Do Higher CLAP Scores Mean Better Audio-Text Matching? A Hachimi Lyric Study

Anonymous ACL submission

Abstract

CLAP-style audio-text models typically use similarity scores between audio and text to determine whether the two modalities are matched. However, whether a higher similarity score truly indicates stronger semantic correspondence remains insufficiently examined. This paper focuses on the "hachimi" song-remix phenomenon on the Chinese internet, in which original lyrics are often replaced by approximate pronunciations, repeated syllables, or sung phrases without stable lexical meaning. Such texts preserve the rhythm and form of lyrics while weakening explicit semantics. Based on this phenomenon, this paper investigates a counterintuitive result: in a Chinese hachimi song dataset, hachimi text receives higher CLAP similarity than semantically meaningful original lyrics. To understand this phenomenon, we construct different text conditions along dimensions such as semantics, token entropy, and caption style, and analyze how these changes in surface form affect model scores. The results show that the high scores assigned to hachimi text do not indicate that the model understands the audio content more accurately. Rather, they may arise because such text, through its character composition and tokenization pattern, is more easily encoded by CLAP into text vectors that are closer to audio representations. Therefore, this paper shows that raw CLAP similarity scores should be interpreted with caution: a higher score may be influenced by surface-form properties rather than stronger audio-text semantic matching.

1 Introduction

CLAP-style audio-text models reduce audio-text comparison to a cosine similarity score between text and audio embeddings. This score is widely used as a proxy for audio-text matching, but its interpretation is not always straightforward.

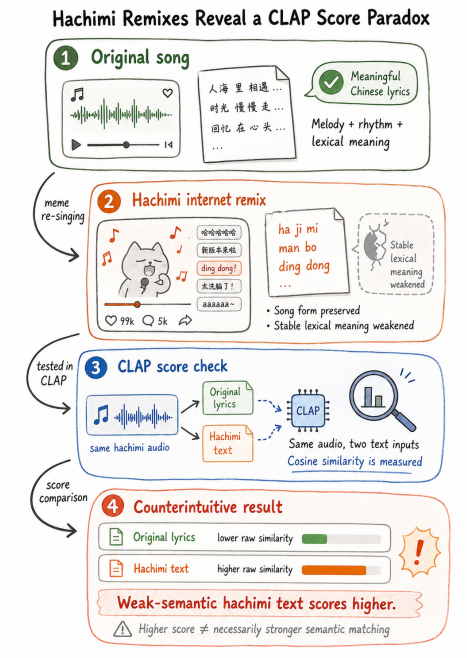

Figure 1: Motivating example of score elevation for weak-semantic hachimi text.

When one text receives a higher score than another for the same audio, it is tempting to assume that the higher-scoring text better corresponds to the audio. This paper examines a counterintuitive case where that assumption becomes difficult to sustain.

Using a 91-song Chinese hachimi dataset, we find that weak-semantic hachimi text receives higher mean true-pair cosine similarity than original lyrics in the CLAP-style models we test. Hachimi text weakens stable lexical meaning while preserving lyric-form cues such as repetition, syllabic texture, and song-shaped phrasing. The surprising point is not simply that CLAP assigns nonzero similarity to weak-semantic text, but that text with weaker human-interpretable meaning can score higher than meaningful original lyrics. This makes hachimi a useful stress test for score interpretation, since cross-modal models can respond

to form-level associations even when literal lexical meaning is not the primary signal (Alper and Averbuch-Elor, 2023).

This observation motivates our central question: what makes weak-semantic hachimi text score higher than original lyrics? We do not treat the higher score as evidence that hachimi text is a better semantic description of the audio. Instead, we treat it as a phenomenon to explain. In particular, we ask whether the elevation is associated with surface-form properties of the text, and how those properties interact with CLAP's embedding space.

To answer this question, we first regress the hachimi–original similarity gap on surface-form features, including token count, token entropy, repetition, Chinese character ratio, and Latin character ratio. We then use the most stable predictors to design a spectrum of controlled conditions that progressively modifies original lyrics along relevant surface-form dimensions. Among these conditions, the Latin-ratio/full-profile step corresponds to the largest observed raw similarity increase. Finally, we use embedding-space diagnostics to examine how this surface-profile shift moves text into a more tokenizer-friendly regime and why the resulting score gain varies across models and audio sources.

We refer to this pattern as surface-form-driven similarity elevation: a raw CLAP score increase associated with weak-semantic lyric-like text, without sufficient evidence that the text has become a stronger semantic match to the audio. Our contributions are threefold. First, we document a counterintuitive score elevation in which weak-semantic hachimi text scores higher than original lyrics. Second, we identify surface-form predictors, especially Latin character ratio, that correlate with this elevation and test them through controlled condition design. Third, we show that the effect is better understood as an interaction between tokenizer-sensitive text shifts and model-specific embedding behavior than as stronger semantic correspondence alone.

2 Related Work

2.1 Audio-language alignment models

CLIP-style cross-modal alignment has also been extended to audio, including Wav2CLIP, AudioCLIP, and CLAP-style systems that learn joint embeddings for audio and text so that matched pairs receive higher cosine similarity than mismatched ones (Wu et al., 2022; Guzhov et al., 2022; Wu et al., 2023c). This paradigm has been extended to music-language settings (Wu et al., 2023b) and multilingual models (Elizalde et al., 2023), where retrieval and representation alignment are central goals. A recurring finding is that raw cosine similarity alone is insufficient for evaluating model understanding: Wu et al. (2023a) showed that current audio-text models do not reliably exploit natural language richness, while Selvakumar et al. (2025) found that retrieval accuracy degrades substantially under linguistic variation. Our work identifies a complementary failure mode: non-semantic but structurally lyric-like text can raise scores *above* those of semantically meaningful text, without a consistent gain in pair-specific retrieval.

2.2 Tokenization and script effects

Modern text encoders usually operate over subword units rather than full words, a design popularized by byte-pair encoding for rare-word handling (Sennrich et al., 2016). This makes script and segmentation behavior relevant to representation quality: tokenizer quality affects multilingual model performance (Rust et al., 2021), and languages and writing systems can incur unequal tokenization costs under commercial tokenizers (Ahia et al., 2023). These findings motivate our use of token count, token entropy, and Latin-character ratio as surface-form diagnostics for Chinese lyrics and Latinized hachimi text, without treating tokenization alone as a complete explanation of CLAP scores.

2.3 When high cross-modal similarity scores mislead

The concern that cosine similarity overstates semantic correctness is well established in vision-language research. Early work with CLIP (Radford et al., 2021) revealed that strong zero-shot performance coexists with systematic failures on compositional reasoning, attribute binding, negation, and spatial relations. Dedicated compositional benchmarks make this concern more explicit: Winoground tests whether models can distinguish captions with the same words but different compositional relations, while ARO shows that vision-language models can behave like bag-of-words systems with limited sensitivity to attribution, relations, and word order (Thrush et al., 2022; Yuksekgonul et al., 2023). Kang et al. (2025) showed concretely that incorrect captions can outscore correct ones, and attributed this to the model responding to salient surface-level features rather than full

relational meaning. Separately, prior work on embedding anisotropy, common mean vectors, and dominant directions motivates diagnostic checks of representation geometry, although such diagnostics do not by themselves explain CLAP score failures (Mu and Viswanath, 2018; Li et al., 2020; Godey et al., 2024).

Our paper studies a related but distinct failure mode in the audio-text domain. Whereas prior CLIP work mainly shows that *wrong* captions can score high, we show that text with *no* stable meaning can score *higher* than source-song lyrics that preserve stable lexical semantics in CLAP-style models. We argue that this elevation is better explained as a surface-form-sensitive score shift than as evidence of stronger grounding.

3 Method

Figure 2 summarizes the experimental pipeline used to diagnose the surface-form-driven similarity elevation introduced in Figure 1.

3.1 Dataset and Condition Design

We use a cleaned 91-song dataset of paired Chinese music with structured text conditions. The primary comparison is between original lyrics (ORIG) and weak-semantic hachimi text (HACHIMI); bridge probes are used as diagnostic probes. Dataset collection, filtering, and cleaning are described in Appendix A; text condition generation and quality validation are in Appendix B and C. The two primary conditions are:

- **Original lyrics** (ORIG): The original Chinese song lyrics, used as the semantic reference.
- **Hachimi text** (HACHIMI): Phonetically derived syllables that preserve the rhythmic and structural texture of the original lyrics while carrying no stable lexical meaning. Critically, HACHIMI is *not* random noise: repetition, syllabic texture, line structure, and song-shaped phrasing remain intact; only stable word-level semantics are removed.

The dataset spans 91 songs with full-song audio-text comparisons. The primary labels are operational probes rather than semantic gold labels: original lyrics provides the original lyric reference, while hachimi text weakens stable lexical meaning while preserving lyric form. Dataset-level feature statistics are reported in Appendix C. We treat these coupled form changes as part of the phenomenon to diagnose, not as nuisance variation to hide.

All text conditions used in the paper are summarized in Table 1.

Paraphrase (PARA) and ZHNOISE (structure-preserving Chinese-script nonsense) serve as semantic controls used in Section 4.2 to test whether the surface-form elevation is specific to hachimi text. We evaluate an auxiliary caption control (audio-description style) separately in Section 4.3.

3.2 Models and Similarity Measurement

The active model roster comprises three CLAP-style systems, all used as frozen pretrained encoders with no parameter updates: **LAION-CLAP** (Wu et al., 2023c), **LAION-CLAP Music** (the `laion/larger_clap_music` checkpoint), and **MS-CLAP** (`microsoft/msclap`) (Elizalde et al., 2023). Checkpoint identifiers, tokenizer names, and package versions are in Appendix G.

For each model and text condition, audio and text embeddings are computed using the respective frozen encoder, and cosine similarity is computed over ℓ_2-normalized embeddings. For a fixed model m and text condition c, we formalize this computation as a similarity matrix $S^{(m,c)} \in \mathbf{R}^{N\times N}$:

$$S_{ij}^{(m,c)} = \left\langle \frac{g_m(t_i^{(c)})}{\|g_m(t_i^{(c)})\|_2}, \frac{f_m(a_j)}{\|f_m(a_j)\|_2} \right\rangle, \quad (1)$$

where g_m and f_m are the frozen text and audio encoders. Each combination of model, audio source, and text condition yields such an $N \times N$ text-audio similarity matrix ($N = 91$): diagonal entry (i, i) is the *true pair* (text i matched to audio i), and off-diagonal entry (i, j) with $j \neq i$ is a *false pair*. Main matrices use hachimi audio; original audio is used only for robustness checks (Appendix E). The central question is whether score increases across conditions correspond to more reliable pair-specific matching.

3.3 Metrics

All main diagnostics are in the text-to-audio (T2A) direction unless otherwise noted. We first distinguish raw score elevation from pair-specific improvement using three matrix-level quantities: true-pair mean, false-pair mean, and pair-specificity

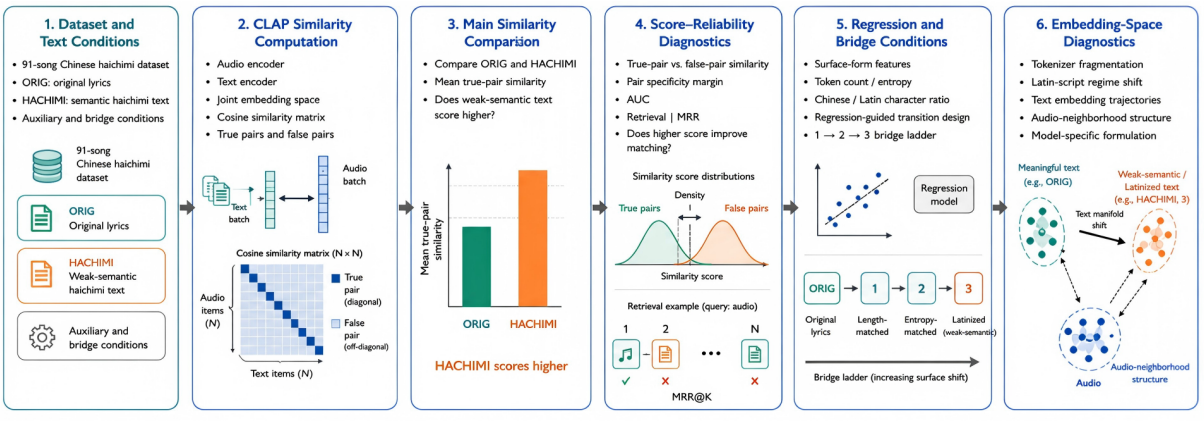

Figure 2: Overall experimental framework for measuring, validating, and diagnosing hachimi-driven similarity elevation.

Name	Construction	Role
ORIG	Original Chinese song lyrics	Semantic reference
HACHIMI	Phonetically derived hachimi text; weakens stable lexical meaning while preserving lyric form	Primary comparison
PARA	Meaning-preserving paraphrase of ORIG	Semantic control
ZHNOISE	Structure-preserving Chinese-script nonsense text	Semantic control
BRIDGE-LEN	Length-matched Chinese text; same token count as HACHIMI	Bridge probe: token count
BRIDGE-ENT	Length + entropy matched; same token count and entropy as HACHIMI	Bridge probe: token count, entropy
BRIDGE-FULL	Weak-semantic Latinized text; progressively matches HACHIMI's per-song token count, entropy, and Latin-character ratio	Bridge probe: full surface profile

Table 1: Text conditions used in the paper. ORIG/HACHIMI form the primary comparison; PARA/ZHNOISE and bridge probes provide controls and diagnostics.

margin:

$$\mathrm{True}(m,c) = \frac{1}{N}\sum_{i=1}^{N} S_{ii}^{(m,c)}, \quad (2)$$

$$\mathrm{False}(m,c) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i} S_{ij}^{(m,c)}, \quad (3)$$

$$\mathrm{Margin}(m,c) = \mathrm{True}(m,c) - \mathrm{False}(m,c). \quad (4)$$

We additionally report empirical AUC and text-to-audio MRR as ranking-based reliability diagnostics. Audio-to-text MRR (A2T MRR) is consistent with T2A results and is defined and reported in Appendix D. Z-score normalization is computed within each model × audio-source × condition similarity matrix. Full metric definitions and formulas are in Appendix D.

3.4 Bridge-Probe Construction

Bridge probes are constructed as a sequential surface-form intervention ladder. Starting from ORIG, we first match HACHIMI's length profile, then token entropy, and finally Latin-character ratio/full surface profile. These probes are not intended to be natural lyric alternatives; they are controlled interventions for testing which surface-form changes reproduce the HACHIMI elevation.

Algorithm 1 Compressed Bridge-Probe Construction

Require: Orig text ℓ_0; Hachimi targets L^*, H^*, R^*
$\ell_{\text{len}} \leftarrow \textsc{MatchLength}(\ell_0, L^*)$
$\ell_{\text{ent}} \leftarrow \textsc{MatchEntropy}(\ell_{\text{len}}, H^*)$
$\ell_{\text{full}} \leftarrow \textsc{MatchLatinProfile}(\ell_{\text{ent}}, R^*)$
return BRIDGE-LEN $= \ell_{\text{len}}$, BRIDGE-ENT $= \ell_{\text{ent}}$, BRIDGE-FULL $= \ell_{\text{full}}$

3.5 Diagnostic Logic

Our analytical strategy separates raw similarity elevation from reliable matching using two diagnostic types.

Score-reliability diagnostics. We compare raw true scores with empirical AUC, pair-specificity margins, z-scored true scores, and text-to-audio retrieval metrics (A2T results in Appendix D). These diagnostics test whether higher scores translate into stronger pair-specific evidence rather than only larger aggregate similarity.

Semantic controls and bridge probes. We use semantic controls, surface-form bridge probes, and embedding-space diagnostics to constrain mechanistic interpretation without elevating exploratory analyses above headline findings.

An exploratory song-level regression over eight surface-form features identifies Latin-character ratio as the most stable positive correlate of the HACHIMI–ORIG similarity gap across models. We use this result to guide bridge-probe construction, not as causal evidence; full Spearman, ridge, OLS, and collinearity diagnostics are reported in Appendix F.

4 Experiments

All experiments use the 91-song dataset and three models: LAION-CLAP, LAION-CLAP Music, and MS-CLAP. The primary comparison is between original lyrics (ORIG) and hachimi text (HACHIMI), with PARA/ZHNOISE treated as semantic controls. All main similarity matrices are computed against hachimi audio; key original-audio comparisons are reported inline, with full robustness details in Appendix E. Bridge probes (BRIDGE-LEN/BRIDGE-ENT/BRIDGE-FULL) and embedding-space diagnostics serve as supporting mechanistic investigations.

4.1 Main Observation and Score Reliability

Table 2 summarizes the main ORIG–HACHIMI comparison across all three models. The headline pattern is consistent: HACHIMI achieves higher true-pair similarity than ORIG across all three models (Figure 3). Under original (non-hachimi) recordings, LAION-CLAP retains a significant gap ($d = 0.22$, $p = 0.036$); LAION-CLAP Music shows a positive but non-significant difference ($d = 0.13$, $p = 0.233$); and MS-CLAP shows no meaningful effect ($d = -0.03$, $p = 0.782$). Thus, the cross-model consistency of Table 2 is specific to the hachimi-audio condition, a distinction further examined in the embedding-space diagnostics.

Figure 3 shows the corresponding song-level distributions. The distributions overlap substantially,

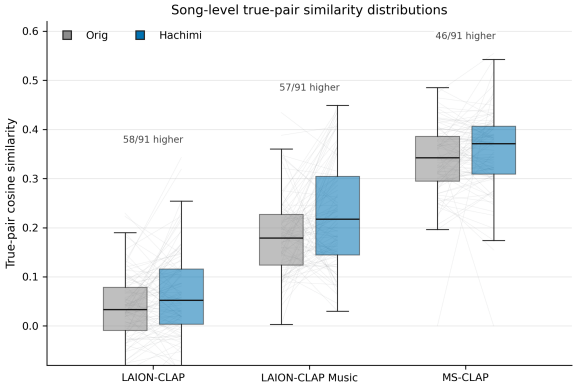

Figure 3: Song-level true-pair similarity distributions on hachimi audio ($N = 91$). Boxes show the distribution over songs, and faint connecting lines link ORIG and HACHIMI scores for the same song.

but the HACHIMI condition shifts upward in aggregate across all three models; this makes the main result a distributional score elevation rather than evidence that every individual song is better matched by HACHIMI.

The central question is not whether hachimi text raises scores, but whether it improves pair-specific matching. Full primary-condition diagnostics are reported in Appendix Table 6. In all three models, HACHIMI has the highest raw true-pair score, but false-pair means rise alongside true-pair means when moving from original lyrics to hachimi text, leaving margins small. The hachimi text elevation is consistent with a *broad affinity shift*: higher similarity to all audio, not specifically the matching one. It is not a consistent improvement in pair-specific correspondence. This section reports score-reliability diagnostics; mechanistic interpretation is developed in Sections 4.2–4.4.

Figure 4 visualizes the same pattern using true-pair and false-pair means. Movement close to the diagonal indicates broad score movement rather than selective improvement for the matching audio. For LAION-CLAP, HACHIMI raises raw true score and modestly improves AUC and T2A MRR, but retrieval remains weak in absolute terms. For LAION-CLAP Music and MS-CLAP, HACHIMI again leads on raw score, while AUC declines and T2A MRR changes only slightly. Across all three models, the raw similarity elevation under HACHIMI is not accompanied by a consistent gain in pair-specificity, discrimination, or retrieval.

Additional relative-ranking comparisons provide supplementary evidence in the same direction and are reported in Appendix E. Audio-to-text ranking

Model	ORIG	HACHIMI	HACHIMI–ORIG gap
LAION-CLAP	0.037	0.064	+0.027
LAION-CLAP Music	0.181	0.227	+0.046
MS-CLAP	0.339	0.354	+0.016

Table 2: Main 91-song hachimi-audio comparison from the retrieval-matrix summary. Hachimi exceeds Orig across all three models.

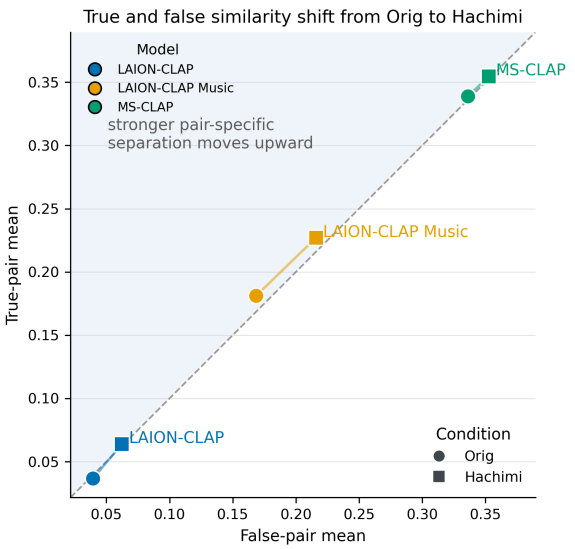

Figure 4: True-pair and false-pair mean similarities for ORIG and HACHIMI. Colors denote models, markers denote conditions, and movement close to the diagonal indicates broad score elevation rather than selective matching improvement.

Model	ORIG	HACHIMI	PARA	ZHNOISE
LAION-CLAP	0.037	0.064	0.032	0.019
LAION-CLAP Music	0.181	0.227	0.174	0.133
MS-CLAP	0.339	0.354	0.340	0.079

Table 3: Semantic-control summary on hachimi audio ($N = 91$). Para remains close to Orig, while ZhNoise is substantially lower.

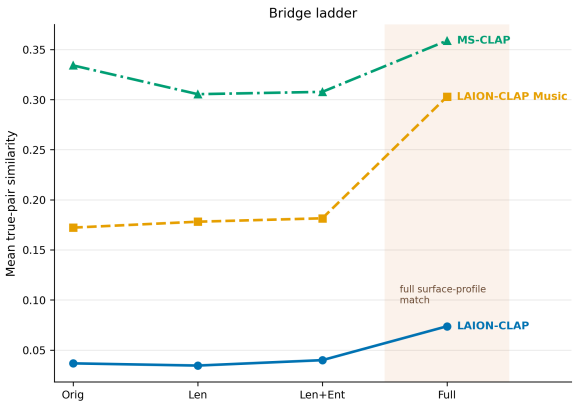

Figure 5: Hachimi-profile bridge ladder. The largest gain appears at the full surface-profile match.

diagnostics show the same qualitative pattern and are reported in Appendix D.

4.2 Semantic Controls and Surface-Form Bridge Probes

Semantic controls test whether the HACHIMI elevation can be explained by simple changes in meaning alone. Table 3 compares PARA, a meaning-preserving paraphrase of ORIG, with ZHNOISE, a Chinese-script nonsense control that preserves rough textual structure while replacing content words with random non-semantic Chinese sound words.

The paraphrase control remains close to ORIG across models, indicating that synonym-level rewriting does not by itself reproduce the HACHIMI elevation. In contrast, ZHNOISE is substantially lower, especially for MS-CLAP. This pattern matters for interpretation: not every non-semantic text produces elevated CLAP similarity. The HACHIMI effect therefore depends on a more specific combination of semantic content and modality-specific cues rather than on semantic absence alone.

Before constructing bridge probes, we use the exploratory regression described in Section 3.5 to identify which surface-form features are most stably associated with the HACHIMI–ORIG gap. This analysis motivates the length–entropy–Latin-profile ladder.

Using the bridge-probe construction defined in Section 3.4, we test whether progressively matching HACHIMI's surface profile reproduces the raw-score elevation. Within this ladder, length and entropy matching produce limited changes, while the largest gain appears at the Latin-ratio/full-profile step under hachimi audio.

Figure 5 summarizes this pattern as a ladder plot, highlighting that the largest increase occurs when the full Hachimi-profile surface-form match is introduced.

Model	BRIDGE-FULL	Zh cap.	En cap.	En margin
LAION-CLAP	0.1469	0.0494	0.1260	+0.0166
LAION-CLAP Music	0.2496	0.1536	0.1620	+0.0378
MS-CLAP	0.3711	0.2522	0.4376	+0.0230

Table 4: Auxiliary caption controls vs. BRIDGE-FULL on hachimi audio ($N = 91$). Values are raw true-pair means; En margin is true–false margin.

Model	ρ	p	Ridge $\hat{\beta}$	OLS Adj. R^2
LAION-CLAP	0.28	0.008	$0.04^{\dagger}$	0.096
LAION-CLAP Music	0.30	0.003	1.12	0.175
MS-CLAP	0.36	< 0.001	0.29	0.134

Table 5: Association between Hachimi Latin-character ratio and the Hachimi–Orig gap across 91 songs.

4.3 Caption-Style Controls

Auxiliary caption controls address a potential text-domain confound: CLAP-style models are trained primarily for audio-text or audio-caption matching, whereas song lyrics are not always literal descriptions of audible content. Therefore, we test whether semantic text in an audio-description style produces stronger pair-specific matching than lyric-like or surface-form bridge conditions. This concern is consistent with common audio-caption training resources such as AudioCaps and Clotho, where text descriptions are written to describe audible events, scenes, and sound sources rather than song lyrics (Kim et al., 2019; Drossos et al., 2020).

We generate Chinese and English audio-description captions from hachimi audio and compare them with BRIDGE-FULL on the same 91-song hachimi-audio setting (Appendix E).

Table 4 shows that Chinese captions remain close to no true–false separation, while English captions improve pair-specific separation and broader ranking diagnostics, suggesting better alignment with CLAP's intended audio-description behavior. The large gap between Chinese and English captions is notable because both are caption-style semantic descriptions; one plausible contributor is tokenizer mismatch, where Chinese captions are fragmented into less favorable token profiles than English descriptions. In MS-CLAP, English captions exceed BRIDGE-FULL in raw true-pair similarity while also producing a positive true–false margin, illustrating a different diagnostic signature from surface-form score elevation.

This comparison separates caption-style semantic matching from surface-form-driven score elevation: BRIDGE-FULL raises raw true-pair similarity, whereas English captions show stronger pair-specific separation and ranking quality. Figure 6 visualizes this raw-score–reliability trade-off.

4.4 Embedding-Space Diagnostics

Text-space analysis. We use tokenizer-sensitive text features as operational diagnostics and find that BRIDGE-FULL shifts toward a Latin-script profile while true-pair and false-pair similarity rise together. We therefore treat the Hachimi-profile bridge as associated with text-space shifts, not as proof of a tokenizer-side mechanism.

Exploratory song-level regression further identifies Latin-character ratio as the most stable positive correlate of the HACHIMI–ORIG gap across models (Table 5).

The association is modest but consistent across the tested models, so we treat Latin-character ratio as a diagnostic cue rather than as a standalone causal factor. The low adjusted R^2 values also indicate that surface form explains only part of the song-level variation.

Full per-condition numeric results are reported in Appendix E.

Audio-space analysis (exploratory diagnostic). Audio-side diagnostics show that the direction and magnitude of gain depend on audio source and model context: in several settings the sign flips between hachimi and original audio for the same text-side shift. This indicates model/audio-dependent modulation and suggests that a purely text-side account is incomplete; the full audio-space figure is reported in Appendix E.

5 Discussion

5.1 What the Pattern Means

We interpret the main empirical pattern of higher HACHIMI than ORIG raw similarity as surface-form-driven similarity elevation: a raw score increase associated with the observed hachimi condition that does not automatically imply stronger pair-specific matching. In our 91-song Chinese setting, score-reliability diagnostics (false-pair mean, margin, AUC, and T2A MRR) do not improve in parallel, making the pattern more consistent with broad embedding-space score elevation than with selective matching improvement.

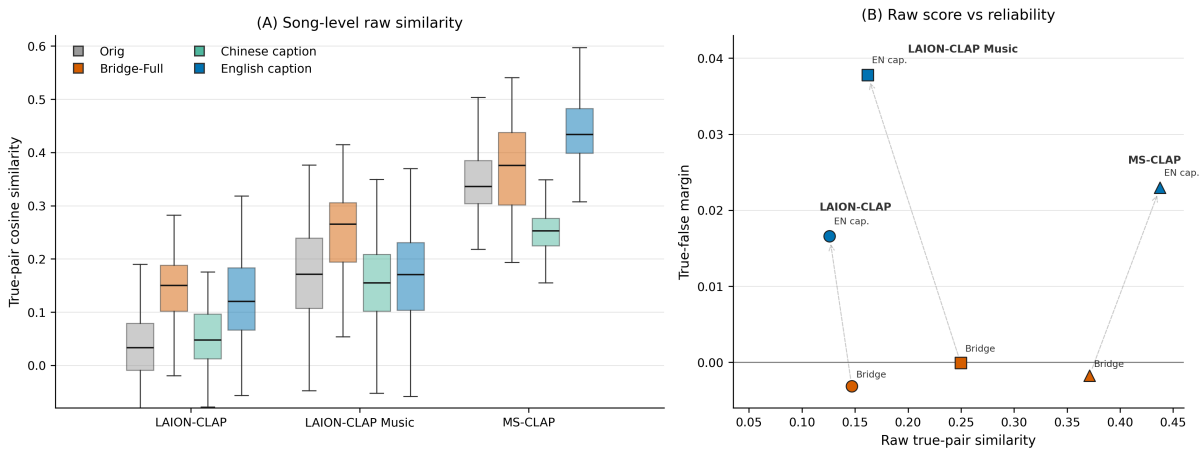

Figure 6: Caption controls on hachimi audio ($N = 91$). Left: song-level true-pair similarity distributions for ORIG, BRIDGE-FULL, Chinese captions, and English captions across the three tested models. Right: raw true-pair similarity versus true–false margin for BRIDGE-FULL and English captions, showing that surface-form bridge conditions can raise raw scores while caption-style semantics yield stronger pair-specific reliability.

5.2 Surface Form, Tokenization, and Model Heterogeneity

The bridge and embedding-space analyses suggest that surface-form factors, particularly script and tokenization-sensitive properties, contribute to the elevation. Yet Latinization-like transformations also lift similarity to non-matching audio, arguing against a simple tokenization-only explanation. The semantic controls and auxiliary caption control reinforce this limited-control interpretation: PARA remains close to ORIG, ZHNOISE does not reproduce the HACHIMI pattern, and Figure 6 shows that English captions improve pair-specific retrieval while BRIDGE-FULL mainly raises raw similarity.

The observed pattern is also model- and audio-dependent. If the phenomenon were only a text-side tokenization artifact, the same text shift would be expected to behave more uniformly across audio conditions. Instead, direction and magnitude vary by audio source and model, suggesting that text-side shifts are filtered through model-specific embedding behavior. Geometry-oriented diagnostics motivate this interpretation but do not establish a complete causal mechanism.

5.3 Implications

The main implication is that raw CLAP similarity should not be treated as direct evidence of semantic audio-text matching. Surface-form properties may shift the score distribution even when pair-specific matching does not become more reliable. Thus, CLAP scores remain useful, but only when interpreted together with controls that distinguish semantic matching from surface-form-driven score elevation.

6 Conclusion

In the main hachimi-audio setting, the core empirical result is that **Hachimi yields higher raw CLAP true-pair similarity than Orig** across the three *tested* models (LAION-CLAP, LAION-CLAP Music, MS-CLAP). Para is used as a semantic control and does not serve as a co-equal primary target with ORIG/HACHIMI. The practical takeaway is that elevated CLAP similarity should be interpreted with score-reliability diagnostics rather than raw cosine alone.

After adding the auxiliary caption control, the same caution remains but becomes sharper: English caption-style text can improve pair-specific retrieval behavior, especially in MS-CLAP, while BRIDGE-FULL mainly produces raw-score elevation. Surface-form factors appear to contribute, but the cross-model and cross-audio heterogeneity indicates that the mechanism is not a single-cause process.

This leaves a broader question beyond the hachimi setting: when text provides little stable meaning, what else can alignment models rely on? Future work can examine this question across CLAP, CLIP, and other models that use text as an alignment signal, especially in settings where semantic content and modality-specific cues can be varied independently. Such studies would clarify how cross-modal alignment behaves when the usual semantic signal in text is weakened or absent.

7 Limitations

This study is based on a 91-song Chinese hachimi dataset and three CLAP-style models. The findings should therefore be interpreted as evidence from a specific lyric-like and weak-semantic setting, rather than as a universal claim about all audio-text models, languages, or genres. In particular, hachimi text is an operational probe: it weakens stable lexical meaning while preserving lyric form, repetition, and syllabic structure. Similarly, paraphrases, bridge probes, and caption controls are used as diagnostic conditions rather than as definitive semantic gold standards. Caption controls are especially operational because audio descriptions can vary across annotators and should not be treated as unique ground truth.

The proposed explanation is also diagnostic rather than fully causal. Our results suggest that surface-form properties, including script and tokenizer-sensitive text profiles, can raise raw CLAP similarity without producing stronger pair-specific matching. However, the bridge conditions and embedding-space analyses do not isolate a single mechanism. The observed effect also varies across models and audio sources, indicating that text-side shifts interact with model-specific embedding behavior.

For this reason, the paper should be read as a caution about interpreting raw CLAP cosine similarity as semantic matching evidence on its own. The reliability diagnostics used here bound what can be inferred from higher raw scores, but they do not replace a full lyric-retrieval benchmark or a complete account of semantic audio-text matching.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Morris Alper and Hadar Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. In *Advances in Neural Information Processing Systems*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta. Association for Computational Linguistics.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. AudioCLIP: Extending CLIP to image, text and audio. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. 2025. Is CLIP ideal? No. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Irene Martín-Morató and Annamaria Mesaros. 2021. What is the ground truth? reliability of multi-annotator data for audio tagging. In *29th European Signal Processing Conference (EUSIPCO)*, pages 76–80. IEEE.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *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)*, pages 3118–3135, Online. Association for Computational Linguistics.

Ramaneswaran Selvakumar, Sonal Kumar, Hemant Kumar Giri, Nishit Anand, Ashish Seth, Sreyan Ghosh, and Dinesh Manocha. 2025. Do audio-language models understand linguistic variations? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 899–913, Albuquerque, New Mexico. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salamon. 2023a. Audio-text models do not yet leverage natural language. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning robust audio representations from CLIP. In *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023b. CLaMP: Contrastive language-music pretraining for cross-modal symbolic music information retrieval. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, pages 157–165.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023c. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

A Dataset Construction and Cleaning

Source. The dataset consists of 91 Chinese popular songs drawn from the *hachimi* meme (romanised *ha ji mi*) community on the Chinese internet. Hachimi is a viral internet meme in which creators re-sing well-known songs by replacing the original Chinese lyrics with phonetically improvised syllables that follow the melody's contour but shed lexical meaning. We collected paired data, including original lyrics, hachimi transcriptions, and audio, for songs that had achieved notable spread within this meme community.

Filtering. Songs were retained only when all three components were simultaneously available: (1) a complete set of original Chinese lyrics, (2) a hachimi audio recording of sufficient quality for manual transcription, and (3) an original studio recording (used in audio-source robustness checks; Appendix E). Songs for which any component was missing or of insufficient quality were excluded, yielding the final 91-song collection.

Original lyrics (Orig) text cleaning. Raw lyrics obtained from public lyric databases often contain residual metadata: standalone artist or producer names prepended to the lyric body, parenthetical annotations, and platform-specific text (e.g., advertising copy, caption attributions). We applied a two-stage cleaning pipeline: (1) rule-based normalization of whitespace and punctuation; (2) an LLM-based auditor that examines the first ≈150 characters and returns a JSON decision (`keep` or `remove_prefix`) with the exact prefix span to strip, using a strict instruction set that forbids touching lyric content, ad-libs, or vocal interjections. Appendix B gives an abbreviated prompt summary for brevity; the full prompt is included in the supplementary repository.

Audio pairing. Each retained song is represented by two audio recordings: a *hachimi audio* recording (the primary source used in all main experiments) and an *original audio* recording (used only in robustness checks; Appendix E).

Metadata and copyright. The released research artifacts identify songs by internal song identifiers and retain derived condition metadata needed to reproduce the analyses. The anonymized supplementary repository is available at `https://anonymous.4open.science/r/hachimi-alignment-9F7E`. Because the corpus involves copyrighted songs, raw audio recordings and raw lyric text are not redistributed. The released artifacts are limited to internal song identifiers, derived text features, condition labels, similarity matrices, aggregate scores, figures, and reproduction scripts.

B Text Condition Generation Details

Orig (original lyrics). Orig is produced by the two-stage cleaning pipeline described in Appendix A. Algorithm 2 gives the pseudocode.

Algorithm 2 Orig Cleaning Pipeline

Require: Raw lyrics string ℓ from public lyric database
  $\ell \leftarrow \textsc{NormalizeSpacing}(\ell)$ ▷ strip/normalize whitespace
  $w \leftarrow \ell[:150]$ ▷ prefix window for LLM inspection
  $r \leftarrow \textsc{LLMAudit}(w)$ ▷ returns `{decision, text}`
  if $r.decision =$ `remove_prefix` **then**
    $\ell \leftarrow \ell[len(r.text):]$
    $\ell \leftarrow \textsc{LStripPunct}(\ell)$ ▷ strip leading `-:|`
  end if
  return ℓ

System prompt (LLM auditor; abbreviated for brevity):

> *You are a strict metadata-cleaning auditor for lyrics. When a standalone artist or creator name appears at the very beginning and can be separated cleanly from following lyrics, remove just that name. Do not remove vocal intros, ad-libs, interjections, spoken lines, or any lyric content.*

The user prompt supplies the first ≈150 characters of the raw lyric and requests a JSON response with fields `decision` (`keep` or `remove_prefix`) and, when removing, the exact prefix `text` to strip. Decisions are validated for prompt-leakage and off-label values before being applied.

Hachimi (hachimi text). Hachimi is produced by manual transcription. An author-side annotator listens to each hachimi audio recording and writes down the syllables sung, preserving line breaks and repetition structure from the original song. No automatic speech recognition is used; all 91 songs are transcribed by hand. The hachimi meme's distinctive phonetic style means that syllables follow the melodic contour but carry no stable Chinese lexical meaning. The resulting text is syllabically rich but

semantically weak, which is the defining property exploited in our experiments.

Para (paraphrase). Para is generated by prompting DeepSeek (`deepseek-v4-flash`) to produce a light synonym paraphrase of each song's Orig lyrics. Algorithm 3 gives the pseudocode; the key prompt constraints are summarized below for brevity.

Algorithm 3 Para Paraphrase Generation

Require: Cleaned lyrics ℓ (Orig), retries R=3
  $msgs \leftarrow$ BUILDPARAPROMPT(ℓ)
  for attempt $= 1 \ldots R$ **do**
    $y \leftarrow$ DEEPSEEKV4FLASH($msgs$, T=0.3)
    $y \leftarrow$ PARSEGENERATEDPARA(y) ▷ extract from JSON
    if $y \neq$ null **then break**
    end if
  end for
  return y

Prompt (Para generation; abbreviated constraints):

- Lightly paraphrase, preserving sentence order, clause count, line count, rhythm, and information density.
- May substitute synonyms or reorder phrases; may not introduce new characters, places, brands, or metadata.
- Keep proper nouns, character names, core imagery, and key actions unchanged.
- Output only the paraphrased text (no explanation), returned as JSON field `generated_c8`.

Auxiliary caption control (`C_OPENAI_CAPTION`). The auxiliary caption control is generated from audio using an OpenAI audio-capable model pipeline. For each song, three 30-second clips (front/middle/back) are extracted from hachimi audio; each clip is captioned with a constrained prompt that requests only audible musical attributes (style, texture, instrumentation, vocal presence, energy) and forbids guessing singer/title/lyrics metadata or precise technical values. This is an abbreviated description for brevity; the full prompt is included in the supplementary repository. The resulting clip captions are not merged by another generative summarization step. Instead, a deterministic selection rule chooses a representative clip caption, preferring repeated captions and otherwise using token-overlap consensus, caption length, and clip order as tie-breakers. The selected caption is inserted as `C_OPENAI_CAPTION` for retrieval comparison against ORIG/HACHIMI/BRIDGE-FULL. Chinese and English caption-control artifacts are included in the anonymized supplementary repository as derived captions and retrieval outputs.

Bridge probes. The main text gives the compressed bridge-probe construction. This appendix provides implementation details. All bridge probes are built deterministically: each per-song random number generator is seeded by SHA-256($song_id \parallel stage \parallel 42$).

ZhNoise (random-character structure clone). Replaces every non-whitespace run in Orig with a run of equal length sampled uniformly from a pool of informal Chinese sound words (*ha*, *ji*, *mi*, *man*, *bo*, etc.), preserving all whitespace positions. The per-song seed is SHA-256($song_id \parallel 42$).

BRIDGE-LEN, BRIDGE-ENT, BRIDGE-FULL (sequential statistical-matching pipeline). Starting from Orig, each stage adjusts one surface feature to match Hachimi's value. Algorithm 4 gives the pseudocode.

Algorithm 4 Bridge Condition Pipeline (BRIDGE-LEN → BRIDGE-ENT → BRIDGE-FULL)

Require: Orig text ℓ_0; Hachimi features: L^* (char count), H^* (token entropy), R^* (Latin char ratio)
  Stage 1 – length matching (BRIDGE-LEN)
  $\ell_{11} \leftarrow$ COMPRESSTOCHARBUDGET(ℓ_0, L^*)
  Stage 2 – entropy matching (BRIDGE-ENT)
  $\ell_{12} \leftarrow$ LOWERENTROPY(ℓ_{11}, H^*)
  Stage 3 – Latin-ratio matching (BRIDGE-FULL)
  if $R^* \geq 0.8$ and Hachimi has Latin tokens **then**
    $\ell_{13} \leftarrow$ LEXICALIZE(ℓ_{12}, Hachimi)
  else
    $\ell_{13} \leftarrow$ LATERALLYLATINIZE(ℓ_{12}, R^*)
  end if
  return ℓ_{11} (BRIDGE-LEN), ℓ_{12} (BRIDGE-ENT), ℓ_{13} (BRIDGE-FULL)

BRIDGE-GLOBAL (global-profile anchor). Identical to Stage 3 of the pipeline, but the Latin token vocabulary is drawn from the aggregated token pool of all 91 Hachimi texts rather than the per-song Hachimi text. This allows assessment of whether uplift is more associated with per-song Latin surface form or with a global CLAP bias toward Latin characters.

C Condition Quality Validation

Hachimi (weak-semantic). Hachimi text is defined as a construction-based weak-semantic condition. It is manually transcribed from Hachimi vocal recordings, where the original meaning-bearing Chinese lyric characters are replaced, character by character or phrase by phrase, with phonetic syllables that do not form stable lexical content. This construction preserves melody-following syllabic timing, repetition, line structure, and song-shaped phrasing, but disrupts the stable word-level semantics of the original lyrics. Thus, "weak-semantic" is used operationally: the condition weakens conventional lexical meaning while retaining vocal-form and lyric-form cues, rather than serving as random noise or as a semantic gold label.

Para (paraphrase quality). Paraphrase quality is validated by checking semantic similarity between Orig and Para using automated similarity metrics and author-side spot checks. We treat Para as an operational semantic control rather than a perfect semantic gold label.

Caption control quality. Caption-style controls are audited for format and content risk (identity guessing, metadata leakage, hallucinated technical measurements, and low cross-segment consensus). Because audio descriptions can vary across annotators and reference labels in audio tasks are not always uniquely defined (Martín-Morató and Mesaros, 2021), we treat `C_OPENAI_CAPTION` as an operational semantic audio-description control rather than a ground-truth caption label.

Token statistics. Key feature statistics across conditions confirm that Hachimi differs systematically from Orig in token count, entropy, and Latin character ratio, and that Para remains close to Orig on these dimensions.

D Full Metric Definitions

All main-text diagnostics default to the text-to-audio (T2A) direction. Audio-to-text (A2T) metrics are supplementary and defined below for completeness.

Let $s(t_i, a_j)$ denote the cosine similarity between the ℓ_2-normalized text embedding t_i and audio embedding a_j for song i, audio j, over $N = 91$ songs.

$$\begin{aligned}
\text{true score}_i &= s(t_i, a_i) \\
\text{false mean}_i &= \frac{1}{N-1} \sum_{j \neq i} s(t_i, a_j) \\
\text{margin}_i &= \text{true score}_i - \text{false mean}_i \\
\text{AUC} &= \hat{P}[s(t_i, a_i) > s(t_i, a_j),\ j \neq i] \\
\text{T2A MRR} &= \frac{1}{N} \sum_i \frac{1}{\text{rank}(a_i \mid t_i)} \\
\text{A2T MRR} &= \frac{1}{N} \sum_i \frac{1}{\text{rank}(t_i \mid a_i)}
\end{aligned}$$

where $\text{rank}(a_i \mid t_i)$ is the rank of the true audio a_i among all N audios when scored by $s(t_i, \cdot)$, and $\text{rank}(t_i \mid a_i)$ is defined analogously for the audio-to-text direction. Z-score normalization subtracts the per-model mean and divides by the standard deviation computed over the full $N \times N$ similarity matrix.

A2T results are consistent with T2A results throughout and are omitted from main-text tables for brevity.

E Full Experimental Results and Robustness Checks

This appendix summarizes the robustness checks referenced in the main text, including original-audio vs. hachimi-audio comparisons, bridge and Latinization condition results, auxiliary caption control comparisons, semantic-control results (PARA/ZHNOISE), and relative-ranking diagnostics. Complete machine-readable outputs are retained in the project artifacts.

Original audio robustness. Under original (non-hachimi) recordings, LAION-CLAP retains a significant Hachimi–Orig gap ($d = 0.22$, $p = 0.036$); LAION-CLAP Music shows a positive but non-significant difference ($d = 0.13$, $p = 0.233$); MS-CLAP shows no meaningful effect ($d = -0.03$, $p = 0.782$). The cross-model consistency seen in Table 2 is therefore specific to the hachimi-audio condition.

Auxiliary caption control. Caption-style semantic text (audio-description format) is compared with ORIG/HACHIMI/BRIDGE-FULL under the same hachimi-audio setup. Chinese captions remain below BRIDGE-FULL in raw true-pair similarity across all three models. English captions are stronger: they remain below BRIDGE-FULL in the

Model	Cond.	True mean	False mean	Margin	AUC	T2A MRR
LAION-CLAP	ORIG	0.037	0.039	−0.003	0.499	0.048
	HACHIMI	0.064	0.062	+0.002	0.511	0.072
LAION-CLAP Music	ORIG	0.181	0.169	+0.013	0.537	0.050
	HACHIMI	0.227	0.216	+0.011	0.528	0.060
MS-CLAP	ORIG	0.339	0.336	+0.003	0.536	0.055
	HACHIMI	0.354	0.353	+0.002	0.521	0.060

Table 6: Score-reliability diagnostics on hachimi audio ($N = 91$; T2A direction). Hachimi raises raw true-pair means, but reliability metrics do not consistently improve.

two LAION models, but exceed BRIDGE-FULL in MS-CLAP and improve retrieval-style reliability across models.

Audio-space diagnostic. The exploratory audio-space analysis examines whether the same text-side shifts behave differently across audio sources and model contexts. Figure 7 shows the full diagnostic.

Semantic controls (PARA/ZHNOISE). Para (paraphrase) and ZhNoise (Chinese-script non-semantic) are summarized as semantic controls. In tested settings, these controls do not reproduce the Hachimi pattern. They narrow the interpretation but do not by themselves establish which surface features are necessary or sufficient.

Relative-ranking diagnostics. Three-way win-count comparisons consistently show Hachimi ranking first in raw-score comparisons, while the AUC/MRR evidence in Table 6 shows that this raw-score advantage does not translate into consistently stronger pair-specific retrieval.

F Regression and Collinearity Diagnostics

This appendix reports full OLS and ridge regression coefficients, Spearman correlations, and VIF values for the exploratory song-level regression described in Section 3.5.

The eight predictors are: token count, token entropy, trigram repetition rate, line repetition rate, pinyin type-token ratio, average syllables per line, Chinese character ratio, and Latin character ratio (all measured on Hachimi text). Songs are split into a dev set (used for hypothesis generation) and a held-out test set (used for bridge condition evaluation). Predictors are standardized before fitting.

Latin character ratio (`c1_latin_char_ratio`) is the only predictor with a consistent positive direction across all three models under Spearman correlation and ridge regression. Token entropy and token count have larger raw OLS coefficients in MS-CLAP (0.55 and −0.53 respectively) but reverse sign across models and show VIF values above 5, indicating collinearity rather than stable effects.

G Reproducibility Details

All reported CLAP scores use frozen pretrained encoders with no parameter updates. The three evaluated models are: **LAION-CLAP**, using the default non-fusion LAION-CLAP package configuration; **LAION-CLAP Music**, using the non-fusion `HTSAT-base` music configuration with checkpoint `laion/larger_clap_music`; and **MS-CLAP**, using the released 2023 configuration from the `microsoft/msclap` model family. Model-specific tokenizers are the tokenizers loaded by the corresponding released packages/checkpoints.

The software environment is Python-based. Core dependencies are `torch` 2.0 or later, `numpy` 1.24 or later, `scipy` 1.10 or later, `librosa` 0.10 or later, `soundfile` 0.12 or later, `laion-clap` 1.0 or later, `msclap` 1.0 or later, `sentence-transformers` 2.2 or later, `transformers` 4.58.0 or later, `accelerate` 1.0 or later, `matplotlib` 3.7 or later, and `seaborn` 0.12 or later. Experiments are deterministic at the condition-generation level: scripts use base seed 42, with per-song random generators derived from `SHA-256(song_id || stage || 42)` for bridge probes and analogous `SHA-256(song_id || 42)` seeds for random-character controls. Embedding extraction uses frozen encoders; small device-dependent floating-point differences may occur across CPU/GPU backends but do not affect the reported qualitative comparisons.

The anonymized supplementary repository includes scripts for the primary ORIG/HACHIMI/PARA comparison, bridge-probe construction, semantic-control evaluation, caption-control generation, retrieval evaluation,

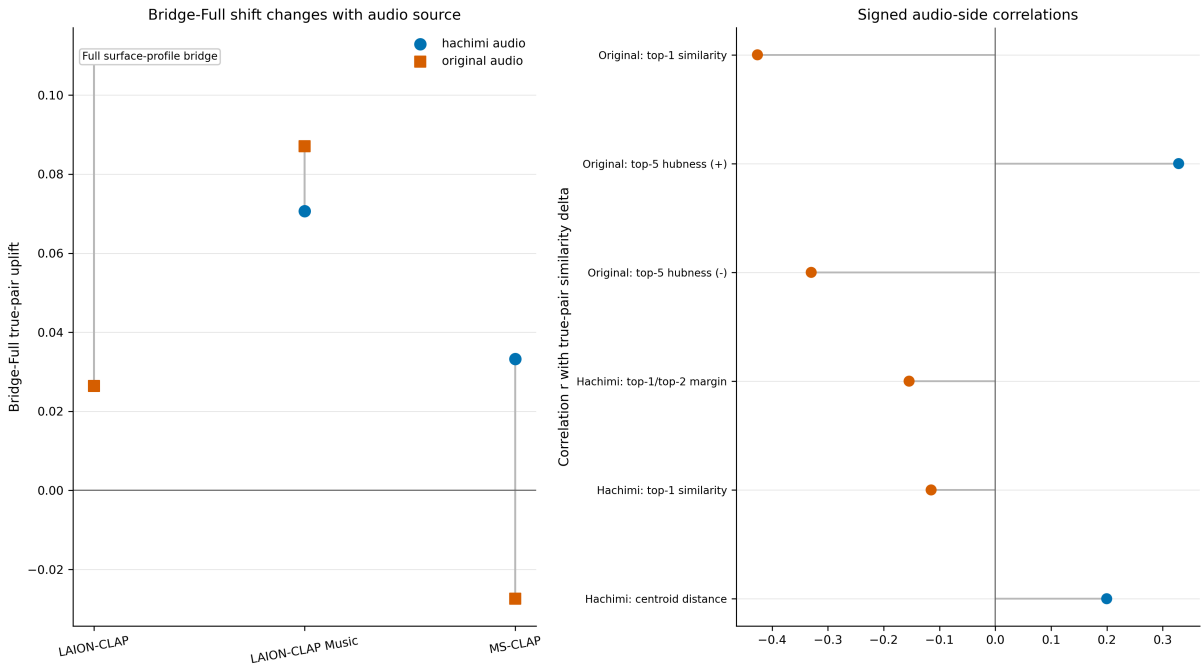

Figure 7: Audio-space diagnostics for bridge effects across audio source and model context.

and figure rendering. Caption-control artifacts are included as derived captions and retrieval outputs; raw copyrighted audio and raw lyric text are not redistributed.