WHEN FROZEN REPRESENTATIONS TRANSFER: TEMPORAL AND STATISTICAL INTERFACE EFFECTS

Anonymous authors

ABSTRACT

Can one frozen model's internal representation become useful to another through only a small learned interface? We study acoustic representation transplantation into frozen language-model prefixes and speech-decoder memory. On a fixed 1,000-example test, a linear Whisper-to-Qwen bridge reaches 5.51–5.53% word error rate across two training seeds, compared with a 4.79% native Whisper reference. Omni-to-Qwen reaches 8.86–9.40%, whereas Omni-to-Whisper remains at 63.40% despite a substantial aggregate advantage over shuffled and zero inputs. A matched development factorial separates local averaging from shorter recipient-facing presentation. Mean pooling improves word error rate by 13.45 points over stride selection at matched rate; shorter presentation of pooled content adds 4.45 points. Fit-derived norm or coordinate-moment calibration recovers about 72% of the pooling advantage, so that advantage cannot be attributed uniquely to temporal integration. A separately selected pooling recipe confirms on test at 8.68%, versus 59.48% for raw states, a paired 50.80-point improvement with 95% interval [49.03,52.57]. These results support selective, directed functional portability: coordinate alignment can expose useful source information, but its usefulness also depends on presentation, feature statistics, and the recipient's native interface.

1 INTRODUCTION

Learned representations are usually consumed by the components with which they were trained. Their apparent incompatibility with other models may reflect missing information, an unsuitable coordinate system, or a mismatch in the interface through which that information is presented. We examine these possibilities using a deliberately restricted intervention: freeze both pretrained endpoints and learn only a small connecting bridge.

This is not a new connector architecture. Model stitching already uses learned connections to test functional compatibility (Lenc & Vedaldi, 2015; Bansal et al., 2021); frozen multimodal interfaces have a substantial history (Tsimpoukelli et al., 2021; Li et al., 2023). In speech recognition, SLAM-ASR demonstrates that a frozen speech encoder and language model can be connected using only a trainable linear projector (Ma et al., 2024). Our question is therefore not whether such a projector can work, but *which aspects of its interface make it work, and how narrowly a successful transplant should be interpreted.*

Audio is useful for this distinction because a representation is a sequence, not simply a point in a vector space. Reducing its rate simultaneously changes the content of individual vectors, their statistics, and the number of positions the recipient must consume. Moreover, a language-model prefix and a specialist decoder's cross-attention memory are different interfaces even when both accept continuous tensors. Treating these differences as coordinate alignment alone obscures experimentally separable effects.

We provide three connected results. First, a directed source–recipient matrix shows strong Whisper/Omni-to-Qwen transfer alongside weak Omni-to-Whisper transfer, retaining correct, shuffled-donor, and zero controls. Second, a matched temporal factorial and fixed-statistics controls separate presentation length from the effects of averaging; much of averaging's advantage is recoverable by simple calibration. Third, fixed artifacts confirm on a 1,000-example test, without new training or selection. A successful bridge supports task-level portability, not equivalent representations (Smith et al., 2025), universal latent geometry, or an independently composable interlingua.

2 FUNCTIONAL TRANSPLANTATION AND ITS CONTROLS

Let E be a frozen source encoder, T a fixed sequence transform, and R a frozen recipient. We learn an interface B_θ and predict

$$p_\theta(y \mid x) = R\big(y \mid B_\theta(T(E(x)))\big). \tag{1}$$

For the Qwen paths, $B_\theta(z) = Wz + b$ projects each 1,280-wide source vector into the 5,120-wide language-model embedding space. The recipient receives the projected sequence through its continuous prefix interface. For the Whisper recipient, the bridge instead supplies encoder memory to decoder cross-attention. Native Whisper is an untrained reference, not another learned transplant.

We distinguish *source dependence* from *useful generalization*. Correct input uses the current utterance. Shuffled input uses a manifest-frozen different utterance. Zero input removes source information while preserving the recipient input shape. An affine bridge can produce a nonzero output from zero input; zeros are applied *before* the bridge. A correct-input advantage supports example-specific source use, but poor correct-input transcription can still make that interface practically ineffective.

The controls preserve the established recipient-specific implementations. For Qwen, donor trajectories are linearly interpolated to the target input length with `align_corners=False`. For Whisper adapters, donors retain their own lengths. Consequently, these are not universally length-matched interventions. Native Whisper receives no artificial donor or zero condition. Each reported cell is one directed interface; the shared-source comparison across Qwen and Whisper is not a test of opposite-direction communication.

3 EXPERIMENTAL DESIGN

3.1 DATA, ENDPOINTS, AND SELECTION

The public-data corpus has 6,000 training, 600 validation, and 1,000 test examples. The test combines 300 LibriSpeech, 300 VCTK, 150 VoxPopuli, 150 Speech Commands, and 100 FLEURS examples, with 948 English and 52 Danish examples. The latter datasets are described by Wang et al. (2021), Warden (2018), and Conneau et al. (2022), respectively. AMI contributes training data only. Source, transcript provenance, checksum, split and donor identity are recorded per example. Donors are chosen within split, source and language, preferring the same duration bin and falling back to nearest duration. No source exceeds 30% of training rows.

The train/validation audit finds no sample-ID, recording, checksum or known speaker overlap. Speaker IDs cover 520 of 600 validation examples; missing FLEURS identities prevent a global speaker-disjointness claim. Temporal development uses 1,200 fit and 300 development rows drawn only from training. Its subsequent 600-row validation comparison reuses the primary selection set and is labelled validation confirmation, not untouched testing. The final test was opened only after artifact, decoding and reporting choices froze. Historical mechanical preflights had parsed split manifests; we do not claim the test manifest had never been read. Unknown foundation-model pretraining exposure cannot be excluded.

The endpoints are Whisper-large-v3-turbo (Radford et al., 2023), Qwen3-Omni's audio encoder (Xu et al., 2025), and Qwen3.8-27B. We use Omni's 1,280-wide `pre_proj2_1280` representation, approximately 12.5 positions per second, and Whisper's approximately 50-Hz valid encoder trajectory. Qwen is frozen in NF4; its 6,558,720-parameter bridge is FP32. Endpoint revision identifiers and selected artifact hashes are preserved with the evaluation. This study does not isolate architecture from pretraining objective or interface semantics, nor does it compare these models under equal total parameter counts.

The full-scale Qwen-facing runs train for 5,000 updates with AdamW, learning rate 10^{-3}, weight decay 0.01, 3% warmup, cosine decay, gradient clipping at 1.0, and effective batch four. Bridge weights start at $\mathcal{N}(0, 0.02)$ with zero bias. Candidates occur every 500 updates and are selected by validation WER, then CER, then later update. Seeds 42 and 69 remain separate fixed artifacts on test. The per-example objective is

$$\mathcal{L}_i = \ell(y_i \mid x_i) + 0.25 \max\{0, 0.10 + \ell(y_i \mid x_i) - \ell(y_i \mid x_{d(i)})\}, \tag{2}$$

where $d(i)$ is the frozen donor. The hinge is applied before microbatch averaging. Source and recipient parameters are never updated.

Table 1: Fixed-artifact test WER (%). Lower is better; insertions can raise WER above 100%. The native reference has no learned bridge.

Interface	Seed/recipe	Correct/native	Shuffled	Zero
Omni → Qwen	42	8.86	109.52	144.06
Omni → Qwen	69	9.40	108.96	161.88
Whisper → Qwen	42, mean4	5.53	109.74	194.08
Whisper → Qwen	69, mean4	5.51	109.61	173.39
Omni → Whisper	matched v2	63.40	113.99	102.20
Whisper → Whisper	native	4.79	—	—

Temporal arms use 1,000 updates, candidates at 250/500/750/1,000, and the same seed42, fit rows, target order, donor mapping and four-example optimizer groups. They estimate matched within-seed effects, not between-seed robustness. The Whisper-recipient matched-v2 artifact is retained separately; it is not presented as an equal-compute comparison with the Qwen-facing training recipe.

3.2 Temporal interventions and scoring

The four-arm factorial compares raw 50-Hz states, stride-selected states near 12.5 Hz, local means near 12.5 Hz, and those means repeated to approximately 50 Hz. The names `mean12` and `stride12` denote approximate presentation rates, not pooling factors. Pooling uses nonoverlapping valid FP32 means, retains a final partial window, and casts back to the input dtype. The repeated condition approximately preserves pooled content while changing presentation length. A separate rate curve measures factors 1/2/3/4/5/8. Two fixed calibration controls transform stride-selected states using either a fit-derived global norm scalar or coordinatewise mean/standard-deviation matching to pooled states. Neither transform uses development or test labels.

WER and CER are corpus additive edit counts divided by reference counts; exact match uses normalized transcripts. All paths use greedy decoding with at most 96 new tokens. Learned paths use batch one; native Whisper uses language-grouped batches of eight and the recorded transcription language. The final allocation contains eight learned systems times three controls, plus native Whisper: 25,000 predictions over the same 1,000 examples.

The frozen report contains 25 paired contrasts overall and within each source and language. Intervals use 10,000 paired-utterance bootstrap draws, seed42, manifest order, no stratification, and noninterpolated sorted quantiles. These intervals condition on fixed artifacts; they are not training-seed uncertainty, equivalence tests or familywise-adjusted inference. All allocated results are reported, including unfavorable contrasts. Independently implemented scorers agree on all 3,868 compared report values; this checks arithmetic and reporting, not independent correctness of their shared specification.

4 Results

4.1 Selective directed portability confirms on test

Both Qwen-facing source families support useful transcription (Table 1). Whisper-source bridges have a small absolute gap to native Whisper, but not demonstrated equivalence: transplant-minus-native is $+0.74$ WER points [0.18,1.23] for seed42 and $+0.72$ [0.13,1.26] for seed69. Whisper-source Qwen bridges outperform the corresponding Omni-source artifacts by 3.33 points [2.40,4.48] and 3.89 [2.92,5.06]. Neither test result selects between seeds. The independent Whisper source also shows that successful transplantation does not require shared Qwen-family ancestry.

Omni-to-Whisper remains weak at 63.40%, versus 4.79% native. Correct input nevertheless beats shuffled by 50.59 points [47.22,54.06] and zero by 38.80 [36.30,41.42]. This is a central negative result, not an excluded failed cell: transmission of causally relevant information does not guarantee a robust recipient-compatible representation.

Table 2: Matched single-seed development interventions. The factorial separates averaging at matched rate from presentation length. Calibration shows that averaging changes more than temporal content. These are not test results.

Treatment	Correct WER	Shuffled WER	Zero WER
Local mean, short (mean12)	16.30	109.03	118.73
Local mean, repeated (repeat50)	20.74	110.33	103.38
Stride, short (stride12)	29.74	111.12	125.28
Raw, long (raw50)	76.58	117.45	136.41
Stride + fit-derived norm match	19.94	grounding retained	
Stride + fit-derived moment match	20.08	grounding retained	

Table 3: Categorical development rate curve and fixed test confirmation. Rates are measured development rates, not fitted optima. Small-budget mean4 is not the full-scale mean4 bridge in Table 1.

Pool factor	Measured Hz	Development WER	Selected update
1	50.07	76.58	1000
2	25.07	17.31	1000
3	16.77	14.35	750
4	12.62	16.30	1000
5	10.09	17.12	750
8	6.39	17.88	1000

Test artifact	Correct	Shuffled	Zero
meanpool3	8.68	110.15	185.56
raw50	59.48	107.71	168.29
mean4	9.66	109.65	134.10

All eight learned systems, including the temporal block, beat both controls overall and within each source corpus. This does not imply uniform language robustness. On 52 Danish examples, Omni-to-Whisper correct/shuffled/zero WER is 105.93/140.47/95.43%; zero-minus-correct is -10.51 points [-29.55,2.76], with no exact matches in any condition. Reliable improvement over zero is not established for this subgroup. Both absolute performance and subgroup controls belong in the interpretation.

4.2 Averaging and presentation are separable, but averaging is not a mechanism

At matched short presentation, local means improve WER over stride selection by 13.45 points. Repeating pooled states back to a long sequence loses 4.45 points relative to presenting those states briefly. Thus fewer positions alone do not explain the averaging advantage, while approximately fixed pooled content does not eliminate a presentation-length effect (Table 2).

However, averaging also changes norms, coordinate moments, smoothness and optimization conditioning. Norm matching improves stride WER by 9.80 points [6.69,13.38], closing 72.9% of the stride-to-mean gap. Moment matching improves it by 9.67 points [5.22,14.12], closing 71.9%. Both pass the predeclared material-closure criterion and retain source grounding. Much of what appears to be a temporal-pooling advantage is therefore recoverable through first-order statistics under this finite training budget. The residual could reflect local integration, higher-order statistics or optimization differences; it does not uniquely identify any one of them. A linear bridge's ability to absorb affine calibration in principle does not guarantee equivalent finite-budget learning.

4.3 A SELECTED LOW-RATE RECIPE GENERALIZES WITHOUT ANOTHER SEARCH

The development curve selects factor3 among six measured settings. Factor2 improves over raw50 by 59.27 points [53.77,64.91]; factor3 improves a further 2.96 [1.57,4.74]. Adjacent differences among factors4/5/8 are smaller and their intervals include zero. These categorical observations support a low-rate region, not a universal continuous optimum.

On the final test, the fixed meanpool3 artifact reaches 8.68%, compared with 59.48% for raw50: the primary improvement is 50.80 points [49.03,52.57]. The advantage and its interval remain positive within every source and both language strata. The secondary mean4-minus-meanpool3 contrast is 0.98 points [0.45,1.51] overall, while several subgroup intervals cross zero. This confirms the chosen recipe's held-out advantage without another rate sweep. It does not disentangle pooling from calibration, nor does it compare the small-budget factor3 model with full-scale factor4 training at equal compute.

4.4 THE SPECIALIST RECIPIENT ALSO IMPOSES A NATIVE MEMORY CONTRACT

Even Whisper's own states become difficult to decode when their native memory presentation is changed. On public validation, full padded memory at 1,500 positions reaches 5.53% WER; pooling it by four reaches 25.11%. Keeping only the valid prefix reaches 213.66%, or 255.11% after pooling. The 5.53% reference is the explicit-memory-injection evaluation and is distinct from the historical 5.31% standard-generation reference. We do not silently mix them.

Post-training expansion of the coordinate-only Omni-to-Whisper bridge is not a rescue: repetition changes validation WER from 62.72% to 62.45%, interpolation to 66.53%, and interpolation with zero padding to 63.70%. Repetition also fails the strict 99% normalized-output invariance requirement (97.33%). Matched training with a zero tail reaches 57.83%, and with a learned, sample-invariant tail 56.47%. Both miss the 15% rescue criterion and the ten-point meaningful-support criterion; their 1.36-point difference is inconclusive. These failures constrain interpretation without proving an intrinsic inability of Whisper to accept foreign representations.

5 RELATED WORK AND INTERPRETATION

Model stitching tests functional compatibility between frozen components (Bansal et al., 2021), but successful stitching need not demonstrate equal information (Smith et al., 2025). We therefore use task performance and causal controls to characterize a directed interface, not to claim representational identity. Frozen multimodal models (Tsimpoukelli et al., 2021; Li et al., 2023) establish the broader connector paradigm; our contribution is an experimental decomposition rather than a new multimodal architecture.

SLAM-ASR already connects frozen speech encoders and language models using a linear projector (Ma et al., 2024). Alignment-only audio-language training is also established by Zhou et al. (2026). Strong ASR from a small connector or the usefulness of reduced audio rate is not itself our novelty. The differentiating evidence is the conjunction of recipient-selective transfer, source-dependence controls, a matched averaging/presentation factorial, and fit-only calibration with a fixed-artifact held-out confirmation.

A parallel literature asks whether independently trained representations are converging. Huh et al. (2024) argue that models increasingly measure distance between datapoints alike, hypothesize a shared statistical model of reality, and draw a practical corollary: "when two representations are aligned, transitioning from one to the other should be a simple function that's easily obtained." They cite cross-modal stitching as evidence, including a single linear projection that connects a vision model to a language model, and hold that aligned models "can readily process *representations* of new modalities."

Our Omni-to-Whisper cell tests that corollary directly. The transition is the simple function it describes, an affine bridge, and the source information reaches the recipient: correct input beats shuffled by 50.59 points [47.22,54.06] and zero by 38.80 points. Word error rate nevertheless remains at 63.40% against 4.79% native, and post-training expansion leaves it there. Three properties therefore come apart in one measured system: structural alignment, causal influence on the recipient, and

usable performance. The authors anticipate part of this boundary. They restrict their argument to "bijective projections," exclude "lossy or stochastic observation functions," and allow that special-purpose intelligences might not converge. A specialist speech decoder falls under that exclusion, which places our result inside a limitation the hypothesis already admits.

Recent work finds the convergence evidence itself to be regime-dependent: alignment degrades as evaluation scales and survives as coarse semantic overlap, a pattern that extends to text–audio pairings (Koepke et al., 2026). Our measurements approach the same question through task performance, and they also locate the missing ingredient. Supplying Whisper's *own* encoder states to Whisper's own decoder on public validation, word error rate is 5.53% under the native full padded memory of 1,500 positions, 25.11% when that memory is pooled by four, and 213.66% when only the valid prefix is kept. Source and recipient are fixed across those three conditions, so presentation alone spans the range. A recipient can thus refuse its own representation once the presentation contract is broken. Usability depends on an interface contract over sequence length, padding and feature statistics, and similarity between representations leaves that contract unmeasured.

Similarity indices such as CKA (Kornblith et al., 2019) compare representations while leaving open whether either is consumable by a particular recipient, and stitchability can itself mislead about shared information (Smith et al., 2025). Relative representations take a different route, encoding each sample by its similarities to shared anchors and achieving zero-shot stitching under latent isometries and rescalings (Moschella et al., 2023). Our calibration result addresses the same question from the recipient's side. Were invariance to rescaling sufficient, fixed norm or coordinate-moment matching would leave a presentation effect largely intact; instead it recovers about 72% of the pooling advantage. The recipient responds to first-order feature statistics, which those invariances discard.

Work on continuous text compression shows that vector norm can matter to frozen language-model decoding (Zeng & Tan, 2026). Our calibration controls supply a related acoustic result, without asserting an identical mechanism or direction of norm change. Taken together, the experiments suggest three distinctions: source information may be *present*, may *affect* a recipient, and may be *useful* under a particular temporal/statistical interface. None of these alone proves universal compatibility, and convergence of measured structure does not imply any of them.

6 Limitations and conclusion

The matrix is small, acoustic, and asymmetric. Recipient architecture, objectives and injection surfaces are confounded; we identify tested interfaces, not a universal ranking of models. Main Qwen-facing replication uses two seeds, while the temporal factorial uses one. Subgroup sizes are unequal and Danish evidence is limited. The dataset is a custom mixture, not a state-of-the-art ASR benchmark. Validation and development were used for selection; the final test confirms only the frozen allocation. Historical checkpoint and runtime limitations remain explicit in Appendix C. Agreement of independently implemented scorers cannot detect a shared specification defect.

Within these limits, small interfaces can transplant useful acoustic structure between independently pretrained frozen models, but portability is selective. Source dependence alone does not guarantee usability. Local averaging and recipient-facing sequence length have separable effects, and first-order calibration explains much of the averaging advantage. The practical implication is to treat representation transfer as an interface problem involving coordinates, statistics and sequence presentation, rather than infer compatibility from dimensionality or a successful connector alone.

Reproducibility statement

The experiment record binds data manifests, donors, selected artifacts, endpoint revisions, source files and reporting rules. Final predictions are immutable; both scorers recompute all reported comparisons from the same complete export. Appendices describe decoding, data allocation, provenance limitations and compute. The accompanying local research package contains aggregate reports and hashes; public availability and redistribution approval are not claimed here.

ETHICS STATEMENT

Only public corpora are used; their licences and transcript provenance are recorded per source. Upstream licences do not automatically authorize every redistribution. Foundation-model training exposure is unknown. Unequal language coverage and recognition errors limit consequential uses; generated text may be incorrect or offensive. The experiments do not establish suitability for surveillance, identity inference or high-stakes transcription.

AI USE STATEMENT

AI coding assistants were used substantially to propose and refine hypotheses, design and critique experiments, implement methods and dataset processing, diagnose failures, write tests and analysis code, interpret results, check literature, and draft and edit the manuscript and tables. Independent scoring refers to separately implemented software, not independent human replication. Responsibility for scientific claims, citations, authorship, data use and submission remains with the human authors. This statement requires their final review and attestation before submission.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech, 2022. URL https://arxiv.org/abs/2205.12446.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024.

A. Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A. Efros. Back into plato's cave: Examining cross-modal representational convergence at scale, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence, 2015. URL https://arxiv.org/abs/1411.5908.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, and Xie Chen. An embarrassingly simple approach for llm with strong asr capacity, 2024. URL https://arxiv.org/abs/2402.08846.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 55972–55998. PMLR, 2025. URL https://proceedings.mlr.press/v267/smith25a.html.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/01b7575c38dac42f3cfb7d500438b875-Abstract.html.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021. URL https://arxiv.org/abs/2101.00390.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition, 2018. URL https://arxiv.org/abs/1804.03209.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Yunsheng Zeng and Yongmei Tan. Frozen LLMs are native decoders for high-norm semantic vectors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37028–37043. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1717. URL https://aclanthology.org/2026.acl-long.1717/.

Xuanru Zhou, Yiwen Shao, Jiahong Li, and Dong Yu. Alignment is all you need: Instruction-free training for general audio-language models, 2026. URL https://arxiv.org/abs/2608.18132.

A Data allocation and reporting details

Source	Train	Validation	Test
VCTK 0.92	1200	180	300
LibriSpeech	1500	170	300
AMI Edinburgh	1500	0	0
Speech Commands 0.02	600	90	150
FLEURS English/Danish	500	80	100
VoxPopuli English	700	80	150

Table 4: Fixed corpus allocation. AMI is training-only because earlier project development had exposed its room domains.

The final bootstrap samples paired utterances in the bound population order with Python's seeded random generator. The sorted lower and upper indices are 249 and 9749 for 10,000 draws. Every contrast is reported overall and in each allocated source/language stratum. Empty normalized references are invalid execution, not silently dropped rows. Teacher-forced losses are descriptive cell means and do not select a generated result. Exact-match and CER results accompany WER in the aggregate reports. The 25 planned contrasts include 16 source-control contrasts, five learned-primary-minus-native contrasts, two source-family comparisons, and two temporal comparisons. No familywise adjustment or equivalence margin was predeclared.

B Fixed artifacts

The following artifacts were selected before the test allocation was evaluated. The Qwen-facing bridges each contain 6,558,720 FP32 parameters; the Whisper-recipient bridge contains 1,639,680. SHA-256 values identify serialized checkpoint bytes, not canonical tensor-only hashes. Native Whisper has no learned bridge checkpoint.

Fixed artifact	Selected update
`omni_qwen_42` `f39c2622bac75481fcef54ce8529384990afaa024ca212ac2931a6e035917fba`	3000
`omni_qwen_69` `816facaba5507893cdff3d3b3be61eb27fc82169694c59cb38db85022e0bbef7`	4500
`whisper_qwen_42` `8a7430b80df2e8c1ffcd9463b1fb9c9f5de897bc8c0cc57b1966d09368aa876b`	3000
`whisper_qwen_69` `d2f806f50406cff7f0f8c957fd51c1c9eb496948a31ab41834ff72af794c4c4f`	5000
`omni_whisper_v2` `74bf03832596b4526a72730480fa51d0142aaa6efc23d4c510e9ed772fb4ff57`	12000
`temporal_meanpool3` `1b429212961f7734611eec1e468add15bacfd8d9390c8a6b5cbe12abeb9c6bec`	750
`temporal_raw50` `ea5616f9a0dc2c2b0b059e02599576b2849d09cc200a90b07a2be37b8837f64b`	1000
`temporal_mean4` `25bfdb8724c3a615119725ecdbc0abe80506ea88dbe028a9093047f7d8ae492d`	1000

C Provenance, numerical identity, and compute

Historical validation reports are preserved rather than overwritten by the new test. Omni-to-Qwen seed42's historical evaluation-to-checkpoint byte binding remains unresolved. A present-day replay

of the recovered artifact reached 9.70% correct validation WER versus historical 9.57%; shuffled was 108.37% versus 108.39%, and zero 152.87% versus 195.66%. The cause of the zero-input drift is not isolated. Current hashes do not establish what an old process loaded. A recovered ledger associates the Whisper-to-Qwen seed42 reports with its checkpoint, but is not independent runtime attestation.

A later public implementation comparison agreed on all transcripts, error counts and 1,176 learned-path loss pairs but failed its strict byte criterion on two pairs. The observed difference was signed-zero bits in four of 87 old/fresh raw Whisper tensors, including two donor-only rows. That failure is retained. The final test used a separately frozen single-path extraction and raw-byte cache identity; it neither retroactively passes the old gate nor regenerates inputs during scoring.

The final run records PyTorch 2.13.0+cu130, CUDA 13.0, Transformers 5.15.0 and bitsandbytes 0.50.1 on an RTX 4090. Historical environments are not uniform: the archived Omni seed42 training record reports PyTorch 2.8.0+cu128. Endpoint repository revisions are:

- Qwen3.8-27B: `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`;
- Qwen3-Omni: `26291f793822fb6be9555850f06dfe95f2d7e695`;
- Whisper-large-v3-turbo: `41f01f3fe87f28c78e2fbf8b568835947dd65ed9`.

The final 25-cell inference took 30,206.92 seconds, plus 131.34 seconds for feature construction. These are process elapsed times, not energy or utilization measurements. Full-scale Whisper-to-Qwen seed69 used 7.17 optimizer GPU-hours; the temporal factorial used 7.59 optimizer and 3.81 evaluation accelerator-hours; six recovered rate/calibration training reports add 10.76 optimizer GPU-hours. These scoped measurements are not an exhaustive total across all historical attempts or provider billing. The final prediction output contains 49,260 files and 48,629,862 bytes, transferred and verified in two owner-side copies before local scoring. The successful scoring step took 766.96 seconds. No model update or additional inference was performed during scoring.