Localizing the Verbalization Bottleneck in Music Quality Prediction

Anonymous ACL submission

Abstract

When multimodal LLMs evaluate music, their judgments must pass through language—but how much evaluative signal survives that passage? We investigate this on SongEval by comparing how well expert quality ratings across five dimensions can be predicted from audio embeddings versus from text embeddings of LLM-generated song descriptions. Text-mediated prediction consistently falls short. Using a shared-backbone encoder pair, we decompose this gap and find that aligning an audio encoder to language does not degrade its predictive value; the loss occurs when evaluation is routed through explicit text descriptions. The pattern holds under both linear and nonlinear probes and is stable across description models, text encoders, and random splits. A prompt-format ablation adds nuance: structured prompts recover far more signal than open-ended ones, showing that the bottleneck is partly a matter of how descriptions are elicited, not only what language can express.

1 Introduction

Multimodal LLMs are increasingly deployed as evaluators of audio content, rating speech quality (Zhang et al., 2026), scoring AI-generated music (Tjandra et al., 2025), benchmarking audio understanding (Wang et al., 2024a), and serving as judges in generation benchmarks (Zheng et al., 2023). In many such systems, evaluation is mediated by language: the model generates textual descriptions or produces judgments shaped by language-trained representations. A natural concern is that this dependence introduces systematic blind spots. Prior work supports the concern: images degrade when described and reconstructed (Kodathala and Vunnam, 2024), visual relationships are lost in language projection (Li et al., 2025), and Nguyen Quang et al. (2026) showed that multimodal LLMs encode perceptual distinctions they fail to express, with the gap especially pronounced for audio. What these studies leave open is *where* in the pipeline the loss occurs: does language-supervised alignment itself degrade the perceptual representation, or does the aligned representation retain the relevant signal while the conversion into explicit text discards it?

We disentangle these two possibilities on SongEval (Yao et al., 2025), a benchmark of 2,399 songs rated by 16 experts across five quality dimensions, by decomposing the prediction gap into an *alignment* component (self-supervised vs. language-supervised encoder) and a *verbalization* component (language-supervised encoder vs. explicit text). Music is a natural testbed: song quality depends on temporal, affective, and holistic properties that cognitive science predicts to be hard to verbalize (Levinson and Majid, 2014; Majid et al., 2018), and a shared-backbone encoder pair (MuQ and MuQ-MuLan (Zhu et al., 2025), which differ only in a contrastive language-alignment head) allows us to isolate language-supervised alignment with architecture held constant.

Our analysis shows that aligning the audio encoder to language does not degrade its predictive value; the loss is concentrated in the text-mediated pipeline. This conclusion is stable under both linear and nonlinear probes, across description models and text encoders. A prompt-format ablation further reveals that the gap depends substantially on how descriptions are elicited, indicating that the bottleneck is not a fixed property of the dimensions being evaluated. We note that these results characterize recoverability under a specific pipeline (LLM description $\rightarrow$ text encoder $\rightarrow$ probe), not the inherent limits of language.

2 Related Work

The idea that some perceptual experiences resist verbal encoding has deep roots in cognitive science. Levinson and Majid (2014) formalized this

intuition as *differential ineffability*, and Majid et al. (2018) gave it empirical grounding across 20 languages, finding that auditory properties are among the hardest to name in any of them. This asymmetry has begun to surface in multimodal AI as well: images degrade when described and reconstructed (Kodathala and Vunnam, 2024), visual relationships are lost in language projection (Li et al., 2025; Koh et al., 2020), and, most relevant to our setting, Nguyen Quang et al. (2026) showed that omnimodal LLMs encode perceptual distinctions they cannot articulate, with the gap especially pronounced for audio. These studies establish *that* routing perception through language loses information, but leave open *where* in the pipeline the loss occurs.

That question has practical stakes for audio evaluation. The LLM-as-judge paradigm has expanded rapidly to speech quality (Zhang et al., 2026) and music aesthetics (Tjandra et al., 2025), yet existing evaluations report aggregate performance without decomposing the source of errors. Within music specifically, recent work has begun to map the terrain: Carone et al. (2026) and Weck et al. (2024) probed what multimodal models understand about music perception, Zang et al. (2025) showed that text-only LLMs achieve competitive accuracy on music QA benchmarks without any audio input, Wang et al. (2024b) examined the limits of textual music description, and Zhu et al. (2026) developed embedding-based quality metrics that bypass language altogether. Our work draws on contrastive audio-text alignment (Huang et al., 2022) and complements these efforts by localizing the loss, distinguishing what is sacrificed at alignment from what is sacrificed at verbalization.

3 Method

Dataset. We use SongEval (Yao et al., 2025), a collection of 2,399 songs rated by 16 expert listeners on five dimensions (vocal naturalness, memorability, coherence, structural clarity, and overall musicality), each on a 1–5 scale. We split the data into 2,000 training and 399 test songs (seed 42) and verify stability across seeds 1–5.

Description generation. Each song is described by Gemini 2.5 Pro (Comanici et al., 2025) and Qwen3.5-Omni-Plus (Qwen Team, 2026) using a structured prompt that elicits 2–3 sentences on each of five facets: vocal quality, melody/rhythm, harmony, structure, and overall impression. Both models describe every song; the main tables report Qwen3.5-Omni-Plus,[1] the most capable open-weight audio-language model at the time of writing, paired with Qwen3-Embedding-8B,[2] the top-performing open-source text encoder. Additional evaluation with Gemini 2.5 Pro as the description model confirms the pattern holds across models.

Prediction conditions (Figure 1). We compare three ways of predicting expert ratings from the same songs. In the first condition (**C1**), we extract 1,024-dimensional embeddings from MuQ-large (Zhu et al., 2025), a self-supervised audio encoder, and fit a probe to predict each dimension's ratings. In the second (**$\mathbf{C1_{LS}}$**), we swap in MuQ-MuLan-large (Zhu et al., 2025), which shares MuQ's backbone but adds a 512-dimensional contrastive projection head trained against language. In the third (**C2**), we route each song through an LLM description and encode the resulting text with Qwen3-Embedding-8B (4,096-d; Qwen Team, 2025), then fit the same probe. All conditions use two probes: ridge regression (5-fold CV, $\alpha \in [10^{-6}, 10^{6}]$) and a single-hidden-layer MLP (128-d ReLU) that projects every encoder to a common dimensionality before regression. The difference $\mathrm{C1}-\mathrm{C1_{LS}}$ isolates the effect of language-supervised alignment; $\mathrm{C1_{LS}}-\mathrm{C2}$ captures what we call the "verbalization term," though this label is shorthand for the full text-mediated pipeline: description generation, text encoding, and probing combined. As a robustness check, we also swap in OpenAI's text-embedding-3-large (3,072-d) as an alternative text encoder.

4 Results

The decomposition in Table 1 tells a clear story. Under ridge regression, the alignment term averages just $+.009$ across the five dimensions, with no confidence interval excluding zero, while the verbalization term averages $+.237$ with all five CIs comfortably above zero. Switching to an MLP probe does not change the picture: alignment stays near zero ($+.016$, still 0/5 significant) and the verbalization term, if anything, grows slightly ($+.249$). Since the MLP projects each encoder to a shared 128-d space before regression, this comparison also rules out the possibility that the near-zero alignment gap is an artifact of the 1,024-to-512 dimensionality change between MuQ and MuQ-MuLan. Across the five dimensions, coherence shows the largest

[1] https://qwen.ai/blog?id=qwen3.5-omni
[2] https://qwen.ai/blog?id=qwen3-embedding

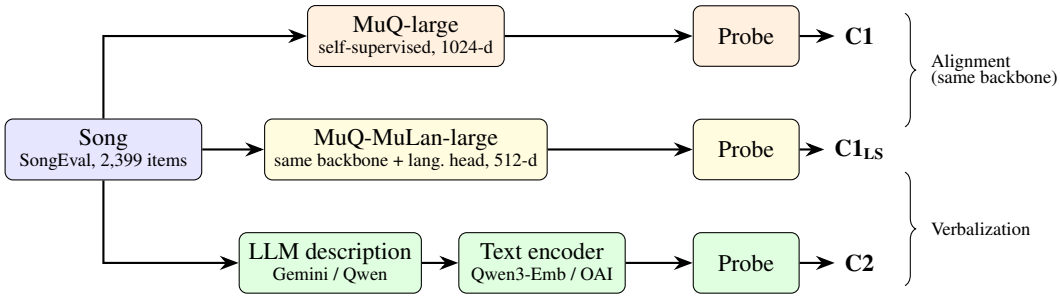

Figure 1: Three-way decomposition for music. Total gap C1−C2 = alignment (C1−$C1_{LS}$) + verbalization ($C1_{LS}$−C2). MuQ and MuQ-MuLan share the same audio backbone.

	Ridge			MLP		
Dimension	**Align**	**Verb**	**CI_v**	**Align**	**Verb**	**CI_v**
Vocal nat.	+.015	+.212	[.16,.27]	+.018	+.224	[.17,.28]
Memorab.	+.008	+.244	[.19,.30]	+.013	+.259	[.20,.32]
Coherence	+.000	+.265	[.21,.33]	+.009	+.278	[.22,.34]
Struct. cl.	+.010	+.241	[.19,.30]	+.019	+.253	[.20,.31]
Overall	+.011	+.225	[.18,.28]	+.020	+.233	[.18,.29]
Mean	**+.009**	**+.237**		**+.016**	**+.249**	

Table 1: Three-way decomposition under ridge and MLP probes. **Align**: C1−$C1_{LS}$. **Verb**: $C1_{LS}$−C2. CI_v: bootstrapped 95% CI for verbalization (all exclude zero). Alignment CIs all include zero under both probes (Appendix B). The MLP projects all encoders to a common 128-d, eliminating the 1024/512 dimensionality difference. C1 ranges from .880–.895; C2 from .615–.659 (ridge).

Dimension	**C1**	**Struct**	**Open**	**Δ**
Vocal nat.	.885	.658	.479	+.179
Memorab.	.882	.631	.463	+.168
Coherence	.880	.615	.468	+.147
Struct. cl.	.885	.635	.461	+.173
Overall	.895	.659	.459	+.200
Mean	**.885**	**.640**	**.466**	**+.174**

Table 2: Prompt-format ablation (Qwen + Qwen3-Emb). **C1**: self-supervised ceiling. Δ: Struct − Open. In paintings, the effect reverses (Appendix D).

verbalization term (+.265 ridge, +.278 MLP) and overall musicality the smallest (+.225, +.233), but the spread is narrow (σ = .020). The gap is not driven by task difficulty: MuQ reaches .880–.895 SRCC on every dimension, so the evaluative signal is present in the audio; the text-mediated pipeline simply does not recover it. A secondary self-supervised encoder (MERT-v1-330M; Li et al., 2024) achieves C1 = .856–.869 with the same gap ordering, confirming the pattern is not idiosyncratic to MuQ.

Dimension	**Gemini**	**Qwen**	**$\lvert\Delta\rvert$**
Vocal nat.	.726	.694	.032
Memorab.	.718	.671	.047
Coherence	.716	.655	.061
Struct. cl.	.717	.672	.045
Overall	.730	.684	.046

Table 3: C2 SRCC by description model (best of two text encoders). Gemini outperforms Qwen, but the dimension-level profile is preserved ($\rho > .95$). Across text encoders ($\rho > .95$) and 6 splits ($\tau \geq .91$), the pattern is stable. Gemini-primary decomposition: verb +.159, align +.009.

How much depends on elicitation? The prompt-format ablation in Table 2 suggests: quite a lot. Switching from structured to open-ended prompts degrades C2 from .640 to .466, widening the verbalization gap by +.174 on average. The structured format was selected before the main analysis, and the magnitude of this difference was unexpected. Open-ended descriptions tend toward generic aesthetic commentary ("a pleasant and well-produced track") that fails to differentiate songs; structured prompts force coverage of each evaluative facet. In paintings, the effect reverses (Appendix D), suggesting that the optimal elicitation strategy is domain-dependent.

Stability across models and splits. Gemini 2.5 Pro produces stronger descriptions than Qwen3.5-Omni-Plus on every dimension (Table 3, $\lvert\Delta\rvert$ = .032–.061), but the dimension-level ordering is preserved ($\rho > .95$): which dimensions lose the most signal is not an artifact of the description model. Under a Gemini-primary decomposition, the verbalization term shrinks to +.159 but alignment remains near zero (+.009), confirming that a better description model narrows the gap without eliminating it. The pattern is also stable across the two text encoders ($\rho > .95$) and six random

Dimension	ZS_{dir}	ZS_{txt}	Δ
Vocal nat.	.349	.259	+.090
Memorab.	.368	.146	+.222
Coherence	.536	.164	+.372
Struct. cl.	.606	.299	+.307
Overall	.380	.147	+.233

Table 4: Gemini 2.5 Pro zero-shot SRCC (no learned probe). ZS_{dir}: scores stimulus directly. ZS_{txt}: scores from own description. Δ conflates verbalization loss with task-format differences, but the direction is consistent with the probe-based decomposition.

train/test splits ($\tau \geq .91$).

What descriptions miss. An inspection of 50 descriptions from the top and bottom rating quartiles reveals two recurring failure modes (Appendix C). *Categorical flattening*: high- and low-rated songs receive nearly interchangeable feature descriptions, collapsing continuous perceptual gradients into coarse verbal categories. *Evaluative hedging*: descriptions use similarly positive language regardless of quality, leaving execution gaps unrecorded. Both patterns are especially stark for coherence, the dimension with the largest verbalization gap, where descriptions name structural elements without conveying how well the piece holds together as a whole.

Zero-shot validation. Table 4 offers a comparison closer to deployed LLM-as-judge systems: Gemini 2.5 Pro scoring each song from audio directly or from its own text description, with no learned probe. SRCC drops on every dimension when the model works from text, with the largest drops on coherence (.372) and structural clarity (.307), the same dimensions with the largest probe-based verbalization terms. This comparison is confounded by task format (direct scoring vs. description generation), but the consistent direction across all five dimensions reinforces the main finding.

5 Discussion

The central finding is consistent across every axis of variation we test: the performance gap between perceptual embeddings and text-mediated prediction in music evaluation falls after the language-aligned audio stage, not at alignment itself. What is perhaps most striking is the uniformity of the effect: all five dimensions lose similar amounts of signal ($\sigma = .020$), even though vocal naturalness, memorability, and coherence seem like qualitatively different kinds of judgment. Whether this uniformity reflects a shared bottleneck in how LLMs describe music or is coincidental awaits further investigation. The prompt-format ablation adds nuance: the gap can be substantially narrowed by structured elicitation, but even the best pipeline (Gemini with structured prompts) leaves a sizable residual, suggesting that better prompting alone will not close it entirely.

For the growing use of LLMs as music evaluators (Tjandra et al., 2025; Zhang et al., 2026), these results raise a practical concern: routing evaluation through text may systematically underweight holistic qualities relative to what perceptual embeddings capture. This complements Zang et al. (2025), who found that text-only LLMs match audio-language models on music QA benchmarks, suggesting that existing pipelines may reward textual reasoning over genuine perception from both directions. If such scores serve as reward signal for music generation, the asymmetry could bias generators toward verbally legible features at the expense of qualities like coherence and memorability. The bridge from our probe-based setup to deployed judgment quality, however, remains to be built.

Limitations

Our main analysis covers a single domain (music) with five quality dimensions. Cross-domain results on paintings (Appendix D) show the same verbalization gap across all ten dimensions, but the encoder comparison there is less tightly controlled because no shared-backbone pair exists for vision. All descriptions are LLM-generated; human-authored descriptions would be needed to disentangle model-specific generation habits from constraints inherent to language. The measured gap varies with description model, text encoder, and prompt format, so the specific magnitudes we report should be read as properties of the pipeline rather than of the dimensions themselves.

Societal considerations. Text-mediated evaluation may systematically undervalue music from traditions whose quality resists verbalization in English-centric registers, particularly traditions where timbral nuance or microtonal inflection carries evaluative weight that standard English vocabulary does not differentiate well. If such evaluators inform reward modeling for music generation, these biases risk narrowing the stylistic diversity of the music that gets promoted and produced.

References

Giuseppe Carone and 1 others. 2026. LLMs can read music, but struggle to hear it. an evaluation of core music perception tasks. *arXiv preprint*.

Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Qingqing Huang, Aren Jansen, Honglak Lee, Ravi Plakal, Matthew Sharifi, and Marco Tagliasacchi. 2022. MuLan: A joint embedding of music audio and natural language. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Xin Jin, Qianqian Qiao, Yi Lu, Huaye Wang, Heng Huang, Shan Gao, Jianfei Liu, and Rui Li. 2024. APDDv2: Aesthetics of paintings and drawings dataset with artist labeled scores and comments. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Sai Varun Kodathala and Rakesh Vunnam. 2024. The describe-then-generate bottleneck: How VLM descriptions alter image generation outcomes. *arXiv preprint arXiv:2509.18179*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, pages 5338–5348.

Stephen C. Levinson and Asifa Majid. 2014. Differential ineffability and the senses. *Mind & Language*, 29(4):407–427.

Wenyan Li, Raphael Tang, Chengzu Li, Caiqi Zhang, Ivan Vulić, and Anders Søgaard. 2025. Lost in embeddings: Information loss in vision-language models. *arXiv preprint arXiv:2509.11986*.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, and Jie Fu. 2024. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations (ICLR)*.

Asifa Majid, Seán G. Roberts, Ludy Cilissen, Karen Emmorey, Brenda Nicodemus, Lucinda O'Grady, Bencie Woll, Barbara LeLan, Hilario de Sousa, Brian L. Cansler, Shakila Shayan, Connie de Vos, Gunter Senft, N.J. Enfield, Rogayah A. Razak, Sebastian Fedden, Sylvia Tufvesson, Mark Dingemanse, Ozge Ozturk, and 7 others. 2018. Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, 115(45):11369–11376.

Trung Nguyen Quang, Yiming Gao, Fanyi Pu, Kaichen Zhang, Shuo Sun, and Ziwei Liu. 2026. Senses wide shut: A representation–action gap in omnimodal LLMs. *arXiv preprint arXiv:2605.13737*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, and 7 others. 2025. DINOv3: A unified framework for self-supervised learning. *arXiv preprint*.

Qwen Team. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Qwen Team. 2026. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. AudioBench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Zihao Wang and 1 others. 2024b. MuChin: A chinese colloquial description benchmark for evaluating language models in the field of music. *arXiv preprint arXiv:2402.09871*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, György Fazekas, and Dmitrij Gašić. 2024. MuChoMusic: Evaluating music understanding in multimodal audio-language models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*.

Jixun Yao, Guobin Ma, Huixin Xue, Hangting Chen, Chongjia Hao, Yu Jiang, Hexin Liu, Ruibin Yuan, Junzhan Xu, Wei Xue, Hao Liu, and Lei Xie. 2025. SongEval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. 2025. Are you really listening? boosting perceptual awareness in music-QA benchmarks. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*.

Leying Zhang, Bowen Shi, Haibin Wu, Bach Viet Do, and Yanmin Qian. 2026. JASTIN: Aligning LLMs for zero-shot audio and speech evaluation via natural language instructions. *arXiv preprint arXiv:2605.04505*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. 2025. MuQ: Self-supervised music representation learning with mel residual vector quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:3653–3666.

Haina Zhu and 1 others. 2026. MuQ-Eval: An open-source per-sample quality metric for AI music generation evaluation. *arXiv preprint*.

A Experimental Details

Models. Self-supervised: MuQ-large (1,024-d; Zhu et al., 2025), MERT-v1-330M (1,024-d; Li et al., 2024). Language-supervised: MuQ-MuLan-large (512-d; same backbone as MuQ; Zhu et al., 2025). Description: Gemini 2.5 Pro, Qwen3.5-Omni-Plus. Text encoders: Qwen3-Embedding-8B (4,096-d; Qwen Team, 2025), text-embedding-3-large (3,072-d). All accessed May 2026.

Prompts. Structured: "You are an expert music critic. Listen to the provided song and write a structured evaluation covering: (1) Vocal quality and naturalness, (2) Melodic and rhythmic elements, (3) Harmonic structure, (4) Structural organization, (5) Overall impression. Write 2–3 sentences per facet." Open-ended: "Describe the song in detail."

Ridge. Scikit-learn `RidgeCV`, 50 log-spaced $\alpha \in [10^{-6}, 10^{6}]$, 5-fold CV, `StandardScaler`. Seed 42; stability seeds 1–5. Bootstrap: 1,000 test-set resamples.

MLP probe. 1-hidden-layer (128 ReLU, dropout 0.2, weight decay 10^{-4}), Adam (lr 10^{-3}, batch 64, early stopping patience 10), same CV and bootstrap. Projects each encoder to common 128-d.

B Component-Level CIs (Ridge)

Dimension	Gap	95% CI
Total gap (C1−C2):		
Vocal nat.	+.227	[**+.176**, +.283]
Memorab.	+.252	[**+.197**, +.312]
Coherence	+.265	[**+.204**, +.326]
Struct. cl.	+.251	[**+.196**, +.311]
Overall	+.236	[**+.183**, +.294]
Alignment ($C1-C1_{LS}$):		
Vocal nat.	+.015	[−.002, +.034]
Memorab.	+.008	[−.009, +.024]
Coherence	+.000	[−.016, +.016]
Struct. cl.	+.010	[−.005, +.026]
Overall	+.011	[−.004, +.026]
Verbalization ($C1_{LS}-C2$):		
Vocal nat.	+.212	[**+.162**, +.266]
Memorab.	+.244	[**+.192**, +.303]
Coherence	+.265	[**+.205**, +.327]
Struct. cl.	+.241	[**+.191**, +.301]
Overall	+.226	[**+.175**, +.283]

Table 5: Bootstrapped 95% CIs under ridge ($n = 1{,}000$). Bold: CI excludes zero.

C Qualitative Analysis

We inspected 50 Gemini descriptions from top and bottom rating quartiles. This analysis is illustrative, not systematic. Three patterns recur:

Categorical flattening. Descriptions name relevant features but collapse continuous variation into coarse bins. Songs rated 4.5 and 1.2 on memorability receive nearly interchangeable descriptions ("smooth, clear tenor with a slightly breathy quality" vs. "smooth, clear tenor voice with a warm, slightly breathy tone").

Evaluative hedging. High- and low-rated songs receive similarly positive evaluative language ("demonstrating strong artistic vision" vs. "demonstrates notable creative ambition").

Holistic compression. For overall musicality, descriptions itemize components but do not convey how well they integrate.

D Cross-Domain Evaluation: Paintings

We repeat the analysis on APDDv2 (Jin et al., 2024), 2,400 paintings subsampled from 10,023 artworks (10 dimensions). Self-supervised: DINOv3 ViT-L (Oquab et al., 2025). Language-supervised: SigLIP 2 ViT-L (Tschannen et al., 2025). Unlike music, the encoders differ beyond language supervision, making the decomposition less diagnostic.

The verbalization gap is positive for all 10 dimensions and exceeds the total gap because negative alignment partially compensates. Prompt-format effects reverse: open-ended outperforms structured (mean $\Delta = -.073$).

Dim.	C1	$C1_{LS}$	C2	Align	Verb
Theme/log.	.647	.712	.535	$-.065$	+.177
Creativity	.645	.756	.501	$-.111$	+.255
Layout/c.	.684	.741	.576	$-.057$	+.165
Space/p.	.671	.692	.573	$-.020$	+.119
Order	.671	.735	.557	$-.064$	+.178
Light/sh.	.702	.726	.617	$-.024$	+.109
Color	.691	.729	.609	$-.038$	+.120
Detail/t.	.754	.801	.657	$-.047$	+.144
Mood	.642	.698	.518	$-.056$	+.180
Overall	.717	.761	.616	$-.044$	+.145
Mean	**.682**	**.734**	**.576**	**$-.053$**	**+.159**

Table 6: Three-way decomposition for paintings (ridge). Verbalization positive for 10/10. Alignment negative for 10/10, likely because the language-supervised encoder (SigLIP 2) is architecturally distinct from the self-supervised one (DINOv3), unlike the shared-backbone pair used for music. Under MLP: verb $+.205$, align $-.073$.