **restaurant,** expressing enjoyment and satisfaction.



speech **112%** text-first 78%
A bustling **urban scene** at a **train station** features a sleek, white **train parked** alongside a pedestrian walkway filled with commuters engaged in.



speech **86%** text-first 81%
A **train station** set amidst a mountainous landscape with a **yellow** and **blue** locomotive at the center, **surrounded** by snowy peaks and a busy road with.



speech **117%** text-first 110%
A **cozy living room** with a **wooden** entertainment center, framed pictures, a **wooden** side **table** with a lamp, an armchair near a TV, a side **table** with.



speech **100%** text-first 120%
A **close-up** of a ***cat's*** face showing its ***eye*** and fur reflecting light from an outdoor ***window,*** revealing the ***cat's*** curious gaze outdoors and a snowy or.



speech **98%** text-first 118%
A black-and-white photo shows a striped ***tabby*** **cat** relaxing on a ***couch*** covered in a soft plush ***blanket,*** with pillows positioned behind it for added.



speech **62%** text-first 104%
A dimly lit bathroom features a clean **white toilet surrounded** by graffiti-covered metal pieces, a partially full silver trash can, and a partially.



speech **67%** text-first 97%
A person walks along a narrow **dirt path** lined with ***lush*** green foliage in a tropical or subtropical area, likely enjoying a leisurely stroll in nature.



speech **18%** text-first 110%
A showroom displays a classic wooden ***bicycle*** mounted on a ***shelf*** above a sports bike on a stand.



speech **87%** text-first 97%
A dimly lit bedroom features a TV-mounted **bed** displaying a dynamic scene on its screen while the rest of the **room** remains mostly unadorned with a.



speech **28%** text-first 92%
A ***man*** plays **frisbee** against an opposing player on a wooden court within a gymnasium.



speech **23%** text-first 107%
A pair of zebras stand together in an arid, ***rocky*** terrain near two ***trees,*** with little vegetation nearby.



speech **78%** text-first 85%
A **baseball game** at night with three players - a batter, a catcher, and an umpire - in an elevated view, capturing the action during a swing.



speech **38%** text-first 83%
A **woman** sips white ***wine*** at a



speech **82%** text-first 77%
A **baseball** game in progress on a well-manicured field with a clear view of the outfield fence and a chain-link fence in the background.



speech **72%** text-first 61%
A **baseball** game scene with spectators watching through a netted fence.

Figure 22: Case study, continued: items 43 to 58.