WATCHING, NOT LISTENING: BENCHMARKING OMNIMODAL VOCAL PERFORMANCE COMMENTARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vocal coaching turns a sung performance into selective, actionable feedback, now within reach of omnimodal large language models (omni models) that watch and listen at once. Omni models have grown into unified audio-visual backbones, while AI for music has moved from understanding to singing assessment. However, vocal commentary needs audio-visual evidence, verifiable expert judgement and selectivity at once, and no benchmark measures all three. We formalize *Omnimodal Vocal Performance Commentary* (OmniVPC), where a model comments on a sung clip from its video and audio as an expert coach would. We release *OmniVPC-Bench*, 1,970 items over 127 coached clips in nine genres, with three tracks under four matched media conditions and each coach's own words as the reference. Across 31 omni models, none reaches an OVS of 55: models gain from the picture rather than the sound, and the coach's transcript in place of the media adds about 40 points. Our software and data are publicly available.[1]

1 INTRODUCTION

A vocal coach who has just heard a student sing answers with a few remarks: which technique faltered, how a phrase was shaped, and what to try next (National Association of Teachers of Singing, 2017; ABRSM, 2021). Such feedback is how singers improve, yet it is scarce. Omnimodal large language models (omni models) (Jiang et al., 2025) now take video and audio at once and answer in language, so a model that comments on a performance like a coach is within reach (Figure 1).

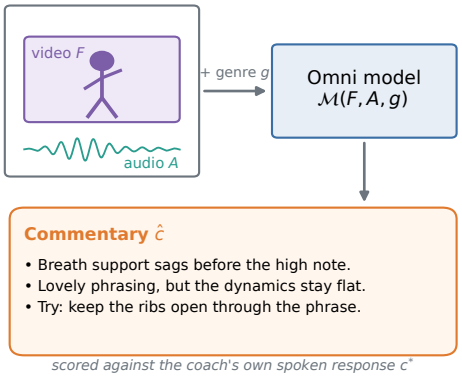


Figure 1: An omni model watches and listens to a singer and comments like a coach.

Along this path, omni models have grown into unified audio-visual backbones, from Qwen-Omni (Xu et al., 2025a;b) to Gemini (Comanici et al., 2025; Google DeepMind, 2026g), with benchmarks moving from joint understanding (Li et al., 2025c; Hong et al., 2026; Zhou et al., 2026b) to whether an answer needs both streams (Gong et al., 2024; Liu et al., 2026a). Music AI has moved from generation and understanding (Copet et al., 2023; Ghosh et al., 2025a) to singing assessment: quality scores (Tang et al., 2024; Yao et al., 2025), critiques (Li et al., 2025a; Zhuang et al., 2026) and pedagogical feedback (Zhang et al., 2025; Bang et al., 2026).

However, a coach's commentary asks for all of this at once. It demands **(i) Audio-visual evidence**: posture and breathing are seen, while intonation and resonance are heard; **(ii) Verifiable expert judgement**: every remark is true or false of this performance, spans technique and artistry, and ends in advice the singer can act on; and **(iii) Selectivity**: a coach addresses 4.3 of fourteen dimensions per clip, so what goes unsaid is part of the answer. As Figure 2A shows, existing lines cover at most two: omni benchmarks ask what is seen and heard, not how well it is sung, lacking (ii) and (iii); singing assessment mostly listens without watching and rarely asks what coaches raise, lacking (i) and (iii); and its judgements may lack auditory support (Zang et al., 2025; Liu et al., 2026b).

[1] Anonymous GitHub link: `https://anonymous.4open.science/r/OmniVPC-Bench`

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

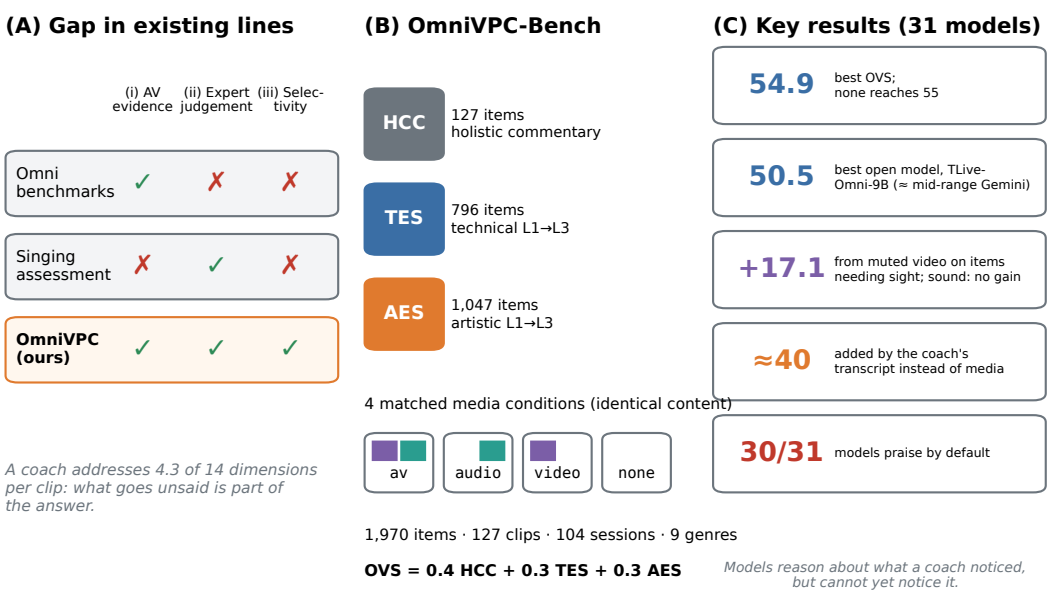


Figure 2: Overview of OmniVPC. **(A)** Existing lines each miss (i), (ii) or (iii). **(B)** OmniVPC-Bench: three tracks under four matched media conditions. **(C)** Key results over 31 omni models.

To close this gap, we formalize *Omnimodal Vocal Performance Commentary* (**OmniVPC**): a model must give the commentary an expert coach would on a sung clip from its video and audio, and is correct insofar as it agrees with that coach's own spoken response, making (i)–(iii) measurable.

We release **OmniVPC-Bench** from real vocal classrooms and master classes: 1,970 items over 127 clips from 104 sessions in nine genres, with no item written by an evaluated model. Three tracks score holistic commentary (HCC) and technical (TES) and artistic (AES) cascades of the coach's verdict, observation and advice, aggregated into OVS; four matched media conditions (`av`, `audio`, muted `video`, `none`) isolate what each stream contributes (Figure 2B).

Across 31 omni models, the best OVS is 54.9, and the strongest open model, TLive-Omni-9B (50.5), matches mid-range Gemini. Probes locate the gap: muted video lifts items needing sight by up to 17.1 points while sound alone does not help; the coach's transcript in place of the media adds about 40 points, whereas swapping the genre line changes 22–30% of verdicts; and 30 of 31 models praise by default. Omni models can reason about what a coach noticed, but cannot yet notice it (Figure 2C).

2 Related Work

Our benchmark joins two lines of work, which Appendix D surveys in full.

Omnimodal large language models. Unified perception (Jiang et al., 2025) is pursued by open dense backbones (Xu et al., 2025a; Team et al., 2026a; Li et al., 2025b; Liu et al., 2025; Fu et al., 2025; Tong et al., 2025; Hu et al., 2026b; OpenBMB, 2025; Cui et al., 2026; Team, 2026a), mixture-of-experts designs (AI et al., 2025; Xu et al., 2025b; NVIDIA et al., 2026; Li et al., 2025d) and the closed Gemini family (Comanici et al., 2025; Google DeepMind, 2025; 2026a;b;d;c;e;f;g), and benchmarked on joint understanding (Li et al., 2025c; Hong et al., 2026; Zhou et al., 2026b) and whether an answer needs both streams (Gong et al., 2024; Liu et al., 2026a), not on performance quality.

AI for music and singing assessment. Music models advanced from generation (Copet et al., 2023) to audio-language understanding (Ghosh et al., 2025a;b; Weck et al., 2024; Li et al., 2022), and singing assessment from quality scores (Tang et al., 2024; Yao et al., 2025) to critiques (Li et al., 2025a; Zhuang et al., 2026) and pedagogical feedback (Zhang et al., 2025; Bang et al., 2026). OmniVPC-Bench joins both lines, scoring audio-visual commentary against coaches.

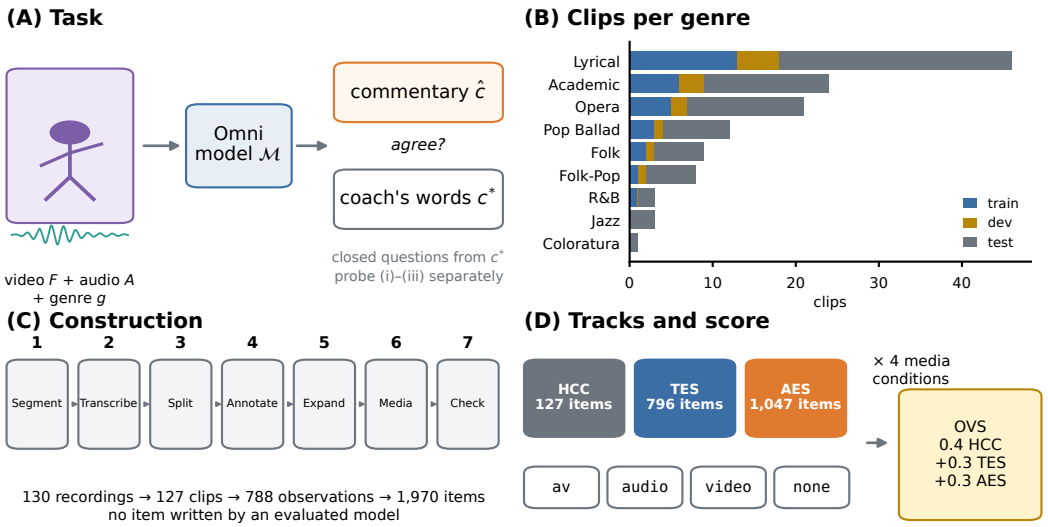


Figure 3: **The OmniVPC task and benchmark.** **(A)** A model comments on a sung clip from its video, audio and genre as its coach would. **(B)** 127 clips from 104 sessions in nine genres. **(C)** Seven steps build 1,970 items. **(D)** Three tracks under four media conditions yield OVS.

3 The OmniVPC Task and Benchmark

3.1 Task Definition

Omnimodal Vocal Performance Commentary (**OmniVPC**) is the task of commenting on a sung performance from its video and its audio at once: given a clip $v = (F, A)$ with visual stream F, synchronised waveform A and a genre tag $g \in \mathcal{G}$ ($|\mathcal{G}| = 9$), a model $\mathcal{M} : (F, A, g) \to \hat{c}$ must produce the commentary an expert vocal coach would give (Figure 3A). The reference is the coach's own spoken response, transcribed into c^*, and a system is correct insofar as its assertions agree with the coach's. The properties of §1 become requirements: *(i)* the evidence is split across F and A; *(ii)* every remark is true or false of the performance, and the commentary must cover technique and artistry and end in advice the singer can act on; and *(iii)* which aspects a commentary leaves out is part of the answer. Closed questions derived from c^* probe each requirement separately.

3.2 Benchmark and Statistics

Source and scale. OmniVPC-Bench is drawn from recordings in which a vocal coach responds to a performance just sung, from Chinese vocal-pedagogy classrooms (69 clips) and international master classes and voice lessons (58). It holds 1,970 items over 127 clips from 104 sessions and 6.68 h of singing in nine genres, split by session into 493 / 184 / 1,293 items over 31 / 13 / 83 clips for train / dev / test (Figure 3B, Table A2).

Dimensions. Every clip is annotated on fourteen dimensions, grouped as in vocal adjudication rubrics (National Association of Teachers of Singing, 2017; ABRSM, 2021): eight technical ones as **Tone** (resonance, vibrato, high register, register transition), **Breath**, **Diction** (articulation) and **Pitch & Rhythm** (intonation, rhythm), and six artistic ones as **Shape** (phrasing, dynamics), **Style & Interpretation** and **Expression & Stage**. A coach addresses 4.3 of them per clip, so "not addressed" is the most frequent label (69.0% of cells; Table A3).

Tracks. HCC (127 items, one per clip) asks for the commentary $\hat{c}$ itself (Figure 3D, Table A1). **TES** (796) and **AES** (1,047) are three-level cascades over the technical and artistic dimensions: L1 asks for the coach's verdict on a dimension, L2 checks one specific observation the coach made,

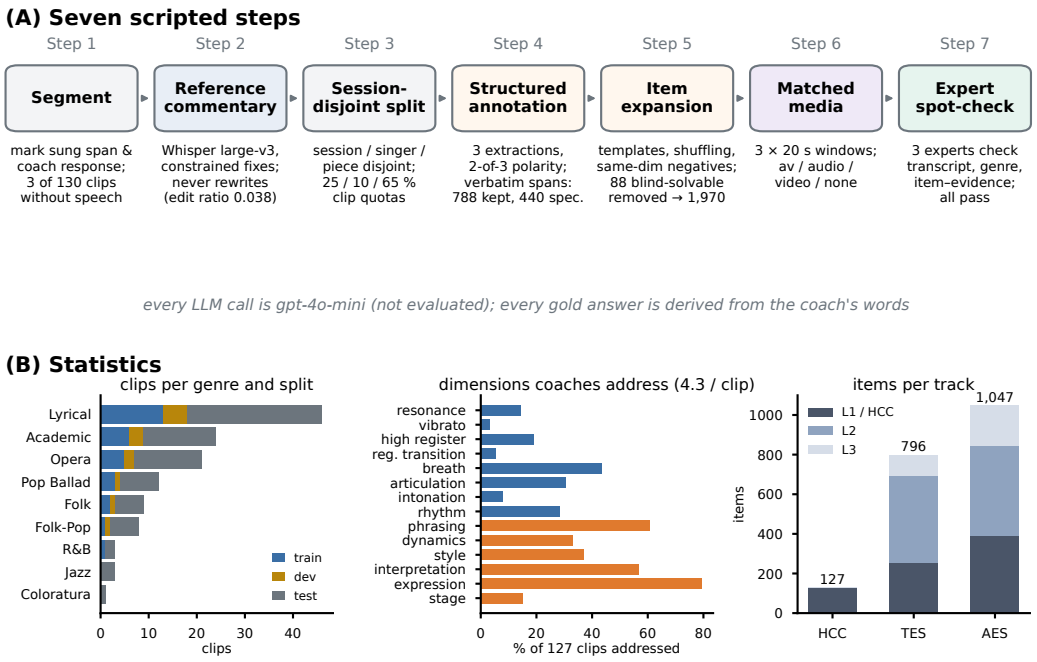


Figure 4: **Construction and statistics of OmniVPC-Bench. (A)** Seven scripted steps, from sourcing to expert spot-check, derive every gold answer from the coach's words. **(B)** Clips per genre and split, dimensions coaches address, and items per track.

and L3 asks for the coach's advice (TES, its own group **Advice**) or for recognition of the coach's description (AES). Each clip ships as `av`, `audio`, muted `video` and `none` with *identical* content.

3.3 Benchmark Construction

Seven scripted, seeded steps build the benchmark (Figure 4A, Appendix A.2); every LLM call in them is made by `gpt-4o-mini`, which is not evaluated, so no system under test authored any data.

Step 1: segmentation. An annotator marks each recording's sung span and coach response and tags the genre; the three of 130 clips without coach speech carry no item. **Step 2: reference commentary.** `faster-whisper large-v3` (Radford et al., 2023) transcribes the response under a vocal-pedagogy prompt, and constrained calls fix recognition errors and keep only the coach's utterances, never rewriting, summarising or translating; the raw transcript is kept beside c^* (median edit ratio 0.038). **Step 3: split.** Components over shared session, singer and piece are assigned toward per-genre quotas of 25 / 10 / 65% of clips, so no recording, singer or piece crosses splits.

Step 4: structured annotation. Each c^* is distilled into

$$\mathcal{A} = \big(\boldsymbol{p}^{\text{tech}},\ \boldsymbol{p}^{\text{art}},\ \mathcal{K},\ \mathcal{S},\ \tau\big), \qquad \boldsymbol{p}^{\text{tech}} \in \Sigma^8,\ \ \boldsymbol{p}^{\text{art}} \in \Sigma^6,\ \ \Sigma = \{+, 0, -, \varnothing\}, \tag{1}$$

where $\varnothing$ marks a dimension the coach never addressed and is itself a gold answer, $\mathcal{K}$ holds observations $(d, p, \tilde{e}, z)$ of a dimension, a polarity, an English rewording and a span $z \subset c^*$, $\mathcal{S}$ the coach's suggestions and τ an optional temporal hint. Of three extraction runs under a *no-new-judgement* prompt, a polarity needs two, and an observation survives only if z is verbatim in c^* (788 kept); the 440 *specific* to the performance, each labelled by the modality it needs, become questions.

Step 5: item expansion. Fixed templates unfold $\mathcal{A}$ into items under four guards: per-item option shuffling; same-dimension negatives, a yes/no negative being an opposite-polarity observation on the *same* dimension from another same-genre clip of the split; class capping in every L1 cell; and removal of the 88 choice items a media-free `gpt-4o-mini` prober solves under all three option orders; after re-capping, $1{,}970$ items remain. **Step 6: matched media.** Each clip is cut to three 20 s windows centred in the thirds of its sung span (the whole span if at most 60 s): `av` is this cut, `audio` its soundtrack, `video` its silent frames, and `none` withholds both, so conditions differ only

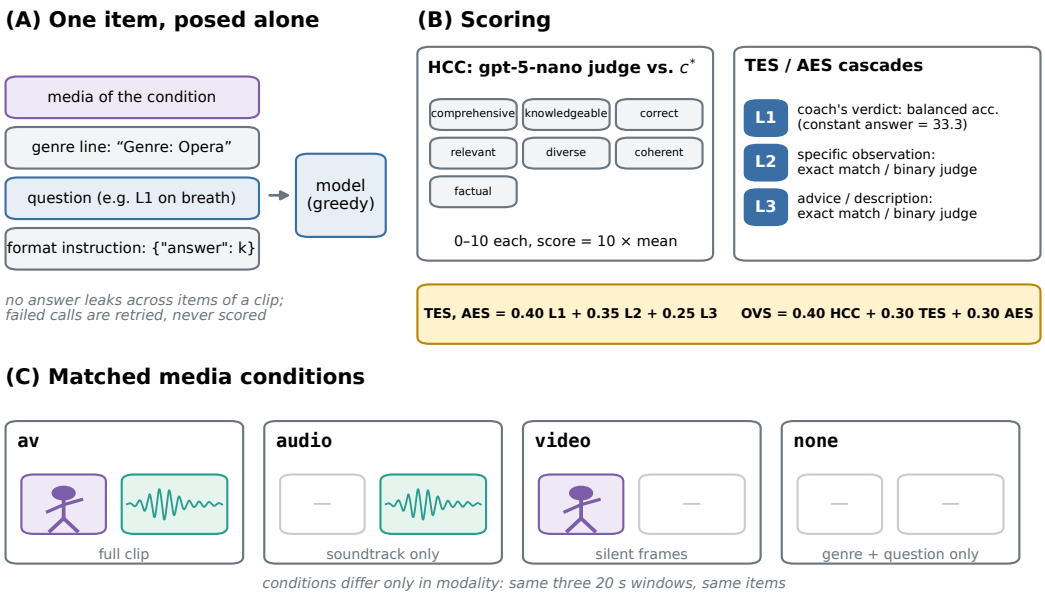


Figure 5: **Evaluation design of OmniVPC-Bench.** **(A)** Each item is posed alone with its media, genre line, question and format instruction. **(B)** HCC is judged on seven rubrics, L1 by balanced accuracy and L2/L3 by exact match or binary judges. **(C)** Matched conditions isolate each stream.

in modality. **Step 7: expert spot-check.** As no step authors free text, three experts from different backgrounds independently spot-check a stratified sample of clips against the coach's audio for transcript fidelity, genre label and item–evidence agreement; every sampled clip passes.

3.4 Evaluation Metrics

Protocol. Each item is posed alone and decoded greedily, with its condition's media, a genre line, the question and a format instruction (Figure 5A, Appendix A.3), so no answer leaks across the items of a clip. Choice replies are read as `{"answer": k}` (by `gpt-4o-mini` for models without a JSON mode), a reply selecting no option is wrong, and a failed call is retried, never scored.

Track scores. Every track scores in $[0, 100]$ (Figure 5B). Closed items are exact-matched, and L1 is balanced accuracy over its three gold classes, so any constant answer scores 33.3. A `gpt-4o-mini` judge credits a short description addressing the reference's aspect without contradicting it, and advice matching the vocal aspect of any coach suggestion. HCC follows the LLM-as-a-judge protocol of Graph-R1 (Luo et al., 2025): `gpt-5-nano` rates $\hat{c}$ against c^* on comprehensiveness, knowledgeability, correctness, relevance, diversity, logical coherence and factuality (0–10, one call each); an item scores ten times their mean, or zero for a refusal or a reply under 50 characters.

Aggregate. Both cascades share one set of level weights, which a dimension group applies to its own items renormalised over its levels, and one score summarises the tracks,

$$\mathrm{TES},\,\mathrm{AES} = 0.40\,\mathrm{L1} + 0.35\,\mathrm{L2} + 0.25\,\mathrm{L3}, \qquad \mathrm{OVS} = 0.40\,\mathrm{HCC} + 0.30\,\mathrm{TES} + 0.30\,\mathrm{AES}. \quad (2)$$

Controls. *Judge sensitivity*: on 80 test clips, a commentary that `gpt-4o-mini` writes from the coach's own annotated points scores 69.0 against that coach's c^* and 61.9 against another same-genre coach's, so the HCC judge rewards agreement with the particular coach; one written from the genre alone scores 67.3, the reference level of a generic commentary. *Linguistic shortcuts*: judge-free items that Gemini-2.5-Flash-Lite answers correctly without media in all three samples are *tagged, not deleted*, so the item set is not fitted to a system under test. *Modality*: a condition's effect is its paired lift over `none` on the same items, and `av−audio` isolates the picture (Figure 5C). Scores carry 95% intervals from 1,000 clip-bootstrap resamples and paired lifts a McNemar test; unless stated otherwise, results are under `av` on all 1,293 test items.

Table 1: **Base models on OmniVPC-Bench** (1,293 test items, `av`). HCC, TES, AES and OVS follow Eq. (2); dimension-group columns (§3.2) weight their own items, so TES/AES are not group means. Per block, the best per column is **bold** and the second underlined. **Price or Params**: OpenRouter USD per 1M input/output tokens (Sept. 2026) or parameter count; decoding in Appendix B.

			Technique (TES)						Artistry (AES)				
Model	**Price or Params**	**HCC**	**Tone**	**Breath**	**Dict.**	**P&R**	**Advice**	**TES**	**Shape**	**S&I**	**E&S**	**AES**	**OVS**
Closed-source API backbones.													
Gemini-2.5-Flash-Lite	\$0.10/\$0.40	69.1	37.2	35.0	24.5	34.7	11.6	27.8	37.9	36.2	34.9	36.5	46.9
Gemini-2.5-Flash	\$0.30/\$2.50	70.5	**40.6**	39.9	33.4	37.7	21.7	33.3	**41.5**	35.0	33.4	37.4	49.4
Gemini-2.5-Pro	\$1.25/\$10.00	71.8	36.9	35.8	36.9	33.2	29.0	33.2	33.7	35.7	39.4	36.5	49.6
Gemini-3-Flash	\$0.50/\$3.00	66.1	36.5	33.2	36.8	37.6	30.4	34.4	34.4	36.0	41.7	38.3	48.2
Gemini-3.1-Flash-Lite	\$0.25/\$1.50	72.2	32.4	35.2	32.8	32.5	33.3	33.4	34.8	39.1	38.7	38.4	50.4
Gemini-3.1-Pro	\$2.00/\$12.00	72.7	36.1	**45.4**	34.4	38.2	29.0	35.4	34.2	35.1	45.7	38.6	51.3
Gemini-3.5-Flash-Lite	\$0.30/\$2.50	70.4	30.9	38.2	25.1	24.9	26.1	29.3	34.0	39.1	45.6	41.0	49.2
Gemini-3.5-Flash	\$1.50/\$9.00	73.7	29.5	32.0	36.2	31.4	29.0	30.8	33.6	39.0	45.9	39.9	50.7
Gemini-3.6-Flash	\$0.75/\$3.75	72.1	38.7	37.0	32.8	38.2	37.7	36.5	38.4	41.6	**50.1**	**44.1**	53.0
Gemini-3.7-Flash	\$0.75/\$3.75	76.1	38.0	39.9	35.6	**39.0**	**46.4**	**39.6**	33.3	43.2	47.4	42.1	**54.9**
Gemini-3.8-Flash	\$0.75/\$3.75	**76.3**	37.2	38.0	**37.4**	37.0	39.1	37.4	35.3	**45.3**	48.4	43.7	54.8
Open-weight omni backbones.													
Gemma-4-E2B	E2B	63.5	29.3	16.5	16.8	19.0	33.3	23.9	22.0	36.2	33.4	30.8	41.8
Gemma-4-E4B	E4B	65.4	24.7	16.7	10.4	13.2	21.7	18.4	27.3	30.8	29.4	28.9	40.4
Qwen2.5-Omni-3B	3B	60.3	34.1	27.7	32.7	22.0	21.7	27.9	33.7	34.9	32.6	34.8	42.9
InteractiveOmni-4B	4B	61.2	33.9	27.9	18.6	20.7	**37.7**	29.2	30.7	34.3	38.7	34.8	43.7
TLive-Omni-4B	4B	68.7	36.8	25.6	33.1	33.4	29.0	30.3	32.0	**41.1**	37.7	37.3	47.8
Qwen2.5-Omni-7B	7B	62.7	38.1	37.0	29.6	36.5	18.8	31.2	32.8	38.7	37.3	36.4	45.4
Baichuan-Omni-1.5	7B	57.2	30.5	27.8	27.5	29.6	14.5	25.1	31.1	29.8	41.2	34.0	40.6
Ola	7B	58.8	28.8	20.5	26.2	23.7	8.7	21.2	27.0	33.3	34.7	32.0	39.5
VITA-1.5	7B	64.3	31.6	27.8	34.8	19.6	10.1	24.0	31.5	33.5	38.0	36.2	43.8
MiniCPM-o-2.6	8B	61.3	24.3	20.6	19.9	19.0	5.8	16.7	25.0	33.0	34.8	31.9	39.1
InteractiveOmni-8B	8B	61.7	39.5	29.5	28.5	26.4	30.4	31.6	31.0	31.9	27.4	29.7	43.1
HyperCLOVAX-SEED-Omni	8B	36.1	27.3	36.2	33.9	24.8	14.5	26.4	33.5	32.1	33.6	33.8	32.5
MiniCPM-o-4.5	9B	65.4	34.5	30.1	30.0	29.3	31.9	32.0	37.1	38.7	40.1	39.7	47.7
TLive-Omni-9B	9B	68.2	40.9	**38.9**	**35.8**	31.8	**37.7**	**36.7**	38.4	39.6	**42.3**	**40.5**	**50.5**
Gemma-4-12B	12B	**72.5**	37.2	36.9	31.0	**39.0**	27.5	33.9	38.4	32.0	36.6	36.5	50.1
Ming-Lite-Omni-1.5	20B-A3B	60.1	31.4	29.7	27.5	25.9	27.5	28.7	30.2	33.0	33.8	32.1	42.3
Qwen3-Omni-Instruct	30B-A3B	67.0	39.0	34.0	32.7	34.9	26.1	33.9	34.5	36.2	40.1	37.4	48.2
Qwen3-Omni-Thinking	30B-A3B	71.1	40.0	33.8	32.1	26.6	34.8	34.6	**39.0**	30.8	40.6	36.7	49.8
Nemotron-3-Nano-Omni	30B-A3B	70.2	**43.7**	33.7	31.1	31.4	33.3	35.2	35.7	34.4	39.4	36.7	49.6
Uni-MoE-2.0-Omni	33B	60.9	39.6	28.0	23.9	25.2	29.0	30.1	36.9	30.6	35.9	34.8	43.8

4 EXPERIMENTS

We answer six RQs: **RQ1:** How well do omni models comment on vocal performances? **RQ2:** Do they listen and watch? **RQ3:** Do their verdicts rest on perception or on priors? **RQ4:** How do they judge a vocal performance? **RQ5:** What do they say? **RQ6:** What varies across models, and why?

4.1 EXPERIMENTAL SETUP

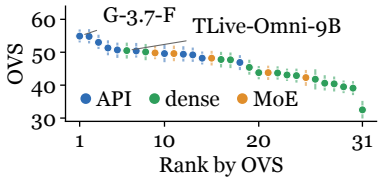
Dataset and Evaluation Metrics. All experiments use the OmniVPC-Bench test split (§3.2, Appendix A): 83 HCC, 523 TES and 687 AES items over 83 clips in nine genres, scored by the metrics of §3.4 and aggregated into OVS by Eq. (2), with every item posed under `av` unless stated otherwise.

Models. We evaluate 31 omni backbones (Table 1, Appendix B): **(i)** eleven closed-source Gemini backbones; **(ii)** fifteen open-weight dense ones (E2B–12B), e.g. Qwen2.5-Omni and TLive-Omni; **(iii)** five open-weight MoE ones (20B–33B), e.g. Qwen3-Omni and Nemotron-3-Nano-Omni. All share one prompt and the same judges, and each receives the clip through its own audio-visual input.

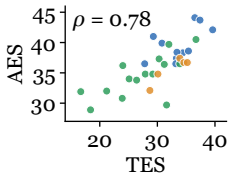
Implementation Details. Items are decoded greedily, one per call, through OpenRouter for API backbones and vLLM or official code on four NVIDIA RTX 6000 Ada GPUs for open-weight ones.

Block		HCC	TES	AES	OVS
API (11)	best	76.3	39.6	44.1	54.9
	mean	71.9	33.7	39.7	50.8
Dense (15)	best	72.5	36.7	40.5	50.5
	mean	61.8	27.2	34.5	43.3
MoE (5)	best	71.1	35.2	37.4	49.8
	mean	65.9	32.5	35.5	46.7

(a) Best and mean per block



(b) OVS with 95% CIs



(c) TES vs. AES

Figure 6: **How well do they comment?** (a) Per block; (b) OVS (95% CI); (c) TES vs. AES.

Probes rerun Gemini backbones changing only the input (Appendix E): *(i)* `audio` and `video` for five backbones; *(ii)* the coach transcript c^* of this or another same-genre clip instead of the media, on the 1,210 TES/AES items; *(iii)* the genre line swapped between classical and pop/jazz; *(iv)* two more identical `av` passes over the 410 L1 items; *(v)* low-effort thinking for Gemini-2.5-Flash-Lite.

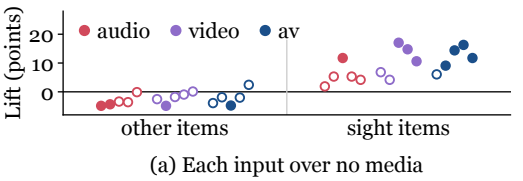
4.2 Main Results (RQ1)

Closed backbones lead, open ones close in. No backbone reaches OVS 55 or TES 40 (Table 1, Figure 6(a)); the best open-weight ones, TLive-Omni-9B (50.5), Gemma-4-12B (50.1) and Qwen3-Omni-Thinking (49.8), match mid-range Gemini-3.1-Flash-Lite (50.4). OVS significantly separates 347 of the 465 backbone pairs (Figure 6(b)), and TES and AES rise together (ρ=0.78; Figure 6(c)).

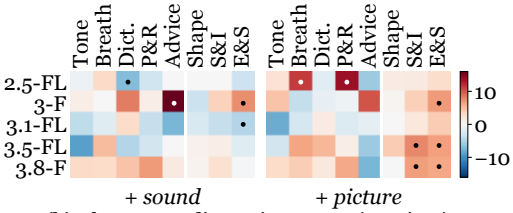
4.3 Do Omni Models Listen and Watch? (RQ2)

Sound alone does not help. Five Gemini backbones answer every item under all four media conditions. On the 946 items whose observation Step 4 did not label as needing sight, the soundtrack changes accuracy over no media by -4.9 to -0.1 points, never a gain (Figure 7(a)).

The picture carries the gain. On the 264 sight items, muted video raises accuracy over no media for all five backbones ($+4.2$ to $+17.1$ points, significant for three). Adding the picture to the sound significantly raises Expression & Stage for three backbones and Style & Interpretation for two (Figure 7(b)). Omni backbones thus judge a vocal performance mainly by what they see rather than by what they actually hear.

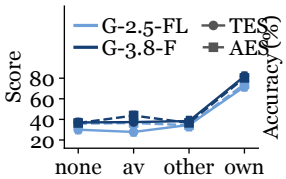


(a) Each input over no media

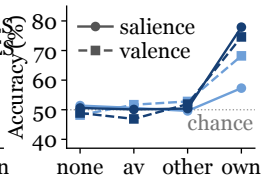


(b) Change per dimension group (Δ points)

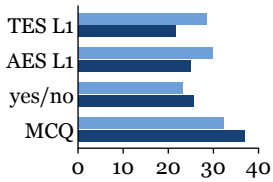
Figure 7: **Listen and watch?** (a) Lift over no media (filled: $p < 0.05$); (b) change per group (dot: CI excludes 0).



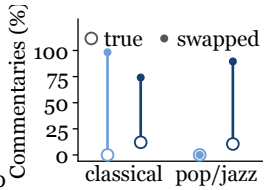
(a) Evidence ladder



(b) L1 salience and valence



(c) Answers changed (%)

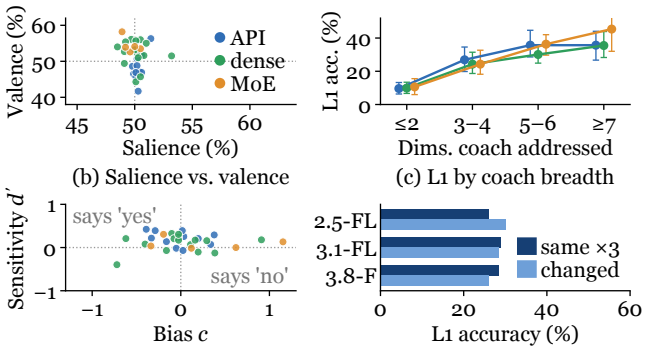


(d) Other-genre terms

Figure 8: **Perception or prior?** (a, b) No media, `av`, or the coach transcript of another same-genre clip (other) or of this clip (own); (c, d) only the genre line swapped (classical vs. pop/jazz).

Backbone	Str.	Wk.	Adq.	N/A
Coach (gold)	29.3	31.0	0.0	39.8
Mean of 31	62.1	19.5	15.0	2.8
Gemini-3.1-Pro	30.0	63.4	5.6	1.0
Gemini-3.7-Flash	92.9	4.9	1.7	0.5
Gemini-3.8-Flash	87.1	9.0	3.4	0.5
MiniCPM-o-4.5	72.9	16.3	8.0	2.7
TLive-Omni-9B	74.6	20.5	3.7	1.2
Gemma-4-12B	63.2	36.6	0.0	0.2
Qwen3-Omni-Think.	57.6	35.4	2.9	4.1
Uni-MoE-2.0	63.2	9.0	24.4	3.4

(a) L1 answer shares (%)



(b) Salience vs. valence (c) L1 by coach breadth (d) Yes/no detection (e) Three identical passes

Figure 9: **How do omni models judge?** (a) L1 answer shares; (b) salience vs. valence (dotted: chance); (c) L1 accuracy (95% CI); (d) yes/no detection; (e) three identical passes.

4.4 Perception or Prior? (RQ3)

Perception is the bottleneck. With this clip's coach transcript in place of the media, TES/AES rise by about 40 points ($p < 0.001$), while another same-genre coach's transcript scores like no media (Figure 8(a, b)): backbones can reason about what a coach noticed, but cannot yet notice it unaided.

The label steers the verdict. Swapping only the genre line changes 21.6–29.8% of L1 answers (identical passes: 7.0–10.1%) and brings pop and jazz technique terms into up to 98.3% of the commentaries on classical clips, up from at most 12.1% under the true genre label (Figure 8(c, d)).

4.5 How Do Omni Models Judge a Performance? (RQ4)

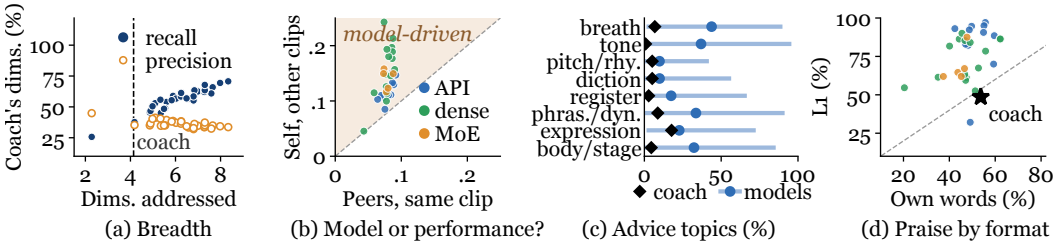
Backbones cannot tell what a coach would address. Salience and valence accuracies stay near chance (48.5–53.2 and 41.7–58.2; Figure 9(b)). "Not addressed", gold for 39.8% of the L1 items, is chosen for 2.8%, while 30 of the 31 backbones prefer "a clear strength" to "a weakness" (Figure 9(a)); L1 accuracy thus climbs from about 10% to 35.5–45.4% with coach breadth (Figure 9(c)).

Specific observations are not verified. On the yes/no items d' spans only -0.40 to 0.42 while the "yes" rate spans 12.2–76.2% (Figure 9(d)), and answers stable over three passes are no more accurate (Figure 9(e)): the verdicts are reproducible priors rather than perceptions of this performance.

4.6 What Do Omni Models Say? (RQ5)

Broad and model-driven. A coach addresses 4.1 of the fourteen dimensions of a clip, the backbones 2.3–8.3 (Figure 10(a)), and a backbone's commentaries on different clips are almost twice as similar as different backbones' commentaries on one clip (TF-IDF 0.141 vs. 0.079; Figure 10(b)).

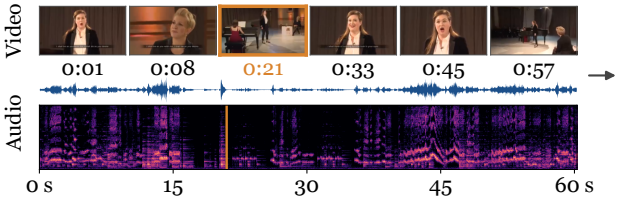
Breath as default advice, praise as default verdict. On average 43.8% of a backbone's advice concerns breath, against 6.8% of the coaches' (Figure 10(c)), and the Advice judge credits it mainly where the coach asked for breath (79.8% vs. 22.0% of the items). Backbones praise about as often as coaches in their own words, yet in L1, 30 of the 31 call a greater share strengths (Figure 10(d)).



(a) Breadth (b) Model or performance? (c) Advice topics (%) (d) Praise by format

Figure 10: **What do omni models say?** Step-4 labels of each commentary. (a) Overlap with the coach's dimensions; (b) TF-IDF similarity; (c) advice topics (line: range); (d) share of praise.

Table 2: **Case study** on test clip `v2_057` (opera, `av`), graded against the coach's words (first row). Blue: agrees with the coach; red: contradicts or is absent; orange: sight cue.



Source	HCC: commentary (excerpt)	HCC	AES L1	Advice (TES)
Coach (c^*, reference)	"… I didn't expect you to look at me but you did this is a great great great thing [stage +] … I want to work breath with you [breath −]"	–	strength	"Work on your breath control."
Gemini-3.7-Flash OVS 54.9; 9/16 items	"… superb dramatic presence … ground your vocal technique firmly in your *appoggio*."	80.0	✓ strength	✓ "… buoyant stream of breath …"
TLive-Omni-9B OVS 50.5; 5/16 items	"… you maintain strong breath support … your articulation is slightly rushed …"	77.1	✗ adequate	✓ "… consistent breath support …"
Gemma-4-12B OVS 50.1; 7/16 items	"… Your facial expressions are engaging … a slightly pinched quality in the higher register."	75.7	✗ weakness	✗ "Maintain a … facial posture …"
Qwen3-Omni-Instruct OVS 48.2; 6/16 items	"Excellent work on the emotional delivery … there's a slight strain …"	71.4	✗ not notable	✗ "… forward-placed tone"
Ola OVS 39.5; 5/16 items	"… your breath control could be better … your vibrato could be more subtle …"	52.9	✗ weakness	✗ "… the vibrato technique …"
MiniCPM-o-2.6 OVS 39.1; 3/16 items	"… a more consistent vibrato … the dynamics are well-balanced … the breath support …"	60.0	✗ adequate	✗ "… controlled vibrato …"

4.7 What Varies Across Models, and Why? (RQ6)

Scale and generation help, thinking less. OVS grows with size among open-weight backbones (ρ=0.47) and along the Gemini Flash line from 48.2 to 54.9, with both Pro backbones below the 3.6- to 3.8-Flash ones (Figure 11(a, b)). Low-effort thinking moves Gemini-2.5-Flash-Lite from 46.9 to 46.4, and Qwen3-Omni-Thinking gains 1.6 over its Instruct version through HCC alone (Figure 11(c)); 103 of the 937 closed items are missed by every one of the 31 backbones (Figure 11(d)).

Case study. In Table 2, only the top backbone, Gemini-3.7-Flash, rates as a strength the stage choice the coach saw and matches the coach on breath in commentary and advice; the others praise the breath support the coach flags, or dwell on points the coach never mentions. It also leads on the clip (HCC 80.0 vs. 52.9–77.1), so OmniVPC-Bench pins each failure to a track and a dimension.

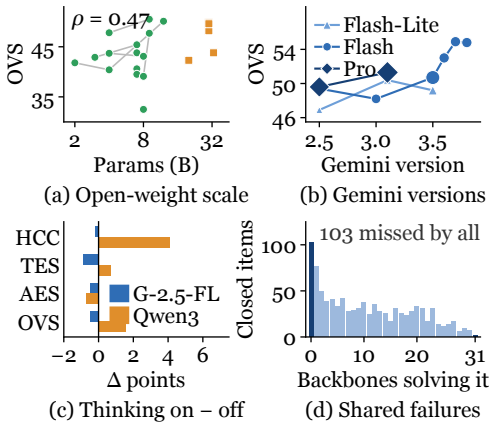


Figure 11: **What varies?** (a) OVS by size; (b) Gemini versions (area: price); (c) thinking; (d) solvers per item.

5 Conclusion

We introduced OmniVPC, commenting on a sung performance from its video and audio as an expert coach would, and OmniVPC-Bench, which grades what a model perceives, judges and advises across fourteen vocal dimensions against the coach's own words. No omni model reaches an OVS of 55, and the gap lies in perception: models gain from the picture but not from the sound, gain about 40 points from the coach's transcript, and otherwise fall back on priors, praising by default, following the genre label and commenting broadly in the model's own style. OmniVPC-Bench thus measures what omni models still lack as vocal coaches: noticing what a coach would single out first.

AI USE STATEMENT

Generative models are part of the method, as `gpt-4o-mini` runs the scripted construction steps and `gpt-4o-mini` and `gpt-5-nano` serve as judges (§3.3, §3.4), and we also used generative AI tools to assist with coding and with drafting and polishing the text. The authors reviewed and verified all AI-assisted code, data and text, and we take full responsibility for the content of this work.

ETHICS STATEMENT

OmniVPC-Bench is built from existing recordings of vocal lessons and publicly posted master classes, collects no new data from human subjects, and takes every gold answer from the coach's own words. It is released for research use only, to evaluate omni models as vocal coaches, and is not intended for assessing or screening individual singers.

REPRODUCIBILITY STATEMENT

The code, data, media and prompts are released at the anonymous link in the abstract, and every construction step with its seeds is given in §3.3 and Appendix A.2, the scoring rules in Appendix A.3, and the configuration of all 31 backbones in Appendix B. The raw reply to every item is stored with its run record (Appendix B.5), so every reported number can be regraded without calling a backbone again, and the protocol of each analysis is detailed in Appendices C and E.

REFERENCES

ABRSM. Singing from 2018: Practical Grades, Requirements and Information (updated April 2021). `https://www.abrsm.org/sites/default/files/2023-09/singing-syllabus-practical-april-2021.pdf`, 2021.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, Furong Xu, GuangMing Yao, Jun Zhou, Jingdong Chen, Jianxin Sun, Jiajia Liu, Jianjiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Libin Wang, Lixiang Ru, Lele Xie, Longhua Tan, Lyuxin Xue, Lan Wang, Mochen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qinglong Zhang, Qiang Xu, Rui Liu, Ruijie Xiong, Sirui Gao, Tinghao Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yunxiao Sun, Yipeng Chen, et al. Ming-omni: A unified multimodal model for perception and generation, 2025. URL `https://arxiv.org/abs/2506.09344`.

Inclusion AI, :, Bowen Ma, Cheng Zou, ChengKun Du, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Chengxiang Fan, Dandan Zheng, Fudong Wang, Furong Xu, Guangming Yao, Haohao Liu, Han Peng, Jun Zhou, Junluan Xia, Jingdong Chen, Jianing Li, Jianxin Sun, Jianjiang Zhu, Jianping Jiang, Jinpeng Ou, Jun Peng, Jin Peng, Kaixiang Ji, Li Tang, Libin Wang, Lixiang Ru, Longhua Tan, Lu Ma, Lan Wang, Mochen Bai, Minghong Cai, Mingxue Yang, Ning Gao, Qingpei Guo, Qinglong Zhang, Qiang Xu, Qin Zhao, Rui Liu, Ruijie Xiong, Ruobing Zheng, Sirui Gao, Shaoxiong Lin, Tao Zhang, Tianqi Li, Tinghao Liu, Tongli Wang, et al. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation, 2026. URL `https://arxiv.org/abs/2510.24821`.

Hayeon Bang, Hounsu Kim, Wonil Kim, and Juhan Nam. Vocalcoachbench: Benchmarking audio-language models on expert feedback for singing, 2026. URL `https://arxiv.org/abs/2609.04241`.

Brandon James Carone, Iran R. Roman, and Pablo Ripollés. Evaluating multimodal large language models on core music perception tasks, 2025. URL `https://arxiv.org/abs/2510.22455`.

Brandon James Carone, Iran R. Roman, and Pablo Ripollés. The muse benchmark: Probing music perception and auditory relational reasoning in audio llms. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16097–16101.

IEEE, May 2026. doi: 10.1109/icassp55912.2026.11462479. URL http://dx.doi.org/10.1109/icassp55912.2026.11462479.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 47704–47720. Curran Associates, Inc., 2023. doi: 10.52202/075280-2066. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/94b472a1842cd7c56dcb125fb2765fbd-Paper-Conference.pdf.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Maryam Dehdashti. Qwen-musicavqa-7b: A multimodal model for music audio-visual qa, 2026. URL https://arxiv.org/abs/2608.11329.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. Omnihalluc-l: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination, 2026. URL https://arxiv.org/abs/2606.03614.

Arthur N. dos Santos and Bruno S. Masiero. A survey on 30+ years of automatic singing assessment and singing information processing, 2026. URL https://arxiv.org/abs/2601.12153.

Milan Liessens Dujardin, Song-Ze Yu, and Kevin Miao. Musp-bench: Advanced multimodal benchmarking of music understanding across score and performance, 2026a. URL https://arxiv.org/abs/2608.28212.

Milan Liessens Dujardin, Song-Ze Yu, Craver Corbyn Thomas-Smith, David M. Chan, and Karina Nguyen. Pitchbench: Measuring pitch hearing in audio-language models, 2026b. URL https://arxiv.org/abs/2605.26176.

Yuxuan Fan, Gyusik Seo, Jing Hao, Jaemin Cho, Mohit Bansal, and Jaehong Yoon. Musebench: Benchmarking intent-level audiovisual arts understanding in mllms, 2026. URL https://arxiv.org/abs/2606.30026.

Kun Fang, Ziyu Wang, and Ichiro Fujinaga. What are you listening to? temporal music grounding for audio-to-text large language models, 2026. URL https://arxiv.org/abs/2608.29480.

Chaoyou Fu, Haojia Lin, Xiong Wang, yifan zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume

38, Main Conference, pp. 75300–75320. Curran Associates, Inc., 2025. doi: 10.52202/085713-2529. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6ce9d51dded7dac82b3b4d3dbb1d73bc-Paper-Conference.pdf.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 41819–41886. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1396. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3babb6b453cb59d87cb58a1219ef914b-Paper-Conference.pdf.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Music flamingo: Scaling music understanding in audio language models, 2025b. URL https://arxiv.org/abs/2511.10289.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, Sang gil Lee, Zhifeng Kong, Jaehyeon Kim, Sungwon Kim, S Sakshi, Ramani Duraiswami, Dinesh Manocha, Andrew Tao, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio-visual flamingo: Open audio-visual intelligence for long and complex videos, 2026a. URL https://arxiv.org/abs/2607.16107.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music, 2026b. URL https://arxiv.org/abs/2604.10905.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, Kevin J. Shih, Siddharth Gururani, Abhinav Shrivastava, Ramani Duraiswami, Dinesh Manocha, Andrew Tao, Bryan Catanzaro, Mohammad Shoeybi, and Wei Ping. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos, 2026. URL https://arxiv.org/abs/2603.14145.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your multimodal llms really understand audio-visual information?, 2024. URL https://arxiv.org/abs/2412.02611.

Google DeepMind. Gemini-3-Flash: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-flash/, 2025. Accessed 2026-09-22.

Google DeepMind. Gemini-3.1-Flash-Lite: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, 2026a. Accessed 2026-09-22.

Google DeepMind. Gemini-3.1-Pro: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026b. Accessed 2026-09-22.

Google DeepMind. Gemini-3.5-Flash: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026c. Accessed 2026-09-22.

Google DeepMind. Gemini-3.5-Flash-Lite: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/, 2026d. Accessed 2026-09-22.

Google DeepMind. Gemini-3.6-Flash: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-6-flash/, 2026e. Accessed 2026-09-22.

Google DeepMind. Gemini-3.7-Flash: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, 2026f. Accessed 2026-09-22.

Google DeepMind. Gemini-3.8-Flash: Official Model Card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026g. Accessed 2026-09-22.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/55dc32df563a35bf406717fb52c118d5-Paper-Conference.pdf.

Guangzheng Hu, Ziyue Jiang, Weixu Qiao, Lixin Zhang, Jianye Kang, Yuru Wu, Rong Bao, Niantong Li, Wei Wang, Ziyi Cheng, Xinfa Zhu, HangRui Hu, Ting He, Bing Zhao, Lin Qu, Hu Wei, and Jin Xu. Omnijudge or omnibias? diagnosing multimodal judges through balanced, decoupled lenses, 2026a. URL https://arxiv.org/abs/2608.24160.

Yibo Hu, Yu Qian, Mao Gu, Yingfan Tao, Yuhao Chen, Yongdong Luo, Zhuoqun Liu, Meiguang Jin, and Junfeng Ma. Tlive-omni: An omni-modal understanding model for e-commerce live streaming, 2026b. URL https://arxiv.org/abs/2608.20958.

Min Jang, Orevaoghene Ahia, Nazif Tamer, Sachin Kumar, Yulia Tsvetkov, and Noah A. Smith. Bass: Benchmarking audio lms for musical structure and semantic reasoning, 2026. URL https://arxiv.org/abs/2602.04085.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. From specific-MLLMs to omni-MLLMs: A survey on MLLMs aligned with multi-modalities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8617–8652, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.453. URL https://aclanthology.org/2025.findings-acl.453/.

Sumit Kumar, Suraj Jaiswal, Parampreet Singh, and Vipul Arora. Automatic detection and analysis of singing mistakes for music pedagogy, 2026. URL https://arxiv.org/abs/2602.06917.

Daeyong Kwon, Qiyu Wu, Shinobu Kuriya, Junghyun Koo, Shuyang Cui, Zhi Zhong, Wei-Hsiang Liao, Hiromi Wakaki, and Yuki Mitsufuji. Mustbench: Benchmarking and advancing temporal grounding in music llms, 2026. URL https://arxiv.org/abs/2605.29300.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, Ying He, Haoxiang Liu, Yuxuan Wang, Qiufeng Wang, Jiafu Tang, Zhenhe Wu, Jiehui Luo, Zhiyu Pan, Weihao Xie, Chenchen Zhang, Zhaohui Wang, Jiayi Tian, Yanghai Wang, Zhe Cao, Minxin Dai, Ke Wang, Runzhe Wen, Yinghao Ma, Yaning Pan, Sungkyun Chang, Termeh Taheri, Haiwen Xia, Christos Plachouras, Emmanouil Benetos, Yizhi Li, Ge Zhang, Jian Yang, Tianhao Peng, Zili Wang, Minghao Liu, Junran Peng, Zhaoxiang Zhang, and Jiaheng Liu. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms, 2026a. URL https://arxiv.org/abs/2510.10689.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19108–19118, June 2022.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Haoxuan Li, Hao Wang, Shijian Wang, Guanting Dong, Jiajie Jin, Yinuo Wang, Yuan Lu, Ji-Rong Wen, Zhicheng Dou, and Zhouchen Lin. Omnigaia: Towards native omni-modal ai agents, 2026b. URL https://arxiv.org/abs/2602.22897.

Xueyan Li, Yuxin Wang, Mengjie Jiang, Qingzi Zhu, Jiang Zhang, Zoey Kim, and Yazhe Niu. Generative multi-modal feedback for singing voice synthesis evaluation, 2025a. URL https://arxiv.org/abs/2512.02523.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, et al. Baichuan-omni-1.5 technical report, 2025b. URL https://arxiv.org/abs/2501.15368.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan WEN, Yanghai Wang, Shihao Li, ZHAO-XIANG ZHANG, Ruibo Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025c. doi: 10.52202/085713-1036. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2c7c4a12a9dcf7dace9896d08154c705-Paper-Datasets_and_Benchmarks_Track.pdf.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, Baotian Hu, and Min Zhang. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data, 2025d. URL https://arxiv.org/abs/2511.12609.

Che Liu, Lichao Ma, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Xuerui Yang, and Fei Tian. Boosting omni-modal language models: Staged post-training with visually debiased evaluation, 2026a. URL https://arxiv.org/abs/2605.12034.

Yu Liu, Jiahui Liu, Zhilin Liu, Cong Cao, Fangfang Yuan, Yuling Yang, Pin Xu, and Yanbing Liu. Music hallucination in audio-language models: A hierarchical formulation and empirical study, 2026b. URL https://arxiv.org/abs/2609.20195.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model, 2025. URL https://arxiv.org/abs/2502.04328.

Haoran Luo, Haihong E, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, and Luu Anh Tuan. Graph-r1: Towards agentic graphrag framework via end-to-end reinforcement learning. *arXiv preprint arXiv:2507.21892*, 2025.

Yishan Lv, Jing Luo, Xinyu Yang, and Zhizheng Wu. Is one score enough? assessing singing quality of songs with temporal score curves, 2026. URL https://arxiv.org/abs/2607.16599.

Menghe Ma, Siqing Wei, Yuecheng Xing, Ziyue Zhu, Zhenghong Lin, Yaheng Wang, Fanhong Meng, Peijun Han, Luu Anh Tuan, and Haoran Luo. Onote: Hypergraph-grounded omnimodal reasoning for computational music science, 2026. URL https://arxiv.org/abs/2604.20719.

National Association of Teachers of Singing. NATS student auditions adjudication form. https://www.nats.org/_Library/NSA_Files/NATS_Adjudication_Form_with_rep_July_11_8_5_by_11.pdf, 2017.

NVIDIA, :, Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, Mike Ranzinger, Greg Heinrich, Guo Chen, Lukas Voegtle, Philipp Fischer, Timo Roman, Karan Sapra, Collin McCarthy, Shaokun Zhang, Fuxiao Liu, Hanrong Ye, Yi Dong, Mingjie Liu, Yifan Peng, Piotr Zelasko, Zhehuai Chen, Nithin Rao Koluguri, Nune Tadevosyan, Lilit Grigoryan, Ehsan Hosseini Asl, Pritam Biswas, Leili Tavabi, Yuanhang Su, Zhiding Yu, Peter Jin, Alexandre Milesi, Netanel Haber, Yao Xu, Sarah Amiraslani, Nabin Mulepati, Eric Tramel, Jaehun Jung, Ximing Lu, Brandon Cui, Jin Xu, Zhiqi Li, Shihao Wang, Yuanguo Kuang, et al. Nemotron 3 nano omni: Efficient and open multimodal intelligence, 2026. URL https://arxiv.org/abs/2604.24954.

OpenBMB. MiniCPM-o-2.6: Official Model Release. https://huggingface.co/openbmb/MiniCPM-o-2_6, 2025. Accessed 2026-09-22.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Neelam Saini and Sourav Ghosh. Listening Like a Judge: A Music-Aware Framework for Automatic Singing Performance Evaluation. In *Interspeech 2026*, pp. 133–137, 2026. doi: 10.21437/Interspeech.2026-912.

Tomáš Sourada, Katia Vendrame, and Jan Hajič jr. Music i care about: Automated multimodal benchmarking of llm music perception skills on (almost) any music, 2026. URL https://arxiv.org/abs/2607.06015.

Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin. Singmos: An extensive open-source singing voice dataset for mos prediction, 2024. URL https://arxiv.org/abs/2406.10911.

Yuxun Tang, Lan Liu, Wenhao Feng, Yiwen Zhao, Jionghao Han, Yifeng Yu, Jiatong Shi, and Qin Jin. Singmos-pro: An comprehensive benchmark for singing quality assessment. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15027–15031. IEEE, May 2026. doi: 10.1109/icassp55912.2026.11460899. URL http://dx.doi.org/10.1109/icassp55912.2026.11460899.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, et al. Gemma 4 technical report, 2026a. URL https://arxiv.org/abs/2607.02770.

Meituan LongCat Team, Bairui Wang, Bayan, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, Chen Chen, Chengxu Yang, Chengzuo Yang, Cong Han, Dandan Peng, Delian Ruan, Detai Xin, Disong Wang, Dongchao Yang, Fanfan Liu, Fengjiao Chen, Fengyu Yang, Gan Dong, Gang Huang, Gang Xu, Guanglu Wan, Guoqiang Tan, Guoqiao Yu, Haibo Qiu, Hao Lu, Hongbo Liu, Hongyu Xiang, Jiaheng Wu, Jian Yang, Jiaxing Liu, Jing Huang, Jingang Wang, Jinrui Ding, Juchao Jiang, Jun Kuang, Jun Wang, Junhui Mei, Ke Ding, Kefeng Zhang, Lei Chen, Liang Shi, Limeng Qiao, Liming Zheng, Lin Ma, et al. Longcat-flash-omni technical report, 2025. URL https://arxiv.org/abs/2511.00279.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, Haozhe Zhao, Hong Liu, Hui Su, Jiaqi Zhang, Jiawei Wang, Jing Li, Kefeng Zhang, Manyuan Zhang, Minhao Jing, Peng Pei, Quan Chen, Taofeng Xue, Tongxin Pan, Xiaotong Li, Xiaoyang Li, Xiaoyu Zhao, Xing Hu, Xinyang Lin, Xunliang Cai, Yan Bai, Yan Feng, Yanjie Li, Yao Qiu, Yerui Sun, Yifan Lu, Ying Luo, Yipeng Mei, Yitian Chen, Yuchen Xie, Yufang Liu, Yufei Chen, Yulei Qian, Yuqi Peng, Zhihang Yu, Zhixiong Han, Changran Wang, Chen Chen, Dian Zheng, Fengjiao Chen, Ge Yang, et al. Longcat-next: Lexicalizing modalities as discrete tokens, 2026b. URL https://arxiv.org/abs/2603.27538.

NAVER Cloud HyperCLOVA X Team. Hyperclova x 8b omni, 2026a. URL https://arxiv.org/abs/2601.01792.

Qwen Team. Qwen3.5-omni technical report, 2026b. URL https://arxiv.org/abs/2604.15804.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue, 2025. URL https://arxiv.org/abs/2510.13747.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. Muchomusic: Evaluating music understanding in multimodal audio-language models. In Blair Kaneshiro, Gautham Mysore, Oriol Nieto, Chris Donahue, Cheng-Zhi Anna Huang, Jin Ha Lee, Brian McFee, and Matthew McCallum (eds.), *Proceedings of the 25th International Society for Music Information Retrieval Conference*, pp. 825–833, San Francisco, California, USA and Online, 2024. ISMIR. doi: 10.5281/zenodo.14877459. URL https://doi.org/10.5281/zenodo.14877459.

Benno Weck, Pablo Puentes, Andrea Poltronieri, Satyajeet Prabhu, and Dmitry Bogdanov. HumMusQA: A human-written music understanding QA benchmark dataset. In Elena V. Epure, Sergio Oramas, SeungHeon Doh, Pedro Ramoneda, Anna Kruspe, and Mohamed Sordo (eds.), *Proceedings of the 4th Workshop on NLP for Music and Audio (NLP4MusA 2026)*, pp. 58–67, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-369-2. doi: 10.18653/v1/2026.nlp4musa-1.9. URL https://aclanthology.org/2026.nlp4musa-1.9/.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. Songbench: A fine-grained multi-aspect benchmark for song quality assessment, 2026. URL https://arxiv.org/abs/2604.25937.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

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, 2025b. URL https://arxiv.org/abs/2509.17765.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. Songeval: A benchmark dataset for song aesthetics evaluation, 2025. URL https://arxiv.org/abs/2505.10793.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, Yuming Lou, Dong Yang, Zhijian Liu, Yukang Chen, Ambrish Dantrey, Ehsan Jahangiri, Sreyan Ghosh, Daguang Xu, Ehsan Hosseini-Asl, Danial Mohseni Taheri, Vidya Murali, Sifei Liu, Yao Lu, Oluwatobi Olabiyi, Yu-Chiang Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu Yin, and Pavlo Molchanov. Omnivinci: Enhancing architecture and data for omni-modal understanding llm, 2025. URL https://arxiv.org/abs/2510.15870.

Wenhao You, Xingjian Diao, Wenjun Huang, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Tingxuan Wu, Ming Cheng, Soroush Vosoughi, and Jiang Gui. Music audio-visual question answering requires specialized multimodal designs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1392–1426, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.69. URL https://aclanthology.org/2026.findings-acl.69/.

Zuyao You, Zhesong Yu, Mingyu Liu, Bilei Zhu, Yuan Wan, and Zuxuan Wu. Gamma: Towards joint global-temporal music understanding in large multimodal models, 2026b. URL https://arxiv.org/abs/2605.00371.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Liumeng Xue, Ziyang Ma, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhu Chen, Jie Fu, Emmanouil Benetos, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, and Yike Guo. ChatMusician: Understanding and generating music intrinsically with LLM. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6252–6271, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.373. URL https://aclanthology.org/2024.findings-acl.373/.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. Are you really listening? boosting perceptual awareness in music-qa benchmarks. In Juhan Nam, Dasaem Jeong, Keunwoo Choi, Li Su, Magdalena Fuentes, Tomoyasu Nakano, Xiao Hu, and Hao-Wen (Herman) Dong (eds.), *Proceedings of the 26th International Society for Music Information Retrieval Conference*, pp. 247–261, Daejeon, South Korea, 2025. ISMIR. doi: 10.5281/zenodo.17811361. URL https://doi.org/10.5281/zenodo.17811361.

Huan Zhang, Vincent K.M. Cheung, Hayato Nishioka, Simon Dixon, and Shinichi Furuya. Llaqo: Towards a query-based coach in expressive music performance assessment. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, April 2025. doi: 10.1109/icassp49660.2025.10890522. URL http://dx.doi.org/10.1109/ICASSP49660.2025.10890522.

Longyin Zhang, Parth Sakhare Mahendra, Chengwei Wei, Ning Zhang, Lim Ming Chong, Sirui He, and Ai Ti Aw. Avtrace: Diagnosing audio-visual temporal reasoning in omni models, 2026a. URL https://arxiv.org/abs/2609.19991.

Sirui Zhang, Tianle Wang, Xinyi Tong, Peiyang Yu, Jishang Chen, Liangke Zhao, Haoxin Zhang, Duo Xu, Xin Jin, Feng Yu, and Songchun Zhu. Madb: A large-scale music aesthetics dataset with professional and multi-dimensional annotations, 2026b. URL https://arxiv.org/abs/2607.06929.

Yizhou Zhang, Wangjin Zhou, Yi Zhao, Wei Tan, Keisuke Imoto, and Zhi Gong. Genre bias or aesthetic perception? identifying and mitigating shortcut learning in music evaluation, 2026c. URL https://arxiv.org/abs/2607.13903.

Jinwen Zhou, Huan Zhang, Weixi Zhai, Jinhua Liang, Aidan O. T. Hogg, and Simon Dixon. Majeppa: Morphing and assessing in a unified piano performance space, 2026a. URL https://arxiv.org/abs/2608.11026.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026b. URL https://arxiv.org/abs/2505.17862.

Ziya Zhou, Shangda Wu, Shenyang Xu, Yutong Zheng, Dafang Liang, Suin Chung, Danbinaerin Han, Junyan Jiang, Yongyi Zang, Ruibin Yuan, Rongxiu Zhong, Shilei Zhang, Junlan Feng, Jinglei Liu, Haotian Zhou, Zijin Li, Dasaem Jeong, Wei Xue, and Yike Guo. Universe: Benchmarking and enhancing lalms on culturally inclusive low-resource music understanding, 2026c. URL https://arxiv.org/abs/2608.17852.

Jiabao Zhuang, Changhao Jiang, Hanchen Wang, Jiahao Chen, Zhixiong Yang, Zhenghao Xiang, Yifei Cao, Jiajun Sun, Hui Li, Ming Zhang, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Musecritic: Learning multi-aspect song rewards through natural-language aesthetic critiques, 2026. URL https://arxiv.org/abs/2608.11755.

APPENDIX

A OMNIVPC-BENCH DETAIL

This appendix expands §3 with the task design (§A.1), the construction pipeline with per-genre and per-dimension statistics (§A.2), and the scoring rule of every item type (§A.3). Throughout, $v = (F, A)$ is a clip with genre $g \in \mathcal{G}$, c^* the coach's transcript, $\mathcal{A}$ its annotation (Eq. 1), $\mathcal{D}$ the fourteen dimensions, and $\kappa \in \mathcal{C}$ one of the four media conditions.

A.1 TASK DESIGN

Every item is a triple $x = (v, q, y^*)$ of a clip, a question and a gold answer, and every model call has the form $\mathcal{M} : (\pi_\kappa(v), g, q) \mapsto \hat{y}$ with $\pi_{\texttt{av}}(v) = (F, A)$, $\pi_{\texttt{audio}}(v) = (\varnothing, A)$, $\pi_{\texttt{video}}(v) = (F, \varnothing)$ and $\pi_{\texttt{none}}(v) = (\varnothing, \varnothing)$, so the four conditions differ in the streams they expose and in nothing else. No question quotes c^*; closed questions carry only English rewordings $\tilde{e}$ of annotated observations. Table A1 lists every item type with an example.

HCC (holistic commentary). $T_{\mathrm{HCC}} : (\pi_\kappa(v), g) \mapsto \hat{c}$ with $y^* = c^*$ asks for 150–200 words in a coach's voice; the reference is the full transcript, so the grader sees every aspect the coach raised.

L1 (verdict). For a dimension $d \in \mathcal{D}$ the model selects the coach's verdict from $\Omega = \{\text{STRENGTH}, \text{ADEQUATE}, \text{WEAKNESS}, \text{NOT ADDRESSED}\}$, with $y^* = \omega(p_d)$ mapping $+, 0, -, \varnothing$ to the four options in order. NOT ADDRESSED is a gold answer in its own right, so L1 asks what a coach would single out, not only how good it sounds; the question names d by a fixed phrase (Table A3).

L2 (observation). L2 grounds the verdict in one observation $o = (d, p, \tilde{e}, z) \in \mathcal{K}$. In the yes/no format $\tilde{e}$ is gold-yes and an opposite-polarity observation on d from another clip is gold-no; the description format (TES) asks which of four rewordings the coach observed; and the short-description format asks for at most 18 words on d, anchored to the region the coach named (opening, middle or closing section, throughout, or one phrase), and is graded against $\tilde{e}$.

L3 (advice and description). TES L3 asks for one imperative sentence, graded against the whole suggestion set $\mathcal{S}$, since each of a coach's several instructions is a valid answer; AES L3 is the four-way description format over artistic observations, recognising the coach's account of an interpretive choice among those of other performances of the same genre.

A.2 BENCHMARK CONSTRUCTION

The seven steps of §3.3 are scripted and seeded (20260828 for the split and the item expansion); every LLM call among them is made by `gpt-4o-mini` in JSON mode.

(1) Sourcing and segmentation. Duplicates are merged and re-encodes of one recording form one session; sung spans are re-encoded to H.264 with AAC at 128 kb/s and responses cut to 16 kHz mono PCM. The 127 clips hold 6.68 h of singing (range 10–545 s) and 6.26 h of coach response (Table A2).

(2) Reference commentary. `faster-whisper large-v3` detects the language and decodes again in it with a vocal-pedagogy initial prompt (e.g., appoggio, passaggio, messa di voce), beam size 5, temperature fallback $\{0, 0.2, 0.4\}$, voice-activity detection; Chinese is normalised to Simplified characters. The repair call (temperature 0) sees each segment with its time stamps and ASR log-probability, fixes only recognition errors, and labels every utterance as coach, singer or other speech (5,256 / 536 / 42 of 5,834); a fixed dictionary of 24 Chinese pedagogy homophones is then applied deterministically. The median c^* has 468 Chinese characters or 457 English words.

(3) Split. Union-find over shared session, singer and repertoire identifiers yields 99 indivisible blocks. Within each genre, blocks are taken largest first (seeded ties) and each goes to the split furthest below its quota. Genres with fewer than three sessions go to test whole, and a final check confirms that no session, singer or piece occurs in two splits. Judge-free items number 354 / 114 / 937 per split.

Table A1: **Item types of OmniVPC-Bench.** *Items*: all splits, test in parentheses, TES/AES for a pair. Examples come from one English-language Lyrical master-class training clip (the yes/no proposition from another Lyrical training clip); questions are abridged and options omitted.

Level	Format	Scoring	Items	Example question ⇒ gold answer
HCC	commentary, 150–200 words	LLM judge against c^*	127 (83)	. . . write the commentary an expert vocal coach would give the singer. ⇒ c^*: *"[. . .] you're not expressing what royally means. [. . .]"*
L1	polarity, 4 options	balanced accuracy	255/389 (162/248)	. . . assessing breath support in this performance. Which statement best matches their assessment? ⇒ *a weakness the coach would flag*
L2	yes/no	exact match	138/260 (108/186)	Regarding breath support: is this an accurate description of what actually happens? "Your breath engagement is stable and should open up more." ⇒ *no*
L2	description, 4 options	exact match	160/– (102/–)	. . . did the coach actually observe about THIS performance? ⇒ *The beginning was too slow and introspective.*
L2	short description, $\leq$ 18 words	judge: same aspect, no contradiction	141/195 (82/122)	. . . describe what happens with rhythm and timing in the opening section. ⇒ *The beginning was too slow and introspective.*
L3	advice, one imperative sentence	judge: aspect of a coach suggestion	102/– (69/–)	Give the single most useful, concrete instruction . . . ⇒ *Deliver the text as a story.* (one of five suggestions)
L3	description, 4 options	exact match	–/203 (–/131)	as for the L2 description ⇒ *You observe the dynamics well, but don't let them dominate meaning.*

Table A2: **Per-genre statistics.** Clips per split (session-disjoint), coaching language (Chinese / English; Pop Ballad has one Italian clip), median sung span, and items, observations ($\mathcal{K}$) and suggestions ($\mathcal{S}$). A session whose clips carry two genre labels is counted in each.

		Clips			Lang.		Span	Items					
Genre	**Sess.**	**Train**	**Dev**	**Test**	**zh**	**en**	**(s)**	**HCC**	**TES**	**AES**	**Test**	$\lvert\mathcal{K}\rvert$	$\lvert\mathcal{S}\rvert$
Lyrical	40	13	5	28	15	31	217	46	312	423	505	305	205
Academic	19	6	3	15	24	0	168	24	188	170	231	147	99
Opera	19	5	2	14	11	10	219	21	124	232	234	153	106
Pop Ballad	9	3	1	8	4	7	35	12	52	59	84	40	30
Folk	9	2	1	6	9	0	178	9	53	93	110	61	36
Folk-Pop	7	1	1	6	5	3	157	8	45	49	84	58	30
R&B	3	1	0	2	1	2	30	3	9	8	15	4	5
Jazz	1	0	0	3	0	3	215	3	11	11	25	15	18
Coloratura	1	0	0	1	0	1	31	1	2	2	5	5	6
Total	**104**	**31**	**13**	**83**	**69**	**57**	**188**	**127**	**796**	**1,047**	**1,293**	**788**	**535**

(4) Structured annotation. The extraction prompt asks only for what the coach expressed, with $\varnothing$ as the stated default ("do not infer a polarity from [. . .] the genre, or from your own opinion of singing"): a polarity per dimension; 3–8 observations, each with an English rewording of at most 18 words, a verbatim span of 6–60 characters and its region; up to three imperative suggestions; and an optional temporal hint. Of three samples at temperature 0.3, polarities need a two-of-three majority, while observations and suggestions are pooled and deduplicated. The annotations hold 788 observations, 535 suggestions over 102 clips and 1,778 polarity cells (288 +, 253 −, 10 neutral, 1,227 $\varnothing$); a second pass (temperature 0) labels each observation as specific (440) or general (348) and by the modality needed to verify it (audio 572, both 184, visual 32; Table A3).

(5) Item expansion. Templates fix the question wording, so items differ only in the clip, the dimension phrase and the observation text (Table A1). A yes/no pair is built only when the clip's polarity on d is $+$ or $-$, with its negative from a clip of opposite polarity on d, so "no" is backed by both commentaries; the 398 yes/no items split 199 / 199. Description distractors are observations on other dimensions from other clips of the same genre and split, so no item text crosses splits. Class capping downsamples $\varnothing$ in every L1 cell (track, dimension) to the mean size of the strength and weakness classes (at least four) and caps each gold class at its runner-up. Re-capping after the text-

Table A3: **Per-dimension statistics.** *Phrase*: how the question names the dimension. *Addr.*: clips whose coach addressed it (% of 127) and their polarity; *Agr.*: mean majority share of the three extraction samples where any sample addresses it. $|\mathcal{K}|$: observations, of which *spec.* are specific. The last row is TES L3, which carries no dimension; its $|\mathcal{K}|$ counts suggestions.

			Addressed clips					Obs.		Items			
Group	**Dimension**	**Phrase**	**Addr.**	+	−	0	**Agr.**	$\lvert\mathcal{K}\rvert$	**spec.**	**L1**	**L2**	**L3**	**Test**
Tone	resonance	*resonance*	18 (14.2)	15	3	0	0.91	25	20	21	44	–	40
	vibrato	*vibrato*	4 (3.1)	1	3	0	0.92	6	4	7	7	–	10
	high register	*the high register*	24 (18.9)	13	11	0	0.91	37	36	35	97	–	78
	register transition	*register transitions*	7 (5.5)	2	5	0	0.86	9	8	10	13	–	11
Breath	breath	*breath support*	55 (43.3)	25	30	0	0.95	59	33	81	80	–	103
Diction	articulation	*articulation and diction*	39 (30.7)	8	29	2	0.94	61	47	44	111	–	109
P&R	intonation	*intonation*	10 (7.9)	3	3	4	0.80	8	6	9	9	–	6
	rhythm	*rhythm and timing*	36 (28.3)	12	24	0	0.91	52	29	48	78	–	97
Shape	phrasing	*musical phrasing*	77 (60.6)	28	49	0	0.91	106	57	104	92	50	157
	dynamics	*dynamic contrast*	42 (33.1)	13	28	1	0.92	50	22	52	31	19	56
S&I	style	*stylistic idiom*	47 (37.0)	37	9	1	0.89	79	32	53	36	27	86
	interpretation	*interpretation of the text*	72 (56.7)	49	22	1	0.92	133	51	88	89	45	145
E&S	expression	*emotional expression*	101 (79.5)	66	35	0	0.97	137	86	72	198	59	221
	stage	*stage presence and body use*	19 (15.0)	16	2	1	0.92	26	9	20	9	3	22
Advice	–	–	102 (80.3)	–	–	–	–	535	–	–	–	102	69

only filter and removing unpaired yes/no items and an undersized temporal-localisation template shrink the pool from 2,130 to 1,976 items; dropping six neutral-gold L1 items leaves 1,970, with L1 gold classes of 72 / 88 / 95 (strength / weakness / not addressed) in TES and 124 / 117 / 148 in AES.

(6) Matched media. The windows are centred at $\frac{1}{6}$, $\frac{1}{2}$ and $\frac{5}{6}$ of the sung span and concatenated into `av.mp4`; the muted `v.mp4` and `a.mp3` copy its video stream and soundtrack, so all three share frames and samples, and a probe confirms that every `v.mp4` is silent and a clip's files agree in length within 0.5 s. The release ships them for all 130 sourced clips with the items and every step's metadata.

A.3 Evaluation Metric Design

Prompt. The system prompt is "You are an expert vocal coach evaluating a singing performance." The user turn holds a head line (under `none`, "A singing performance. No audio or video is provided."), a genre line pairing the label with a short description of its idiom, the question with options numbered from 0, and a format line such as `Reply strictly as JSON {"answer": <option number>}` or "One imperative sentence. Nothing else.", then the media.

Closed items. A reply is mapped to an option $\hat{\imath}$ from its `answer` field or its first valid index (for backbones without a JSON mode, `gpt-4o-mini` reads the chosen option), and $\hat{\imath} = -1$ otherwise. Yes/no and description items score $\mathbb{1}[\hat{\imath} = i^*]$, and L1 is balanced accuracy over its gold classes $\mathcal{Y}$,

$$\mathrm{L1} = \frac{100}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \frac{1}{|X_y|} \sum_{x \in X_y} \mathbb{1}\big[\hat{\imath}(x) = i^*(x)\big], \tag{3}$$

where $\mathcal{Y} = \{\textsc{strength}, \textsc{weakness}, \textsc{not addressed}\}$ and X_y are the items with gold y, so every class weighs equally and any constant answer scores 33.3.

Judged items. Two binary `gpt-4o-mini` judges run at temperature 0: $J_{\mathrm{sd}}(\hat{e}, \tilde{e}) = 1$ iff the response's main focus is the reference's aspect and it does not directly contradict it, and an advice item scores $\max_{s \in \mathcal{S}} J_{\mathrm{adv}}(\hat{s}, s)$, with $J_{\mathrm{adv}} = 1$ iff two suggestions address the same vocal aspect. For HCC, `gpt-5-nano` (low reasoning effort) receives the question, c^* as the golden answer, a 0–10 scoring guide and $\hat{c}$ in one call per rubric $m \in R$ (the seven of §3.4); $\mathrm{HCC}(\hat{c}) = \frac{10}{|R|} \sum_{m \in R} s_m(\hat{c}, c^*)$, or 0 for a refusal or a commentary under 50 characters.

Aggregation and uncertainty. Level weights and OVS follow Eq. 2. Intervals take the 2.5th and 97.5th percentiles of 1,000 resamples of the 83 test clips, and paired lifts over `none` are tested with McNemar's test.

B MODEL DETAIL

This appendix documents how each of the 31 backbones of Table 1 was configured and deployed. We first give the inference protocol that every row shares (§B.1), then the closed-source API backbones (§B.2), the open-weight backbones and their serving stack (§B.3), the audio-visual input path of each open-weight model (§B.4), and the run records kept for every row (§B.5).

B.1 SHARED INFERENCE PROTOCOL

Inputs. Every backbone answers the same 1,293 test items of 83 clips under the `av` condition. A clip is shipped as one file `av.mp4` (H.264, 384 pixels wide at 216–308 pixels high, 25 fps, with an AAC soundtrack at 16 kHz mono) together with its muted video stream `v.mp4` and its soundtrack `a.mp3` (16 kHz mono), which share the same frames and samples. Test clips last 10.1–60.1 s (56.6 s on average, 0.78 MB per `av.mp4`). Each backbone receives the clip through the audio-visual input that its own interface or official code provides, so that picture and sound always come from the same recording and the same time window.

Prompt and decoding. All rows use the system prompt, user prompt and format lines of §A.3, one item per call, and greedy decoding (temperature 0 for API and vLLM rows, `do_sample=False` and `num_beams=1` for official-code rows). The answer budget is fixed per item format: 32 output tokens for multiple-choice and yes/no items, 64 for short descriptions and advice, and 900 for the 150–200-word commentary. Backbones that reason before answering receive 2,048 further tokens on top of each budget, and only the final answer is graded. Official inference code has no JSON mode, so these rows receive 256 tokens for closed items and their choice is read from the `answer` field when the reply is JSON, and otherwise by the `gpt-4o-mini` extractor of §A.3; API and vLLM rows decode closed items in JSON mode. Every row produces text only: speech heads, talkers and vocoders are either not built or not run.

Grading. All rows are graded by one code path with the same judges (`gpt-4o-mini` at temperature 0 for short descriptions, advice and choice extraction, 4 output tokens; `gpt-5-nano` at low reasoning effort for HCC, one call per rubric), accessed through one pinned OpenAI-compatible endpoint. HCC judgements are cached under a hash of the judge model, its settings and the full judge prompt, so an identical judgement is never requested twice. A failed backbone or judge call is retried (up to four attempts with linear back-off, then a fresh session) and is never scored; a row enters Table 1 only when all 1,293 items carry a scored answer.

B.2 CLOSED-SOURCE API BACKBONES

The eleven Gemini backbones are called through the OpenRouter chat-completions endpoint in September 2026 (Table B1). The clip is attached as a base64-encoded `av.mp4` file part after the text of the user turn, so the model receives frames and soundtrack in a single audio-visual stream, as in its native video input. Calls run with 8–16 concurrent workers; each call's charge is taken from the `usage.cost` field of the response and summed per session. Reasoning follows each model's default, except that the backbones whose thoughts would otherwise consume the short answer budgets run at low reasoning effort with 2,048 extra output tokens; their thoughts are excluded from the returned text (`reasoning.exclude`), so only the answer is graded.

B.3 OPEN-WEIGHT BACKBONES

Hardware and serving. All twenty open-weight backbones run locally on one server with four NVIDIA RTX 6000 Ada GPUs (48 GB each, driver 555.42), an Intel Xeon w9-3475X CPU and 503 GB of RAM, one model at a time and always on all four GPUs. Checkpoints are the official Hugging Face releases (Table B2), loaded in bf16 unless the official loader prescribes another type, and copied to a RAM disk before loading. Two routes are used. *(i) vLLM*: the Qwen-Omni models and Nemotron-3-Nano-Omni, whose model cards recommend vLLM, are served through its OpenAI-compatible endpoint (maximum length 32,768 tokens, one video and one audio item per prompt, video decoded at 1 fps up to 60 frames, `use_audio_in_video=true`) and queried by the same client code as the API rows with 64 concurrent requests; prefix caching is off and each prompt is prefilled in one piece (at most 32,768 batched tokens), at 70% GPU-memory utilisation

Table B1: **Closed-source API backbones.** Model identifiers as served by OpenRouter. *Budget*: maximum output tokens for closed items / short descriptions and advice / commentary. Prices are in Table 1.

Backbone	OpenRouter model ID	Reasoning	Budget
Gemini-2.5-Flash-Lite	`google/gemini-2.5-flash-lite`	default	32 / 64 / 900
Gemini-2.5-Flash	`google/gemini-2.5-flash`	low effort	2,080 / 2,112 / 2,948
Gemini-2.5-Pro	`google/gemini-2.5-pro`	low effort	2,080 / 2,112 / 2,948
Gemini-3-Flash	`google/gemini-3-flash-preview`	default	32 / 64 / 900
Gemini-3.1-Flash-Lite	`google/gemini-3.1-flash-lite`	default	32 / 64 / 900
Gemini-3.1-Pro	`google/gemini-3.1-pro-preview`	low effort	2,080 / 2,112 / 2,948
Gemini-3.5-Flash-Lite	`google/gemini-3.5-flash-lite`	default	32 / 64 / 900
Gemini-3.5-Flash	`google/gemini-3.5-flash`	low effort	2,080 / 2,112 / 2,948
Gemini-3.6-Flash	`google/gemini-3.6-flash`	low effort	2,080 / 2,112 / 2,948
Gemini-3.7-Flash	`google/gemini-3.7-flash`	low effort	2,080 / 2,112 / 2,948
Gemini-3.8-Flash	`google/gemini-3.8-flash`	low effort	2,080 / 2,112 / 2,948

(90% for Nemotron). *(ii) Official inference code*: every other model runs through the loader, processor and `generate` call of its model card or repository, wrapped in a thin per-model adapter; one process per GPU (or per GPU pair for models that do not fit one GPU) holds a full copy of the model and answers the items of a quarter (or half) of the clips, sharded by clip so that each clip's media is decoded once. Each model family has its own Python environment pinned to the versions its card or repository requires (PyTorch 2.1.2–2.11.0), with user site-packages disabled so that no other installation can shadow a pinned package; attention runs on FlashAttention-2 or PyTorch SDPA, as the official loader selects.

Table B2: **Open-weight backbones.** *Runtime*: vLLM server, or official inference code on Transformers (TF) given as *HF* (the checkpoint's remote code) or *repo* (the authors' GitHub code base). Weights are bf16, except fp16 for VITA-1.5 as in its official loader. *GPUs*: DPn = n model copies with one GPU each; 2×2 = two copies over two GPUs each; TPn = tensor parallelism over n GPUs. *Frames*: frames per clip at 1 fps, capped at 60 unless the official processor fixes the count. *A–V input*: *joint* = one time-aligned audio-visual stream; *units* = 1-s units of one frame followed by that second of audio; *items* = frames and soundtrack as two inputs of the same user turn, with the audio in consecutive windows of the given length where the encoder requires it.

Model	Checkpoint	Runtime	GPUs	Frames	A–V input
Gemma-4-E2B	google/gemma-4-E2B-it	HF, TF 5.5.3	DP4	≤ 60	items, 30 s
Gemma-4-E4B	google/gemma-4-E4B-it	HF, TF 5.5.3	DP4	≤ 60	items, 30 s
Qwen2.5-Omni-3B	Qwen/Qwen2.5-Omni-3B	vLLM 0.17.1	DP4	≤ 60	joint
InteractiveOmni-4B	sensenova/InteractiveOmni-4B	HF, TF 4.51.3	DP4	≤ 60	items, 30 s
TLive-Omni-4B	TaoLiveAIGC/TLive-Omni-4B	HF, TF 5.2.0	DP4	≤ 60	joint
Qwen2.5-Omni-7B	Qwen/Qwen2.5-Omni-7B	vLLM 0.17.1	DP4	≤ 60	joint
Baichuan-Omni-1.5	baichuan-inc/Baichuan-Omni-1d5	HF, TF 4.45.0	DP4	≤ 60	items, 30 s
Ola	THUdyh/Ola-7b	repo, TF 4.43.4	DP4	≤ 60	joint, 30 s
VITA-1.5	VITA-MLLM/VITA-1.5	repo, TF 4.41.1	DP4	16	items
MiniCPM-o-2.6	openbmb/MiniCPM-o-2_6	HF, TF 4.44.2	DP4	≤ 60	units
InteractiveOmni-8B	sensenova/InteractiveOmni-8B	HF, TF 4.51.3	DP4	≤ 60	items, 30 s
HyperCLOVAX-SEED-Omni	naver-hyperclovax/HyperCLOVAX-SEED-Omni-8B	repo, TF 4.52.4	DP4	≤ 60	items, 30 s
MiniCPM-o-4.5	openbmb/MiniCPM-o-4_5	HF, TF 4.51.0	DP4	≤ 60	units
TLive-Omni-9B	TaoLiveAIGC/TLive-Omni-9B	HF, TF 5.2.0	DP4	≤ 60	joint
Gemma-4-12B	google/gemma-4-12B-it	HF, TF 5.17.0	DP4	≤ 60	items, 30 s
Ming-Lite-Omni-1.5	inclusionAI/Ming-Lite-Omni-1.5	repo, TF 4.52.4	2×2	≤ 60	items
Qwen3-Omni-Instruct	Qwen/Qwen3-Omni-30B-A3B-Instruct	vLLM 0.17.1	TP4	≤ 60	joint
Qwen3-Omni-Thinking	Qwen/Qwen3-Omni-30B-A3B-Thinking	vLLM 0.17.1	TP4	≤ 60	joint
Nemotron-3-Nano-Omni	nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16	vLLM 0.20.0	TP2, DP2	≤ 60	joint
Uni-MoE-2.0-Omni	HIT-TMG/Uni-MoE-2.0-Omni	repo, TF 4.51.3	2×2	32	joint, 30 s

B.4 Audio-Visual Input per Model

We give each model the clip through the richest audio-visual path its official code supports, in three forms.

Joint stream. *Qwen2.5-Omni* and *Qwen3-Omni* receive `av.mp4` with `use_audio_in_video`, which interleaves audio and visual tokens on a shared time axis. *TLive-Omni* follows the card's "video with audio" recipe: its processor decodes the soundtrack from the file itself and interleaves it with the visual tokens of each temporal grid, with the Qwen3-VL video loader at 1 fps and at most 60 frames and thinking off. *Nemotron-3-Nano-Omni* is served by vLLM 0.20.0, the version pinned by its model card, with the card's flags (`--video-pruning-rate 0.5`, `--reasoning-parser nemotron_v3`); the server extracts the soundtrack from `av.mp4` and places its audio context directly after the video of the same turn, and thinking is off. *Uni-MoE-2.0-Omni* follows the "Video Omni Understanding" case of the official cookbook, in which video and audio tokens share its Omni-3D-RoPE time axis; its processor samples 32 frames for a video that comes with audio and cuts the soundtrack into 30-s Whisper windows. *Ola* follows its official `infer.py`: a joint `<speech><image>` placeholder, frames through `process_anyres_video` and the soundtrack through its Whisper-large-v3 and BEATs dual encoder in 30-s chunks.

One-second units. *MiniCPM-o-2.6* and *MiniCPM-o-4.5* run in their official omni mode, which cuts the clip into 1-s units of one frame followed by that second of 16-kHz audio (`get_video_chunk_content` of the 2.6 card; `get_video_frame_audio_segments` of the 4.5 card), with thinking off for 4.5.

Frames and soundtrack in one turn. Models whose released code has no audio-in-video mode receive the frames and the soundtrack as two inputs of the same user message, the video+audio format of their own demos. *Gemma 4* takes `v.mp4` with one frame per second of the clip (the processor would otherwise default to 32) and the soundtrack as consecutive segments of at most 30 s, its stated limit per audio input. *InteractiveOmni* loads frames with its own `load_video` at 1 fps and one 448-pixel tile per frame, InternVL's video convention, which keeps 60 frames within its 40,960-token context, and encodes the soundtrack in 30-s segments; the prompt is assembled exactly as `model.chat()` does, with one audio placeholder per segment. *Baichuan-Omni-1.5* follows the video+audio turn of its official web demo; its processor decodes frames at 1 fps (frame cap 60) and splits the audio into 30-s windows. *VITA-1.5* follows `video_audio_demo.py`, with the demo's 16 frames (`MAX_IMAGE_LENGTH`) sampled over the clip and fp16 weights as in the official loader; the random dither of its filter-bank front end is seeded before each call, so that outputs are reproducible. *HyperCLOVAX-SEED-Omni* is served on the card's OmniServe route (vision encoder, audio encoder and language model), which we run as one process per GPU: frames at 1 fps (at most 60, 378^2 pixels) through its Qwen2.5-VL vision encoder, and the soundtrack through its Qwen2-Audio encoder in 30-s windows and its discrete audio tokenizer, with media placed before the text as in OmniServe's examples, reasoning skipped and decoding restricted to text tokens. *Ming-Lite-Omni-1.5* follows the `v1.5` cookbook of its repository, with 60 uniformly sampled frames, the Whisper audio encoder over the whole soundtrack, the system prompt given as the SDK does, and its layers split over two GPUs.

Common settings. Where a card samples (MiniCPM-o-2.6, Ola) or adds a repetition constraint (Ming-Lite-Omni-1.5), we decode greedily as for every other row, and the benchmark's system prompt replaces built-in persona prompts. The prompt text precedes the media unless the official examples order them otherwise. Qwen3-Omni-Thinking is the only open-weight row that reasons before answering; it receives the 2,048 extra tokens of §B.1, and only the text after its reasoning is graded.

B.5 Run Records

Every row keeps a record for audit. For an API row, the run record stores the command-line settings, the start and end time, the Git revision, and the SHA-256 of the evaluation code and of the item file. For an open-weight row, a generation manifest additionally stores the hashes of the checkpoint's configuration and tokenizer files, the number and size of its weight files, a copy of the adapter, runner or server code with its hash, the environment's package versions and full `pip freeze`, and the GPU inventory. Raw answers are stored per item before grading, so every number of Table 1 can be regraded from them without calling a backbone again. Each item's line in the result file holds the raw reply with its parsed choice and the gold answer, or its judged score, and for commentary items the seven rubric scores of the HCC judge, so every aggregate traces back to the replies and judgements that produced it, item by item and clip by clip.

C EXPERIMENT DETAIL

This appendix details the protocol and the full results of each research question of §4. All scores come from the evaluation code of Table 1 on the 1,293 test items under `av` unless a probe changes the input (Appendix E). Intervals are percentiles of a clip bootstrap (1,000 resamples of the 83 test clips, shared by all backbones and conditions, so differences are paired). Two conditions are compared item by item with an exact McNemar test on the TES/AES items both scored; a lift is the accuracy difference in points, and **bold** marks $p < 0.05$ in every table that reports such a paired test.

C.1 MAIN RESULTS (RQ1)

Protocol. Each backbone answers every item once through the shared prompt of Appendix B and is scored by Eq. (2). Two backbones are separated on a score when the 95% interval of their paired difference excludes zero; the same paired test is applied to every dimension-group column of Table 1.

Table C1: **Leaders of the dimension groups** (Table 1): best backbone with its 95% bootstrap interval, the runner-up, and the mean and range over the 31 backbones.

Group	Leader	Score	95% CI	Runner-up	Score	Mean	Range
Tone	Nemotron-3-Nano	43.7	[37.5, 50.5]	TLive-Omni-9B	40.9	34.8	24.3–43.7
Breath	Gemini-3.1-Pro	45.4	[36.3, 54.1]	Gemini-2.5-Flash	39.9	31.9	16.5–45.4
Diction	Gemini-3.8-Flash	37.4	[33.0, 41.5]	Gemini-2.5-Pro	36.9	29.7	10.4–37.4
Pitch & Rhythm	Gemini-3.7-Flash	39.0	[33.0, 44.4]	Gemma-4-12B	39.0	29.6	13.2–39.0
Advice	Gemini-3.7-Flash	46.4	[34.2, 58.2]	Gemini-3.8-Flash	39.1	26.7	5.8–46.4
Shape	Gemini-2.5-Flash	41.5	[35.1, 48.2]	Qwen3-Omni-Think.	39.0	33.5	22.0–41.5
Style & Interp.	Gemini-3.8-Flash	45.3	[39.1, 50.9]	Gemini-3.7-Flash	43.2	35.8	29.8–45.3
Expr. & Stage	Gemini-3.6-Flash	50.1	[44.0, 56.0]	Gemini-3.8-Flash	48.4	38.7	27.4–50.1

Results. OVS intervals are narrow (± 1.6 to ± 2.8 points): OVS separates 347 of the 465 backbone pairs, and OVS or a dimension group separates 418 (89.9%). The groups add diagnostic information: six backbones lead the eight groups (Table C1), in 60 pairs a group significantly reverses the OVS order, and in 33 pairs each backbone significantly wins a different group. The ranking holds across both coaching languages: OVS on English-coached clips exceeds that on Chinese-coached clips by only 0.2 points on average and is the higher of the two for 18 of the 31 backbones (Figure E1(c)).

C.2 LISTENING AND WATCHING (RQ2)

Protocol. Five Gemini backbones spanning versions 2.5 to 3.8 and both tiers (probe *i*) answer every item under `none` (question and genre line), `audio`, muted `video` and `av`. *Sight items* are the 264 items whose coach observation Step 4 labelled as needing sight, out of the 1,210 TES/AES items.

Table C2: **Paired accuracy lifts** (points) of probe *i* over `none`, and of adding the picture to the sound (`av` − `audio`); **bold**: McNemar $p < 0.05$.

Backbone	**Other items** (946)	**Sight items** (264)			
	`audio` − `none`	`audio` − `none`	`video` − `none`	`av` − `none`	`av` − `audio`
Gemini-2.5-Flash-Lite	**−4.9**	+1.9	+6.8	+6.1	+4.2
Gemini-3-Flash	**−4.3**	+5.3	+4.2	**+9.1**	+3.8
Gemini-3.1-Flash-Lite	−3.4	**+11.7**	**+17.1**	**+14.4**	+2.6
Gemini-3.5-Flash-Lite	−3.6	+5.3	**+14.8**	**+16.3**	**+11.0**
Gemini-3.8-Flash	−0.1	+4.2	**+10.6**	**+11.7**	**+7.6**

Results. The sight items respond to the picture as their label predicts: every backbone gains from muted video on them (+4.2 to +17.1 points), `av` beats `none` by up to +16.3, and adding the picture to the sound raises their accuracy for all five backbones (Table C2). Per group, the picture significantly raises Expression & Stage for three backbones and Style & Interpretation for two (Table E2). The benchmark thus separates what a backbone sees from what it hears, which a single audio-visual score could not reveal; Table E2 lists HCC, TES, AES and OVS under each of the four conditions.

C.3 PERCEPTION OR PRIOR? (RQ3)

Protocol. Probe *ii* replaces the media by the coach transcript c^* of the clip itself (own) or of another same-genre test clip (other; pairing in Appendix E); probe *iii* swaps only the genre line between the classical and pop/jazz families; probe *iv* repeats the identical `av` request twice, setting the noise floor of an answer change. Each probe reuses the system prompt, media files and judges of Table 1.

Items require performance-specific evidence. With the own transcript, the backbones gain 565 and 577 items over `av` and lose 84 and 93, a lift of $+39.8$ and $+40.0$ points ($p < 0.001$), and L1 salience rises to 57.3 and 77.9 (Table E3). Another same-genre coach's transcript, with the coaching language of the genre but not this performance, changes accuracy by only -2.1 (p=0.25) and $+0.2$ points (p=0.97). The items are answerable from what the coach heard and saw in this performance, not from genre knowledge, and the gap to this ceiling then measures perception alone, item by item.

Table C3: **Genre-line swap** (probe *iii*): answers changed (%) per closed format, against the change between identical passes (probe *iv*, L1), and the shifts (→) it causes.

Backbone	Changed by the swap: TES L1	AES L1	yes/no	MCQ	All	Identical passes	L1 "strength" share	Pop/jazz terms, classical clips
Gemini-2.5-Flash-Lite	28.4	29.8	23.1	32.2	28.1	7.0	66.6 → 71.0	0.0 → 98.3
Gemini-3.8-Flash	21.6	25.0	25.5	36.9	27.5	10.1	87.1 → 72.7	12.1 → 74.1

The genre line is a strong cue. The swap changes 27.5–28.1% of all closed answers; on L1 it changes 21.6–29.8%, two to four times the 7.0–10.1% between identical passes (Table C3). HCC registers the label-driven commentary, falling from 76.3 to 73.2 and from 69.1 to 67.5 accordingly.

C.4 HOW DO OMNI MODELS JUDGE A PERFORMANCE? (RQ4)

Protocol. L1 asks for the coach's verdict on one dimension: a clear strength, a weakness, adequate, or not addressed. *Salience* scores addressed versus not addressed as balanced accuracy over the 410 L1 items; *valence* scores strength versus weakness on items the coach addressed and the backbone judged. The paired yes/no items give the sensitivity d' and the bias c of each backbone (Appendix E).

Results. Salience lies at 48.5–53.2 for all 31 backbones (Table C4): none can tell which dimensions a coach would single out. The share of "a clear strength" ranges from 26.8% to 96.3% against 29.3% in the gold answers, so the benchmark exposes each backbone's own bias. Where the coach addressed at most two dimensions, 78.1% of the gold L1 answers are "not addressed" and L1 accuracy is 9.6–10.6% per block; where the coach addressed seven or more, the share is 5.8% and accuracy 35.5–45.4%. The rate of "yes" ranges from 12.2% to 76.2% while d' stays within -0.40 to 0.42: the backbones differ in how often they say "yes", not in what they can detect in the clips.

C.5 WHAT DO OMNI MODELS SAY? (RQ5)

Protocol. The Step-4 annotator labels each backbone commentary with the fourteen dimensions and polarities used for the coach's commentary (Appendix E), drawing a single sample at temperature 0.

Results. Coaches are selective: they address 4.1 dimensions per clip, mostly expression, phrasing and interpretation (Table C5). The backbones address 2.3–8.3, recover 25.7–70.7% of the coach's dimensions at a precision of 32.0–44.9%, over-cover breath, resonance and intonation, and under-cover interpretation, style and rhythm. Each is closer to its own commentaries on other clips than to its peers' on the same clip, by 1.05 to 3.28 times: commentary follows the model and not the clip.

C.6 WHAT VARIES ACROSS MODELS, AND WHY? (RQ6)

Protocol. Spearman ρ relates OVS to parameters (20 open-weight backbones) and to output price (11 Gemini backbones); probe *v* adds low-effort thinking to Gemini-2.5-Flash-Lite, all else equal.

Results. OVS grows with scale (ρ=0.47) and price (ρ=0.57). Thinking redistributes accuracy rather than adding perception: TES L3 rises from 11.6 to 21.7 (Gemini-2.5-Flash-Lite) and from 26.1 to 34.8 (Qwen3-Omni), while TES L1 falls from 31.1 to 22.2 and from 35.4 to 29.6, leaving item

Table C4: **Per-backbone statistics.** Sal./Val.: L1 salience and valence; Str.: L1 "a clear strength"; Yes: "yes" rate; Dims: dimensions per commentary; Rec./Prec.: of the coach's dimensions; Praise: praised share; Self: TF-IDF self-similarity; Breath: advice on breath. In % except d', Dims and Self.

	Judgement (RQ4)					Commentary (RQ5)					
Backbone	**Sal.**	**Val.**	**Str.**	**Yes**	d'	**Dims**	**Rec.**	**Prec.**	**Praise**	**Self**	**Breath**
Gemini-2.5-FL	50.6	51.7	66.6	42.2	0.21	6.2	61.4	38.3	48.2	0.151	18.8
Gemini-2.5-Flash	51.4	56.3	68.5	52.0	−0.02	6.6	56.5	33.3	47.1	0.147	21.7
Gemini-2.5-Pro	49.9	46.5	63.7	49.0	−0.07	5.9	58.9	38.6	59.3	0.106	43.5
Gemini-3-Flash	50.4	47.2	96.3	63.0	0.22	7.0	62.5	34.7	55.7	0.129	24.6
Gemini-3.1-FL	49.8	48.6	91.5	35.4	0.29	7.7	65.9	34.9	42.4	0.130	37.7
Gemini-3.1-Pro	50.2	50.7	30.0	36.7	0.07	6.8	63.4	36.2	49.1	0.111	17.4
Gemini-3.5-FL	50.2	48.8	81.0	61.2	0.39	6.8	55.2	32.0	49.7	0.085	46.4
Gemini-3.5-Flash	50.3	41.7	86.1	56.5	0.14	6.8	58.5	32.8	59.6	0.133	26.1
Gemini-3.6-Flash	50.6	46.9	90.7	65.0	0.42	7.6	60.9	32.4	48.7	0.110	84.1
Gemini-3.7-Flash	50.1	45.7	92.9	48.6	0.26	7.3	62.1	33.5	55.1	0.110	58.0
Gemini-3.8-Flash	50.1	46.9	87.1	49.3	0.39	7.3	63.9	33.8	54.9	0.103	49.3
Gemma-4-E2B	49.1	49.4	29.0	18.0	0.18	5.6	54.0	39.4	44.3	0.177	76.8
Gemma-4-E4B	50.1	44.3	26.8	38.4	0.16	5.1	53.0	41.2	46.8	0.197	71.0
Qwen2.5-Omni-3B	49.7	56.1	46.3	45.6	0.07	5.5	48.9	35.0	47.4	0.148	72.5
InteractiveOmni-4B	49.2	55.4	66.1	76.2	−0.40	6.0	56.7	37.1	45.5	0.189	79.7
TLive-Omni-4B	50.2	52.3	57.6	56.5	−0.07	7.3	63.5	35.1	52.7	0.114	26.1
Qwen2.5-Omni-7B	50.5	56.0	78.8	42.2	−0.10	5.3	55.8	40.5	49.1	0.157	18.8
Baichuan-Omni-1.5	49.1	52.7	56.3	73.3	0.21	4.2	37.7	35.3	62.2	0.115	1.4
Ola	50.0	53.2	32.0	43.5	0.17	4.9	45.5	35.0	20.4	0.243	5.8
VITA-1.5	50.5	45.7	55.1	54.1	0.33	6.2	58.4	36.4	47.1	0.180	20.3
MiniCPM-o-2.6	53.2	51.5	46.3	65.8	0.08	5.3	45.1	34.7	62.2	0.193	7.2
InteractiveOmni-8B	48.5	54.0	37.3	35.0	−0.13	4.9	46.7	36.5	35.1	0.213	50.7
HyperCLOVAX-SEED	50.1	55.7	40.0	51.0	0.24	2.3	25.7	44.9	51.3	0.045	5.8
MiniCPM-o-4.5	50.8	51.6	72.9	53.1	0.20	6.5	56.9	34.3	40.1	0.146	43.5
TLive-Omni-9B	51.0	55.0	74.6	51.0	0.17	7.3	67.8	37.3	55.8	0.100	79.7
Gemma-4-12B	50.3	51.0	63.2	51.0	0.31	5.8	55.4	38.1	46.7	0.126	71.0
Ming-Lite-1.5	48.9	58.2	40.5	63.3	0.04	5.0	50.5	40.3	43.7	0.149	62.3
Qwen3-Omni-Inst.	50.5	53.5	61.2	26.5	0.00	5.5	48.4	35.0	37.4	0.158	40.6
Qwen3-Omni-Think.	49.6	52.6	57.6	12.2	0.13	8.0	69.4	34.3	45.2	0.117	89.9
Nemotron-3-Nano	50.0	54.1	66.1	45.2	−0.02	8.3	70.7	33.6	46.6	0.149	75.4
Uni-MoE-2.0	49.2	53.9	63.2	57.7	0.31	6.2	61.3	38.1	47.7	0.123	30.4
Coach	–	–	29.3	50.0	–	4.1	–	–	53.7	–	6.8

Table C5: **Dimensions addressed** (% of the 83 commentaries) by the coaches and, on average, by the 31 backbones.

	Technique								Artistry					
	Brea.	**Into.**	**Vibr.**	**High**	**Reg.**	**Artic.**	**Reso.**	**Rhyt.**	**Expr.**	**Phra.**	**Dyn.**	**Style**	**Stage**	**Interp.**
Coaches	42.2	4.8	3.6	15.7	3.6	31.3	9.6	30.1	77.1	60.2	28.9	38.6	15.7	53.0
Backbones	84.7	35.1	26.1	56.8	18.0	48.4	51.7	18.0	85.2	64.2	57.9	19.6	30.8	20.7

accuracy unchanged (-0.6, p=0.72; $+0.3$, p=0.84). Errors agree beyond chance (Cohen's κ 0.32; 0.41 within a family), and 103 of the 937 closed items are missed by all 31 backbones, leaving ample headroom for future omni models; only 2 closed items are solved by every one of the 31 backbones.

C.7 Case Study

Protocol. Table 2 uses test clip `v2_057`, an English-coached opera clip whose coach comments on both what was heard and what was seen; script `72_case_study.py` checks that every quoted span occurs verbatim in the backbone's output or in c^*, and records every answer and score shown.

Results. The clip carries one HCC, seven TES and nine AES items. Gemini-3.7-Flash answers six of the nine AES items, against one to three for the other five backbones, alone picks the coach's observation on style among four candidates, and receives 9 of 10 on the correctness and relevance rubrics of HCC, in line with its agreement with the coach on stage presence and breath in Table 2.

D Extended Related Work

D.1 Omnimodal Large Language Models

The Specific-to-Omni survey (Jiang et al., 2025) frames unified perception as modality alignment. Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Qwen3.5-Omni (Team, 2026b), and Gemma 4 (Team et al., 2026a) develop it across generations; Baichuan-Omni-1.5 (Li et al., 2025b), Ola (Liu et al., 2025), and OmniVinci (Ye et al., 2025) study how complementary encoders and training data support joint understanding. Ming-Omni (AI et al., 2025), Ming-Flash-Omni (AI et al., 2026), and Uni-MoE-2.0 (Li et al., 2025d) use expert-based architectures, and LongCat-Flash-Omni (Team et al., 2025) and LongCat-Next (Team et al., 2026b) explore scaling and modality representations. VITA-1.5 (Fu et al., 2025), InteractiveOmni (Tong et al., 2025), MiniCPM-o-2.6 (OpenBMB, 2025), and MiniCPM-o-4.5 (Cui et al., 2026) target interactive audiovisual communication; HyperCLOVA X (Team, 2026a), Nemotron 3 Nano Omni (NVIDIA et al., 2026), Audio-Visual Flamingo (Ghosh et al., 2026a), and TLive-Omni (Hu et al., 2026b) broaden deployment settings. Gemini 2.5 (Comanici et al., 2025), 3 Flash (Google DeepMind, 2025), 3.1 Flash-Lite (Google DeepMind, 2026a) and Pro (Google DeepMind, 2026b), 3.5 Flash-Lite (Google DeepMind, 2026d) and Flash (Google DeepMind, 2026c), and 3.6/3.7/3.8 Flash (Google DeepMind, 2026e;f;g) are closed-source backbones. Modality coverage enables vocal coaching but does not establish the reliability of expert judgements.

OmniBench (Li et al., 2025c) and WorldSense (Hong et al., 2026) test joint understanding; Daily-Omni (Zhou et al., 2026b), OmniVideoBench (Li et al., 2026a), and MMOU (Goel et al., 2026) examine temporally extended audiovisual reasoning, and OmniGAIA (Li et al., 2026b) adds agentic tasks. AV-Odyssey (Gong et al., 2024) and OmniClean (Liu et al., 2026a) question whether answers require both modalities; OmniHalluc-L (Dong et al., 2026) probes evidence binding, OmniJudge (Hu et al., 2026a) diagnoses evaluation bias, and AVTrace (Zhang et al., 2026a) tests temporal grounding. These studies separate answer accuracy from reliable use of perceptual evidence. We add an expert-referenced test of vocal-coaching judgements that isolates modality contributions through matched audiovisual, audio-only, video-only, and no-media conditions.

D.2 AI for Music

MusicGen (Copet et al., 2023) and ChatMusician (Yuan et al., 2024) connect generation with musical representations; Audio Flamingo 3 (Ghosh et al., 2025a), Audio Flamingo Next (Ghosh et al., 2026b), Music Flamingo (Ghosh et al., 2025b), and GaMMA (You et al., 2026b) advance audio-language understanding. MuChoMusic (Weck et al., 2024), BASS (Jang et al., 2026), and Hum-MusQA (Weck et al., 2026) test musical reasoning; MUSE (Carone et al., 2026), Core Music Perception (Carone et al., 2025), PitchBench (Dujardin et al., 2026b), and Music I Care About (Sourada et al., 2026) isolate perceptual skills; MusTBench (Kwon et al., 2026) and MusicGrounding-Bench (Fang et al., 2026) require localizing musical evidence in time. MUSIC-AVQA (Li et al., 2022), specialized Music-AVQA designs (You et al., 2026a), and Qwen-MusicAVQA (Dehdashti, 2026) extend to audiovisual performances; ONOTE (Ma et al., 2026), MuSP-Bench (Dujardin et al., 2026a), MuseBench (Fan et al., 2026), and UniVerse (Zhou et al., 2026c) broaden representational, interpretive, and cultural coverage. Identifying musical content is a prerequisite for coaching but does not show that a model can assess execution and prescribe corrections.

For assessment, the singing-assessment survey (dos Santos & Masiero, 2026) places quality prediction (SingMOS (Tang et al., 2024), SingMOS-Pro (Tang et al., 2026), MusicJudge (Saini & Ghosh, 2026), SongSQA (Lv et al., 2026)) alongside pedagogical needs. SongEval (Yao et al., 2025), SongBench (Wu et al., 2026), MADB (Zhang et al., 2026b), and MAJEPPA (Zhou et al., 2026a) assess aesthetic or performance qualities; VocalCritic (Li et al., 2025a) and MuseCritic (Zhuang et al., 2026) make evaluations linguistically interpretable. Are You Really Listening (Zang et al., 2025), Genre Bias (Zhang et al., 2026c), and MuseDiag (Liu et al., 2026b) caution that plausible judgements may lack auditory support. Singing-mistake detection (Kumar et al., 2026) targets pedagogical errors, LLaQo (Zhang et al., 2025) generates formative feedback, and VocalCoachBench (Bang et al., 2026) evaluates expert singing diagnosis and corrective guidance from audio. We contribute an audiovisual benchmark that jointly evaluates holistic commentary, technical execution, and artistic interpretation, including actionable advice and the aspects coaches elect to address.

E Analysis Details

Probes. Table E1 lists the five probes of §4.1. Each reuses the evaluation code of Table 1 (system prompt, media files, answer budgets, reasoning settings, temperature 0 and judges) and changes only the input; Gemini-3.8-Flash runs at low reasoning effort with 2,048 extra output tokens and the other backbones at their default, as in Table 1. Before any full run, a 12-item pilot per probe and backbone estimated its cost, and each run was capped slightly above that estimate. Calls that failed four times were retried in a second session; every probe finished with all items scored. The commentary judge of each probe keeps its own cache, so no probe reuses or overwrites a judgement of Table 1. Costs are the per-call charges reported by the API and exclude the judges.

Table E1: **Probes of Section 4.** FL: Flash-Lite, F: Flash. *Calls* include the 12-item cost pilots and the retries of failed calls.

Probe	Gemini backbones	Input	Items	Calls	USD
(i) Modality	2.5-FL, 3-F, 3.1-FL, 3.5-FL, 3.8-F	`audio`, `video`	1,293	13,097	19.25
(ii) Transcript	2.5-FL, 3.8-F	a coach's c^*, no media	1,210	4,952	3.63
(iii) Genre swap	2.5-FL, 3.8-F	genre line swapped, `av`	1,020	2,100	6.90
(iv) Repeat	2.5-FL, 3.1-FL, 3.8-F	two identical `av` passes	410	2,588	6.24
(v) Thinking	2.5-FL	low-effort thinking, `av`	1,293	1,305	2.77
Annotation	`gpt-4o-mini`	Step-4 annotator	2,573	2,573	0.91
Audio check	2.5-FL, 3.8-F	`av` vs. `video`	2 clips	8	0.02
Total					**39.72**

Transcript pairing. Under "other coach", the prompt carries the transcript of a test clip of the same genre from another session, of the same coaching language where possible (95.2% of clips) and of the nearest transcript length. Jazz and Coloratura, with one test session each, take an Opera clip (4 clips). The prompt is the `none` prompt with the transcript inserted after its first line and introduced as what the coach said to the singer right after the performance. Commentary items are excluded, since the transcript is the reference c^* of HCC. Table E3 gives the scores per level.

Genre swap. The genre line is replaced by a fixed contrasting genre: Opera, Coloratura and Academic by Pop Ballad; Lyrical by Jazz; Pop Ballad, Jazz, R&B and Folk-Pop by Opera; and Folk by Lyrical. Every target genre carries its description in the prompt. Genre vocabulary is counted with fixed lists: classical technique terms (appoggio, squillo, chiaroscuro, passaggio, messa di voce, covered tone, recitative, Fach, tessitura), pop and jazz technique terms (microphone, belting, groove, riff, swing, scat, improvisation, melisma, vocal fry, conversational, breathy, grit, pocket, laid-back, behind the beat), and genre names, which are excluded from Figure 8(d) because a commentary can repeat them from the label. With genre names included, Gemini-2.5-Flash-Lite names a classical genre in 84.2% of the pop and jazz commentaries labelled opera, up from 0%.

Audio delivery. Because muted video matches `av` in Figure 7, we checked that the soundtrack reaches the backbones. For two clips, Gemini-2.5-Flash-Lite and Gemini-3.8-Flash were asked for the sung language: the API bills about 1,500 audio tokens for each `av` request, and both backbones name German for the Mozart aria of one clip only under `av`. The muted files carry no audio stream (Step 6).

Commentary annotation. The Step-4 prompt of §3.3 labels each dimension of a backbone's commentary as praised, faulted, neutral or not addressed, as it labelled the coach's commentary, with one sample at temperature 0 (the coach annotations took the majority of three samples at 0.3). The 18 commentaries under 50 characters, which HCC scores zero, count as addressing no dimension. Recall and precision in Figure 10(a) compare the addressed dimensions with those of the coach's annotation; praise shares count + among + and −.

Commentary similarity and advice topics. TF-IDF vectors use the words of three letters or more, without a short list of English function words, that occur in at least two of the 2,555 commentaries of 50 characters or more. An advice sentence is assigned every topic whose word list it matches (breath: breath, support, air, inhale, exhale, diaphragm, lungs, abdominal; the lists of the other topics are in the released code).

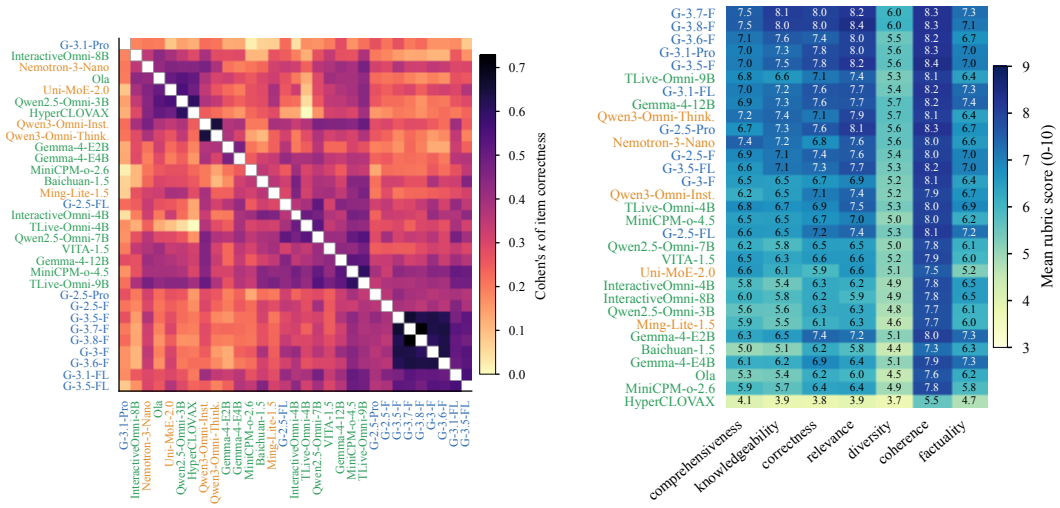
Signal detection. On the yes/no items, a hit is "yes" to a gold-"yes" item and a false alarm "yes" to a gold-"no" item; $d' = z(H) - z(F)$ and $c = -\frac{1}{2}(z(H) + z(F))$ with the log-linear correction for rates of 0 or 1.

Table E2: **Media conditions** (probe *i*). `av` and `none` are the runs of Table 1.

	HCC				TES				AES				OVS			
Backbone	none	video	audio	av	none	video	audio	av	none	video	audio	av	none	video	audio	av
Gemini-2.5-Flash-Lite	70.7	69.7	69.9	69.1	30.0	28.9	24.6	27.8	35.9	37.8	34.5	36.5	48.1	47.9	45.7	46.9
Gemini-3-Flash	59.2	72.8	72.1	66.1	34.8	28.8	32.2	34.4	35.4	35.4	34.2	38.3	44.7	48.4	48.8	48.2
Gemini-3.1-Flash-Lite	75.2	74.8	70.8	72.2	33.6	36.2	36.4	33.4	35.3	41.8	36.5	38.4	50.7	53.3	50.2	50.4
Gemini-3.5-Flash-Lite	72.2	71.3	72.3	70.4	31.6	31.5	29.1	29.3	32.8	40.3	34.3	41.0	48.2	50.0	47.9	49.2
Gemini-3.8-Flash	77.0	77.6	75.6	76.3	37.0	34.1	36.5	37.4	36.5	43.0	39.3	43.7	52.9	54.1	53.0	54.8

Table E3: **Transcripts in place of the media** (probe *ii*), on the 1,210 TES/AES items.

Backbone	**Input**	**TES L1**	**L2**	**L3**	**AES L1**	**L2**	**L3**	**TES**	**AES**
Gemini-2.5-Flash-Lite	no media	28.5	37.7	21.7	30.8	45.5	30.5	30.0	35.9
	`av`	31.1	35.6	11.6	29.1	53.6	24.4	27.8	36.5
	other coach's c^*	29.2	33.6	44.9	33.2	41.9	25.2	34.6	34.3
	own coach's c^*	46.1	87.0	91.3	48.4	88.3	96.2	71.7	74.3
Gemini-3.8-Flash	no media	32.2	41.1	39.1	33.4	48.7	24.4	37.0	36.5
	`av`	31.5	42.8	39.1	31.7	59.7	40.5	37.4	43.7
	other coach's c^*	32.0	33.9	55.1	35.2	44.8	27.5	38.4	36.6
	own coach's c^*	63.8	90.8	98.6	66.5	83.8	97.7	81.9	80.3



(a) Error consistency (b) HCC rubrics

(c) Coaching language and genre

Figure E1: **Further analyses of the 31 backbones.** (a) Cohen's κ between the item-level correctness of two backbones on the 937 closed items, clustered. (b) Mean score of each HCC rubric; across backbones, logical coherence and diversity vary least (standard deviation 0.51 and 0.49) and relevance and knowledgeability most (0.96 and 0.94). (c) OVS on the clips coached in Chinese and in English (the mean gap is 0.2 points, English higher for 18 of 31 backbones), and closed-item accuracy by genre (clips/items in parentheses; dots: backbones; tick: mean).