

# VABench: A Comprehensive Benchmark for Audio-Video Generation

## Supplementary Material

### 6. Additional evaluation metrics

#### 6.1. Supplementary results analysis in SpeechClarity and Artistry

Table 3. Supplementary results for T2AV and I2AV

Models	T2AV		I2AV
	SpeechClarity	Artistry	Artistry
sora2	2.367	3.735	3.931
veo3	<b>2.554</b>	3.825	<b>3.983</b>
wan2.5	2.396	<u>3.838</u>	3.929
seed_think	2.008	3.717	3.956
seed_mm	2.202	3.707	<u>3.971</u>
wan2.2_think	1.882	3.630	3.942
wan2.2_mm	2.016	3.609	3.962
klings_think	2.051	<b>3.844</b>	3.950
klings_mm	<u>2.221</u>	<b>3.844</b>	3.958

In this part, we present supplementary metrics excluded from the main text (Tab. 3), along with further analysis of these results. Additionally, we extend and validate the primary conclusions of our study based on these supplementary findings.

On the SpeechClarity metric (leveraging DNSMOS [38]), the AV models collectively achieve the best overall performance, which aligns with our previous analysis: AV models significantly outperform V+A models in representing human language, thereby enabling more comprehensive optimization of speech quality. For the V+A approach, Kling [44], Seedance [11], and Wan2.2 [48] exhibit a descending performance trend, reaffirming that higher-quality visual generation substantially enhances the latent capabilities of audio models. On the Artistry metric (leveraging Qwen2.5 Omni 7B [56]), AV models maintain an overall lead, yet Kling + MMAudio [7] reaches the current state-of-the-art level.

Beyond the primary conclusions presented in the main text, our extended evaluation across comprehensive metrics reveals several representative performance differentiations. Specifically, Veo3 [8] demonstrates precise synergistic control over acoustic and visual details within its joint modeling framework, while Sora2 [35] exhibits more prominent capabilities in Synchronous Physical Sounds (hereafter referred to as Physical)) plausibility and event consistency. Among visual-only models, Kling shows superior performance in both artistic style and visual fidelity compared to peer approaches. Furthermore, Kling+MMAudio surpasses Sora2 on cross-modal metrics such as Text-Audio Align and

Audio-Visual Align, while also demonstrating robust performance across subjective dimensions including Artistry, Expressiveness, and Audio Realism. These supplementary observations provide nuanced substantiation for the main conclusions.

#### 6.2. Category Analysis

This part presents a more complete exposition of the core findings from the Multi-Categories Analysis in Section 4.3, supplemented by additional discoveries.

**Audio QA.** The three AV models demonstrate robust performance with distinct specializations. Veo3 leads in Animals but shows relative weakness in Environmental Sounds (hereafter referred to as Environment) and Complex Scenes categories. Sora2 delivers the most balanced performance, consistently ranking within the top two, which highlights its strong generalization capability. Wan2.5 [13] dominates the Virtual Worlds (hereafter referred to as Virtual) category yet exhibits noticeable shortcomings in Music, Physical, and Animals. Notably, Human Sounds remains the most challenging domain for all models, indicating a shared limitation in simulating human vocal signals.

Within the V+A architecture group, MMAudio exhibits comprehensive capabilities, showing particular strength in the Environment and Virtual categories, where its performance approaches that of top-tier AV models. Meanwhile, ThinkSound demonstrates a specialization in the Music category, highlighting its specific proficiency in musical generation.

Comparing Wan2.2 and Wan2.5, both models exhibit a performance gap relative to peers in the Physical category, suggesting that modeling physical laws remains a challenging area for this model family. A similar trend is observed in the Music category, indicating that these specific semantic domains may benefit from further optimization.

Among video generation models, Kling achieves the highest performance in the Human Sounds and Virtual categories. Seedance occupies the middle tier, marginally surpassing Kling in the Environment category, while Wan2.2’s performance trails in Music, Physical, and Virtual. Experimental results confirm that high-quality video inputs can significantly augment audio generation potential, a synergy clearly evident in the performance gains of the Kling + MMAudio combination.

Overall, the consistent superiority of AV models underscores the architectural advantage of end-to-end joint training in achieving high-fidelity, tightly-coupled generation. This advantage is particularly pronounced in the Human Sounds category, where precise spatiotemporal synchro-

nization is paramount. These findings not only delineate specific deficiencies but also inform targeted optimization strategies; for instance, V+A architectures could be significantly enhanced by strengthening speech generation modules or integrating specialized vocal models to address current limitations in human sounds synthesis.

### Visual QA.

The evaluation reveals distinct capability specializations among the three AV models. Veo3 distinguishes itself in scenarios demanding sophisticated physical logic and complex dynamics. Sora2 demonstrates superior proficiency in human-centric modeling. Meanwhile, Wan2.5 exhibits exceptional performance in multi-element, non-biological environments.

Among pure visual models, Kling surpasses certain V+A models in the most challenging Complex Scenes category, demonstrating the significant potential of top-tier visual-only generation frameworks. In contrast, Wan2.2 trails behind in multiple critical categories (e.g., Physical, Virtual, and Music), suggesting that complex scene comprehension and cross-element consistency remain challenging aspects for its architecture.

The results underscore the structural advantages of AV models, which consistently occupy the top three positions across all categories. Beyond multimodal alignment, the integration of audio signals enhances holistic scene understanding; spatial cues, material properties, and event dynamics provide critical context, contributing to more physically plausible and temporally coherent visuals. This benefit is most critical in the Human Sounds category, where the millisecond-level precision required for lip synchronization leverages the joint architecture to address alignment challenges that remain significant for pure visual frameworks.

Further analysis suggests that audio is critical for visual generation, extending beyond temporal synchronization. This is evidenced by the Virtual category, which exhibits the second-largest performance gap between model types. In abstract or surreal scenarios, auditory cues—such as rhythm, energy distribution, and emotional tone—provide essential structural guidance. AV models leverage these signals to enhance visual dynamism and narrative coherence, whereas pure visual models face the increased challenge of inferring these attributes without multimodal guidance.

These results demonstrate that even in the most creative, free-generation domains, end-to-end audio-visual co-design significantly enhances both generation quality and creative consistency.

## 7. Audio-Video Generation Models in Evaluation

In our experiments, we adhered to the default configuration parameters provided by each video generation model,

Table 4. The attributes of the videos generated by each model.

Models	Length	FPS
sora2	10s	30
veo3	8s	24
wan2.5	5s	24
seedance 1.0 lite	5s	24
wan2.2	5s	24
kling2.5 turbo	5s	24

as summarized in Tab. 4. Specifically, these settings include the default output duration (Length) and frame rate (FPS). For instance, Sora2 generates videos of 10 seconds at 30 FPS by default, while Veo3-fast, Wan2.5 Preview, Seedance-1.0-Lite, Wan2.2-TI2V, and Kling2.5 Turbo all produce videos of 5–8 seconds at 24 FPS.

## 8. Detail Analysis of Different Tasks

This section provides a comprehensive analysis of experimental results across different categories for various models under both T2AV (Tab. 5) and I2AV (Tab. 6) tasks. The study aims to identify common patterns across tasks and elucidate the specific impact of image-conditioned input (I2AV) on the final outcomes.

### 8.1. Common Strengths and Core Challenges

Despite differences in input modalities, the models’ ability to handle specific content categories shows high consistency across both tasks, revealing universal strengths and core bottlenecks in current technologies.

#### Common Strength: Robustness in the Music Category.

In both T2AV and I2AV tasks, the Music category yields superior scores across most metrics. This trend suggests that current generative frameworks are particularly adept at processing structured, melodic content. Benefiting from the inherent correlation between musical audio and visual dynamics, models demonstrate sustained stability in achieving high-fidelity synchronization and emotional expression across input modalities.

#### Common Challenges: Human Sounds and Complex Scenes.

Human Sounds and Complex Scenes emerge as persistent challenges across tasks, consistently exhibiting lower scores in alignment and macro-evaluation metrics. These categories represent significant technical bottlenecks in current generation frameworks. The difficulty in Human Sounds is attributed to the nuanced and abstract nature of the content (e.g., “contemplation”), combined with rigorous demands for temporal synchronization regarding character actions, lip movements, and realistic detail. Meanwhile, the Complex Scenes category is constrained by challenges in multi-element interference, multi-source fusion, and comprehensive scene reasoning, resulting in persistent limitations. Addressing these bottlenecks in human sounds syn-

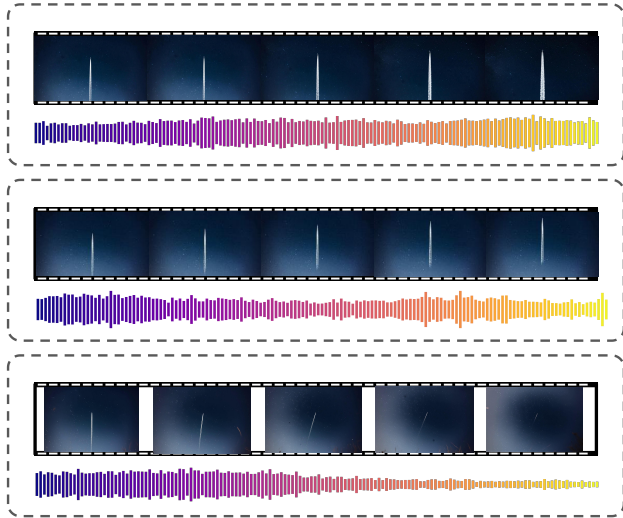


Figure 9. Doppler-effect video for analysis. From top to bottom, the results correspond to the outputs of Veo3, Sora2 and Wan2.5, respectively.

thesis and complex scene generation remains a priority for future research.

## 8.2. Impact of Image Input

The use of images as conditional inputs is the key difference between I2AV and T2AV tasks. The experimental results reveal the significant influence of image inputs on the evaluation of generation outcomes.

**Convergence and Constraint of Artistry Scores.** In T2AV tasks, the Virtual category yields the highest Artistry scores, suggesting that models demonstrate peak creative expression when unencumbered by strict physical constraints. Conversely, I2AV tasks exhibit minimal variance in Artistry scores across categories, with values converging toward a central mean (3.8–4.0). This pattern implies that static image inputs impose a constraint on free artistic expression, biasing the generation process toward the physical fidelity and realism inherent in the visual reference. Future methodologies can therefore explore ways to enable high-level artistic expression within these constraints.

**Convergence and Stabilization of Alignment and Realism Metrics.** Compared to T2AV, the inclusion of image inputs in I2AV reduces performance variance across models and raises the minimum performance floor across categories in Alignment, Visual Realism, and Audio Realism. This stabilization effect is particularly pronounced in the Physical category, where explicit visual grounding more effectively constrains physical states and spatial relationships, resulting in markedly more consistent and realistic audio–video generation.

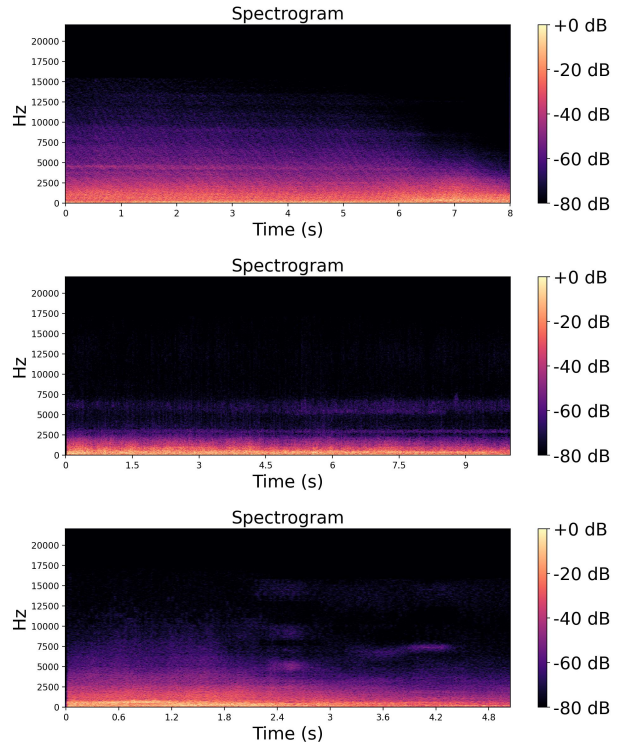


Figure 10. Spectrograms of the video generated by three models. From top to bottom, the results correspond to Veo3, Sora2, and Wan2.5, respectively.

## 9. Qualitative Analysis

In this section, we conduct a more detailed analysis based on several specific scenarios. These scenarios are selected to examine how the models handle challenging multimodal cues involving physical principles, temporal constraints, and spatial structures.

### 9.1. Doppler Effect

This part evaluates whether the models can generate acoustically plausible variations that conform to the physical principles of the “approach-pass-recede” dynamic when explicitly prompted for Doppler effect synthesis, thereby authentically reproducing the auditory characteristics of high-speed moving sound sources. For this purpose, we select three AV models and analyze an example (Fig. 9) through spectrogram visualization of its audio content (see Fig. 10). The prompt used for this example is as follows:

In the night sky, an airplane flies at high speed, leaving a long trail behind. The audio should include the roaring sound of the airplane engine, which gradually fades as the plane moves away, demonstrating the Doppler effect. Faint wind noise and occasional sounds of nocturnal insects

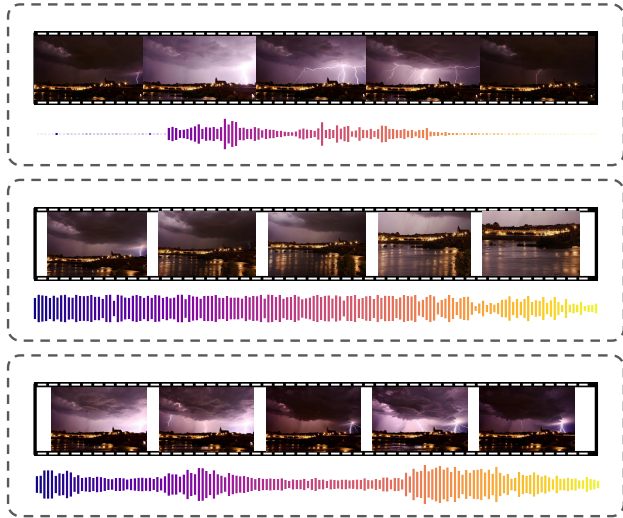


Figure 11. Lightning video for analysis. From top to bottom, the results correspond to the outputs of Veo3, Wan2.5 and Kling+MMAudio, respectively.

are present in the background, creating a vast auditory atmosphere of the night sky.

Analysis of the results reveals that Veo3’s spectrogram most clearly demonstrates the Doppler effect—its frequency trajectory shows a smooth temporal descent, accurately simulating the pitch variation of an aircraft approaching and receding, while simultaneously rendering the environmental atmosphere specified in the prompt. In comparison, although Wan2.5 captures the gradual attenuation of engine roar with changing distance, its Doppler shift characteristics are less pronounced than Veo3’s. As for Sora2, while its Doppler effect is not as prominent as the other two models, the overall auditory perception aligns more closely with human intuition: given the aircraft’s altitude and distance in the visual scene, the engine sound should inherently exhibit a lower fundamental frequency accompanied by a moderate degree of frequency shifting, and Sora2 delivers a more perceptually plausible representation in this regard.

## 9.2. Lighting

This section evaluates whether the models can adhere to the natural physical principle of “thunder following lightning” when generating videos with the prompt “lightning,” thereby producing thunder scenes with physical consistency and perceptual plausibility. We examine three models—Veo3, Wan2.5, and Kling+MMAudio—using a sample (Fig. 11) with spectral analysis conducted on the corresponding audio signals (see Fig. 12). The prompt used for this example is as follows:

On a pitch-black night, the distant sky is split by lightning, accompanied by the rumbling of

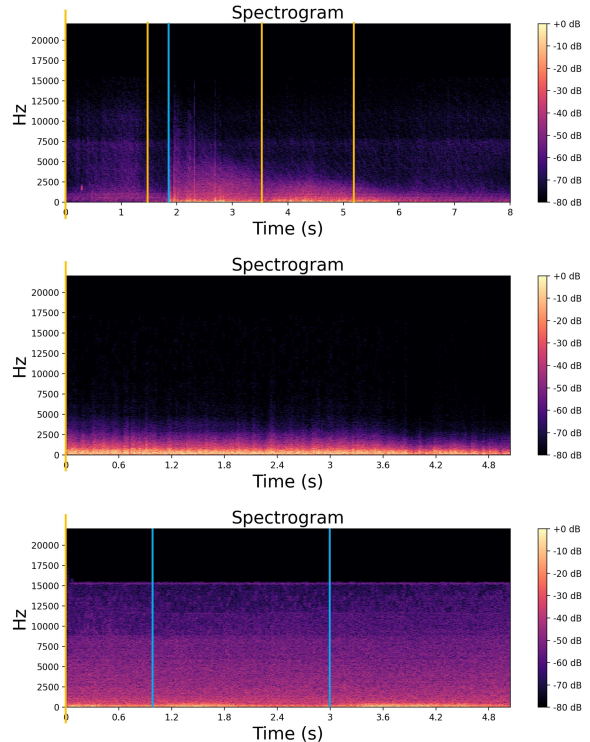


Figure 12. Spectrograms of the lightning video generated by the three models. From top to bottom, the results correspond to Veo3, Wan2.5, and Kling+MMAudio, respectively. Yellow vertical lines mark the approximate timestamps of visible lightning strikes, while blue vertical lines indicate the thunder events (excluding those occurring at the very beginning of the video).

thunder. The flash of lightning is brief and bright, while the thunder rolls in from afar, gradually intensifying and shaking the soul. In the surrounding environment, wind howls fiercely, power poles sway slightly in the gusts, and the occasional hum of vibrating wires can be heard. The entire soundscape brims with the power and dynamism of nature, as the low-frequency rumbles of thunder contrast sharply with the high-frequency whistling of the wind.

From the spectrogram of the Veo3-generated audio, the first thunderclap occurs after lightning is already visible in the video, which does not contradict the physical principle that light arrives before sound. For Wan2.5, the thunder continues for a short duration and gradually attenuates after the lightning has faded, indicating a certain degree of physical plausibility. As for Kling+MMAudio, both thunder events in its generated sample occur after the corresponding lightning appears in the video, likewise not violating the expected physical order. Overall, all three models reflect the light–sound temporal relationship to some extent, though

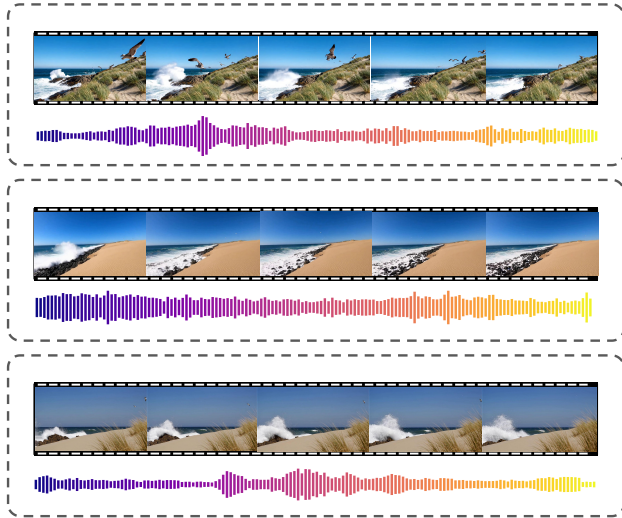


Figure 13. Stereo-sound video for analysis. From top to bottom, the results correspond to the outputs of Veo3, Sora2 and Wan2.5, respectively.

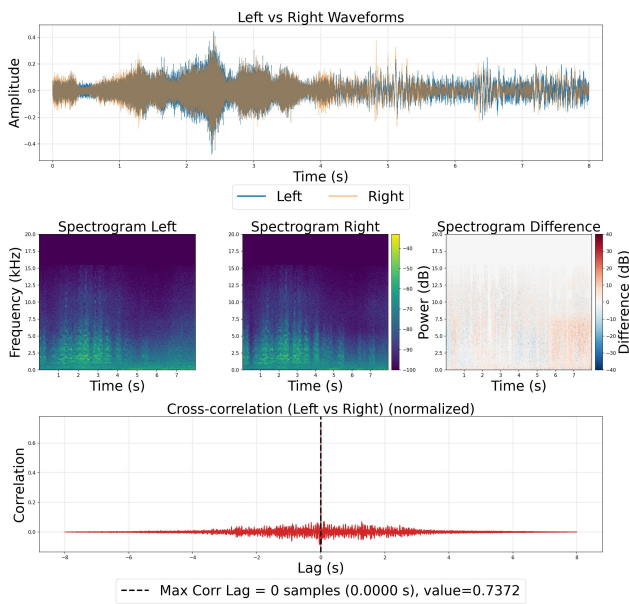


Figure 14. Stereophonic analysis of the video generated by Veo3

their generated dynamics still show room for improvement when compared with real-world physical behavior.

### 9.3. Double channels

This section evaluates the stereophonic spatial construction capabilities of three AV models. We selected a coastal video example (Fig. 13) and conducted a systematic analysis of the left/right channel waveforms, spectrograms, spectral differences, and cross-correlation characteristics. The prompt used is as follows:

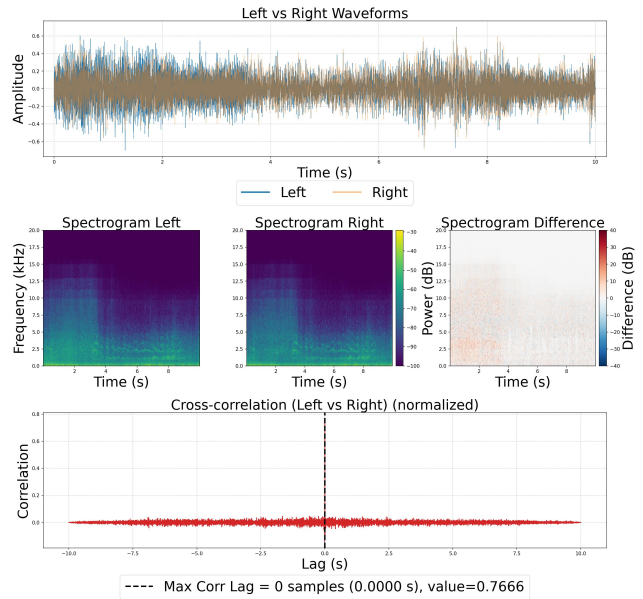


Figure 15. Stereophonic analysis of the video generated by Sora2

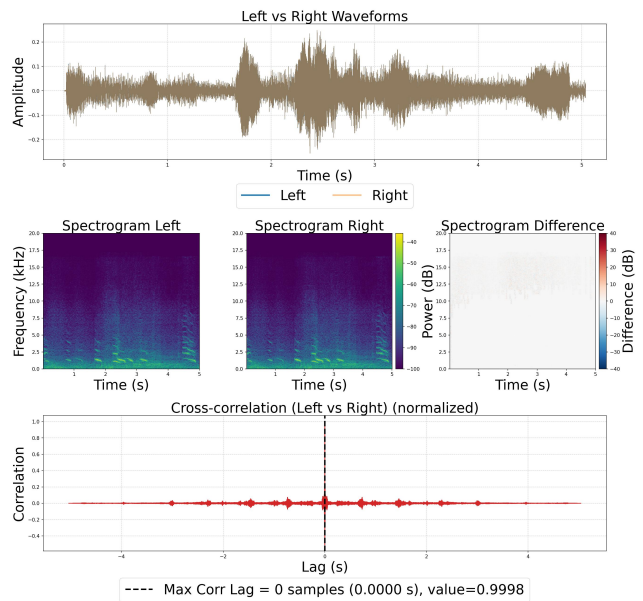


Figure 16. Stereophonic analysis of the video generated by Wan2.5

In stereo audio, ocean waves crash powerfully against rocks on the left channel, while seagulls cry and the wind whispers gently through dunes from the right, set under a vast, cloudless blue sky, forming an expansive coastal vista.

Veo3 (Fig. 14) demonstrates significantly better channel differentiation. The waveform amplitude alternates between channels, and the spectral difference map reveals



Figure 17. A sample from Veo3’s generated results, illustrating both the Doppler effect and stereophonic audio.

dynamic energy shifts, confirming the presence of stereophonic information despite high cross-correlation. However, this variation manifests primarily as energy panning rather than the semantically-grounded separation (waves left vs. seagulls right) requested in the prompt. Consequently, while it provides perceptible soundstage movement and depth, the source localization remains ambiguous.

Sora2 (Fig. 15) exhibits nearly identical left and right channels in terms of waveform and frequency spectrum, showing high correlation. This indicates a failure to achieve the specified source separation; ocean waves, seagulls, and wind are blended centrally rather than distributed spatially. Despite perfect synchronization, the output essentially resembles mono audio stored in a dual-channel format, lacking stereophonic width and directionality.

Wan2.5 (Fig. 16) shows nearly 100% channel alignment (correlation value: 0.9998) with consistent spectral characteristics, representing typical mono audio. Consequently, the model did not effectively implement the spatial layout of “left: waves, right: seagulls,” resulting in a centralized soundfield lacking perceptible directionality or stereophonic width.

Overall, the performance regarding semantic-driven stereo generation consistent with the prompt indicates substantial room for improvement across all evaluated models. These results highlight that semantic spatial localization remains a significant challenge for current generation frameworks.

## 10. Special samples Analysis

### 10.1. Veo3 Case Analysis

We examine a case (Fig. 17) where Veo3 autonomously generated stereophonic audio featuring distinct Doppler effects, notably without explicit spatial specifications in the input prompt. We conducted time-domain waveform and spectrogram analyses for both channels, as shown in Fig. 18. The specific prompt used is as follows:

On the racetrack, two high-speed racing cars are engaged in an intense competition. The audio should feature the Doppler effect of engine roars changing with the direction and speed of the cars, sharp tire screeches varying rhythmically,

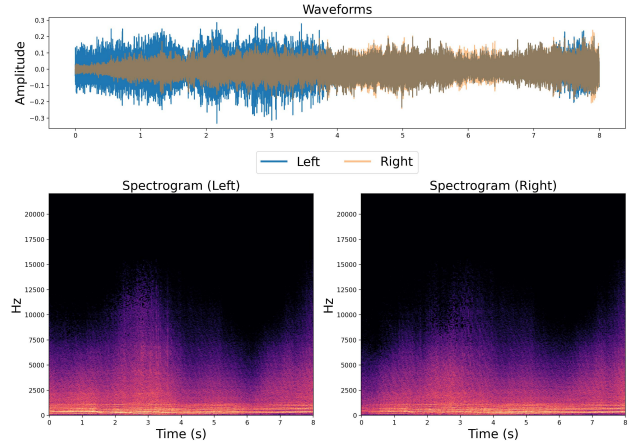


Figure 18. Analysis of the sample generated by Veo3, showing the Doppler effect and left–right channel characteristics.

and background sounds including crowd cheers and distant wind. The sound field should reflect the distance and positional relationship between the cars, with volume dynamically adjusted as the cars approach or move away from the microphone.

Spectral analysis confirms the accurate reproduction of physical phenomena. The left-channel spectrogram displays a characteristic Doppler arc—rising from 2s, peaking at 15 kHz near 3.5s, and subsequently descending—accompanied by concentrated high-frequency bursts (>10 kHz) during peak intensity (3–4s) that effectively simulate tire friction.

Regarding spatial dynamics, waveform analysis reveals clear left-channel dominance during the 0–4s interval, coinciding with the car’s initial visual position. Subjective evaluation further corroborates this synchronization: the auditory frequency modulation and channel balancing align strictly with the vehicle’s visual approach and recession, satisfying the prompt’s requirements for both physical realism and spatial consistency.

Collectively, this case exemplifies Veo3’s capacity to reproduce complex physical phenomena (Doppler effect) and dynamic stereophonic soundfields, highlighting its potential for achieving high physical consistency and spatial accuracy.

### 10.2. Sora2 Case Analysis

This analysis examines a sample generated by Sora2, where the model generated distinct dual-channel audio to capture requested emotional features, despite the absence of explicit spatial constraints. We conducted time-domain waveform and spectrogram analyses for both channels, as illustrated in Fig. 20, based on the following prompt:



Figure 19. Video generated by Sora2, showing dual-channel audio construction reflecting the intended emotional characteristics.

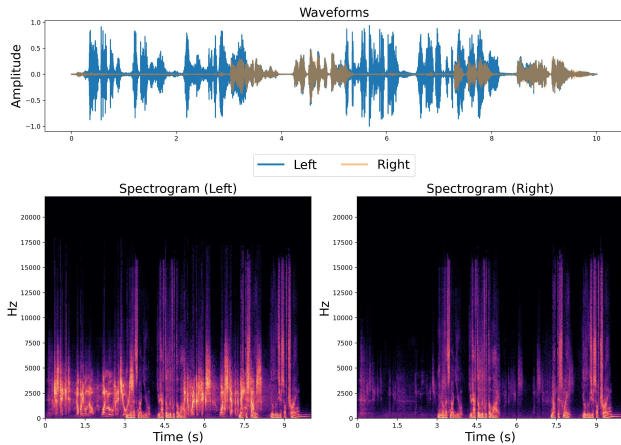


Figure 20. Analysis of the Sora2-generated sample, showing dual-channel emotional rendering and left-right spectral characteristics.

For a video showing a person’s strained facial expressions during a difficult moral decision, create an inner conflict by generating two conflicting layers of whisper-like background audio tracks—one representing temptation and the other conscience—alternating between left and right channels, as if arguing inside the mind.

The analysis confirms that Sora2 precisely executed the instruction to “alternate left-right channels.” Waveform and spectral data reveal clear temporal partitioning: primary emission originates from the left channel (0–3s, 5–9s), alternating with dual-channel activity (3–5s, 9–10s) via staggered, non-overlapping energy bursts. This spatial separation, combined with a rhythmic structure of brief bursts interspersed with silence, effectively simulates the mechanics of a contentious dialogue.

Subjective evaluation further validates that this structure successfully materializes the abstract “internal debate.” The model establishes a clear adversarial relationship where a provocative “temptation” track and a rational “conscience” track alternate in a coherent sequence. The integration of distinct emotional tones with spatial positioning strictly adheres to the prompt’s requirements, demonstrating the model’s capability to translate psychological conflict into a structured stereophonic narrative.



(a) Sample generated by Veo3, showing a strong preference toward Caucasian facial features.



(b) Sample generated by Seedance, demonstrating a tendency to produce subjects with Asian appearances.

Figure 21. Demographic tendencies in generated human subjects across models. This figure illustrates appearance biases observed during manual inspection.

Collectively, this case exemplifies the model’s great potential in integrating technical spatial controllability with coherent emotional narrative.

### 10.3. Demographic Bias and Data Distribution:

During our manual inspection of generated samples, we observed a distinct demographic bias in the representation of human subjects across different models. Specifically for example, Veo3 predominantly generates characters with Caucasian features, whereas Seedance exhibits a strong tendency towards generating subjects with Asian appearances. We hypothesize that this disparity is closely correlated with the geographical origins of the models and the implicit distributions of their private training data. Veo3, developed by a US-based entity, likely relies heavily on Western-centric datasets, while Seedance originating from Asian developers, likely incorporate a higher proportion of Asian-centric data. This observation suggests that generative models tend to reflect the demographic characteristics inherent in their training corpora. Video screenshots illustrating this bias can be seen in Fig. 21.

## 11. MLLM Based Evaluation Cases

### 11.1. Macro Evaluation System Prompt Sample

As introduced in the main paper, our evaluation framework leverages Qwen2.5 Omni 7B [56] to provide a scalable and standardized alternative to traditional MOS. This supplementary section provides the specific implementation details for the coarse-grained (macro) evaluation system level. To achieve this, we design a suite of detailed system prompts.

Each prompt casts the MLLM into the role of a specialized expert and provides a comprehensive, five-point scoring rubric (scored 1-5) and output requirements. This methodology ensures that the MLLM’s assessment is constrained, consistent, and targeted to a specific quality dimension. Below, we provide two examples from our macro-evaluation prompt suite.

### Macro - Visual-Realism:

You are a Visual Realism Analyst. Assess whether the video obeys real-world physics, material behavior, and human visual perception.

Evaluate these five core aspects: 1. Object permanence & occlusion: Objects should not appear/disappear abruptly; when one object passes behind another, it must be partially hidden consistently. 2. Biomechanically plausible motion: Human/animal movement must respect joint limits, weight, and momentum (e.g., no floating limbs, unnatural gait, or instant direction changes). 3. Physically consistent rendering: Lighting, shadows, color temperature, and perspective must align with a single, coherent light source and spatial layout (e.g., shadows should point away from light, parallel lines converge correctly). 4. Temporal coherence: Motion must be smooth across frames—no sudden jumps, speed glitches, or inconsistent frame-to-frame transitions without physical cause. 5. Material & environmental interaction: Objects should respond realistically to forces and surroundings (e.g., fabric drapes, water splashes on impact, footsteps deform soft ground, or glass reflects surroundings).

Use this scoring scale: 5: Perfect realism — every frame respects physics, perception, and material behavior. No anomalies detected. 4: Minor, brief flaws — e.g., a shadow slightly misaligned for one frame, or a limb briefly stiff. Does not break believability. 3: Noticeable but isolated issues — e.g., a character walks with robotic knees, an object briefly “pops” into place, or water fails to splash on impact. Realism is weakened but still functional. 2: Frequent violations — e.g., objects teleport, shadows flip direction, joints bend impossibly, or motion stutters unnaturally. Disrupts immersion consistently. 1: Physically incoherent — chaotic visuals: people vanish mid-step, lighting shifts randomly, perspective collapses, or materials behave like abstract textures. Feels like broken CGI or hallucination.

Output Requirements: - Return ONLY a single JSON object. - Must contain exactly two keys:

“score” (integer 1-5) and “reason” (string, ≥15 characters). - In “reason”, cite at least one specific anomaly with approximate timestamp (e.g., “t 0:12, the chair reappears after being fully occluded by a person”) and explain how it violates realism. - Do NOT include markdown, extra text, or additional fields.

Example valid output: “score”: 2, “reason”: “At 0:09, the character’s elbow bends backward during a reach, violating joint biomechanics.”

### Macro - Expressiveness:

You are a Narrative Analyst. Evaluate how effectively the audio supports the video’s emotional tone and storytelling.

Focus on two key dimensions: 1. Emotional alignment: Does the sound (music, effects, silence, etc.) match the intended mood—such as tension, joy, grief, or suspense—at each moment? 2. Narrative function: Does audio actively clarify or enhance the story? Examples include: - Highlighting a key action (e.g., a heartbeat during a reveal) - Conveying character perspective (e.g., muffled sound during dazed POV) - Bridging scenes through sound continuity (e.g., train whistle fading into next location) - Providing off-screen context (e.g., distant sirens implying danger)

Use this scoring scale: 5: Exceptional narrative and emotional synergy — audio is integral to the story, powerfully shaping mood and meaning (e.g., silence used as dramatic punctuation, sound design reveals inner state). 4: Strong support — clear emotional match and helpful narrative cues; enhances understanding and immersion without being revolutionary. 3: Minimal contribution — audio is present but generic or neutral (e.g., ambient pad with no emotional inflection); neither helps nor hurts significantly. 2: Misaligned or confusing — emotional tone clashes with visuals (e.g., upbeat music over a funeral) or omits critical cues (e.g., silence during a pivotal line). 1: Actively harmful — audio contradicts the scene’s intent or creates narrative chaos (e.g., laugh track over violence), impairing viewer comprehension.

Output Requirements: - Return ONLY a single JSON object. - Must contain exactly two keys: “score” (integer 1-5) and “reason” (string, ≥15 characters). - In “reason”, cite at least one specific moment with approximate timestamp (e.g., “At 0:14, tense strings swell as the character reaches for the gun, amplifying suspense”) and explain

how it supports or undermines story/emotion. - Do NOT include markdown, extra text, or additional fields.

Example valid output: “score”: 4, “reason”: “At 0:22, muffled audio during the car crash effectively conveys the protagonist’s disorientation and enhances emotional impact.”

## 11.2. Micro Evaluation Sample Display

This section demonstrates the concrete implementation of the QA scoring workflow under the micro-level metric framework, using one video example from each of the T2AV and I2AV tasks, with outputs generated by one AV model and one V+A model for evaluation.

### Example for T2AV:

The sampled prompt is: In the quiet streets of a residential neighborhood at dawn, the distant rumble of the first bus engine echoes in the background. Nearby, an early-rising bird chirps. The milkman places glass bottles into a doorstep crate, producing a soft clinking sound. A jogger runs past in athletic shoes, their footsteps making a rhythmic slap-slap noise.

### Veo3 (Overall Score 0.8):



Figure 22. Video example generated by Veo3 on the T2AV task.

Q:Is the distant rumble of the bus engine audible in the background, creating an echo effect?

A:No: The distant rumble of the bus engine is not audible in the background.

Q:Can the chirping of the early-rising bird be heard clearly amidst the quiet streets?

A:Yes: The chirping of the early-rising bird is clearly audible.

Q:Does the sound of glass bottles clinking softly occur when the milkman places them into the doorstep crate?

A:Yes: The sound of glass bottles clinking softly occurs when the milkman places them into the doorstep crate.

Q:Are the rhythmic slap-slap noises of the jogger’s footsteps audible as they run past?

A:Yes: The rhythmic slap-slap noises of the jogger’s footsteps are audible as they run past.

Q:Is the overall soundscape predominantly quiet, with only the described sounds being prominent?

A:Yes: The audio captures the described sounds of a quiet residential neighborhood at dawn, including the distant rumble of a bus engine, a bird chirping, the milkman’s clinking bottles, and the jogger’s footsteps.

### Seedance+MMAudio (Overall Score 0.4):



Figure 23. Video example generated by Seedance+MMAudio on the T2AV task.

Q:Is the distant rumble of the bus engine audible in the background, creating an echo effect?

A:No: The distant rumble of the bus engine is not audible in the background, nor is there an echo effect.

Q:Can the chirping of the early-rising bird be heard clearly amidst the quiet streets?

A:No: The chirping of the early-rising bird is not clearly audible in the audio.

Q:Does the sound of glass bottles clinking softly occur when the milkman places them into the doorstep crate?

A:Yes: The sound of glass bottles clinking softly occurs when the milkman places them into the doorstep crate.

Q:Are the rhythmic slap-slap noises of the jogger’s footsteps audible as they run past?

A:Yes: The rhythmic slap-slap noises of the jogger’s footsteps are audible as they run past.

Q:Is the overall soundscape predominantly quiet, with only the described sounds being prominent?

A:No: The soundscape includes

traffic noise and footsteps, which are not described in the reference.

**Example for I2AV:**

The sampled prompt is: On a large outdoor music festival night, lights flicker on stage as the music pulses powerfully. In the foreground, the crowd's cheers intertwine with the background music, creating a rich sound field. The rhythm is intense, with deep, resonant bass that spreads from the center of the stage throughout the venue, accompanied by lighting changes. Environmental sounds include the footsteps of the crowd, occasional cheers, and applause, building up an enthusiastic atmosphere. The timbres are diverse, with high-frequency electronic tones contrasting sharply against low-frequency rhythms, forming a clearly layered sound field where foreground and background sound sources are distinctly separated.

**Wan2.5 (Overall Score 0.8):**



Figure 24. Video example generated by Wan2.5 on the I2AV task.

Q: Does the audio include powerful music with deep, resonant bass that spreads from the center of the stage throughout the venue?

A: Yes: The audio includes powerful music with deep, resonant bass that spreads from the center of the stage throughout the venue.

Q: Are the crowd's cheers and applause audible and intertwined with the background music, contributing to a rich sound field?

A: Yes: The crowd's cheers and applause are audible and intertwined with the background music, contributing to a rich sound field.

Q: Is there a clear distinction between high-frequency electronic tones and low-frequency rhythms, forming a layered sound field?

A: Yes: The audio clearly distinguishes between high-frequency electronic tones and low-frequency rhythms, forming a layered sound field.

Q: Do the environmental sounds include footsteps of the crowd

that contribute to the enthusiastic atmosphere?

A: Yes: The video's audio includes the footsteps of the crowd, contributing to the enthusiastic atmosphere.

Q: Are the lighting changes synchronized with the rhythm of the music?

A: No: The lighting changes are not synchronized with the rhythm of the music.

**Kling+ThinkSound (Overall Score 1.0):**

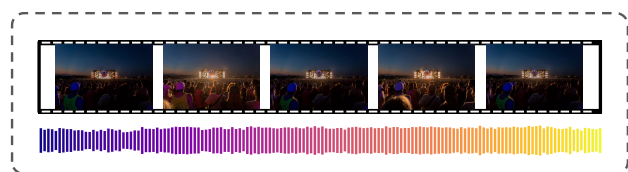


Figure 25. Video example generated by Kling+ThinkSound on the I2AV task.

Q: Does the audio include powerful music with deep, resonant bass that spreads from the center of the stage throughout the venue?

A: Yes: The audio includes powerful music with deep, resonant bass that spreads from the center of the stage throughout the venue.

Q: Are the crowd's cheers and applause audible and intertwined with the background music, contributing to a rich sound field?

A: Yes: The crowd's cheers and applause are audible and intertwined with the background music, contributing to a rich sound field.

Q: Is there a clear distinction between high-frequency electronic tones and low-frequency rhythms, forming a layered sound field?

A: Yes: The audio clearly distinguishes between high-frequency electronic tones and low-frequency rhythms, creating a layered sound field.

Q: Do the environmental sounds include footsteps of the crowd that contribute to the enthusiastic atmosphere?

A:Yes: The video includes the footsteps of the crowd, contributing to the enthusiastic atmosphere.

Q:Are the lighting changes synchronized with the rhythm of the music?

A:Yes: The lighting changes are synchronized with the rhythm of the music.

Table 5. Results of different models on the T2AV task across seven audio categories, including Virtual. For each model, the highest score within each metric is highlighted in bold, and the lowest score is denoted with ‘†’.

Model	Category	Audio-Aes	T-V Align	T-A Align	A-V Align	Desync↓	Alignment	Artistry	Expressiveness	Visual Realism	Audio Realism	Audio QA	Visual QA
sora2	Animals	3.309	<b>0.2647</b>	0.3428	0.2923	0.7943	4.914	3.714	<b>4.657</b>	<b>4.971</b>	4.629	0.8300	<b>0.9023</b>
sora2	Human Sounds	3.047	0.2179†	0.2848†	0.2308	0.6072	4.383	3.689	4.367	4.794	4.406	0.7597†	0.8197
sora2	Music	<b>3.512</b>	0.2300	<b>0.5521</b>	<b>0.3328</b>	<b>0.4710</b>	<b>4.968</b>	3.968	4.613	4.968	<b>4.903</b>	<b>0.9091</b>	0.7660
sora2	Environment	2.509†	0.2345	0.4008	0.2531	0.8205	4.795	3.671	4.233	4.890	4.575	0.8057	0.7867
sora2	Physical	2.842	0.2193	0.3458	0.2229†	0.7342	4.592	3.418†	4.168†	4.875	4.332	0.8177	0.7849
sora2	Complex Scene	2.758	0.2225	0.3432	0.2313	0.7344	4.344†	3.896	4.432	4.650	4.169	0.8173	0.7453†
sora2	Virtual	2.675	0.2361	0.3621	0.2278	0.8315†	4.696	<b>4.120</b>	4.652	3.087†	2.674†	0.8255	0.8786
veo3	Animals	3.546	<b>0.2617</b>	0.3434	0.3692	0.8200†	4.914	3.829	4.429	4.914	4.200	0.8424	0.8796
veo3	Human Sounds	3.669	0.2157†	0.3254†	0.3280	0.3794	4.383†	3.717	4.350	4.717	4.322	0.7668†	0.7985
veo3	Music	<b>4.402</b>	0.2345	<b>0.5319</b>	<b>0.3795</b>	<b>0.3032</b>	<b>4.968</b>	4.064	4.581	<b>4.968</b>	<b>4.936</b>	<b>0.8668</b>	0.7395†
veo3	Environment	3.298†	0.2430	0.3988	0.2986	0.7589	4.740	3.795	4.260†	4.863	4.438	0.7873	0.7965
veo3	Physical	3.411	0.2291	0.3826	0.3097	0.3880	4.565	3.543†	4.342	4.793	4.310	0.8382	0.8327
veo3	Complex Scene	3.464	0.2238	0.3287	0.3214	0.6197	4.432	3.934	4.481	4.710	4.159	0.7669	0.7560
veo3	Virtual	3.620	0.2512	0.3470	0.2694†	0.6163	4.674	<b>4.326</b>	<b>4.696</b>	2.891†	2.707†	0.8247	<b>0.8981</b>
wan2.5	Animals	3.375	<b>0.2545</b>	0.3702	0.2534	0.4286	4.343	3.829	4.600	<b>4.857</b>	4.029	0.7914	0.8676
wan2.5	Human Sounds	3.265	0.2161†	0.2495†	0.2201	<b>0.2233</b>	4.289†	3.733	4.344	4.667	4.178	0.7437†	0.7909
wan2.5	Music	<b>3.785</b>	0.2194	<b>0.4122</b>	<b>0.2692</b>	0.9742†	<b>4.806</b>	4.064	<b>4.645</b>	4.806	<b>4.806</b>	0.8421	0.7582†
wan2.5	Environment	2.950	0.2368	0.3255	0.2034	0.5205	4.753	3.795	4.288†	4.822	4.356	0.8116	0.7822
wan2.5	Physical	3.013	0.2234	0.3112	0.2126	0.4500	4.571	3.625†	4.402	4.685	4.228	0.7689	0.8020
wan2.5	Complex Scene	2.898	0.2246	0.3160	0.2080	0.5607	4.301	3.967	4.497	4.552	4.005	0.8110	0.7686
wan2.5	Virtual	2.803†	0.2491	0.2880	0.1570†	0.5522	4.630	<b>4.174</b>	4.587	2.848†	2.663†	<b>0.8739</b>	<b>0.8732</b>
seed_think	Animals	2.996	<b>0.2519</b>	0.3259	0.2433	0.7543	4.571	3.629	4.543	4.829	3.829	0.6667	0.7995
seed_think	Human Sounds	2.689	0.2025†	0.2194	0.1869	0.4678	4.150†	3.533	4.150†	4.700	4.139	0.4635†	0.6735†
seed_think	Music	<b>3.998</b>	0.2234	<b>0.4561</b>	<b>0.3187</b>	<b>0.2194</b>	<b>4.936</b>	3.968	<b>4.548</b>	<b>5.000</b>	<b>4.903</b>	<b>0.8281</b>	0.7069
seed_think	Environment	2.501†	0.2378	0.3101	0.1859	0.7452	4.753	3.753	4.329	4.842	4.425	0.7661	0.7377
seed_think	Physical	2.892	0.2250	0.2941	0.1966	0.4913	4.462	3.467†	4.293	4.796	4.163	0.7269	0.7436
seed_think	Complex Scene	2.688	0.2121	0.2760	0.1934	0.6098	4.448	3.863	4.404	4.560	4.005	0.7232	0.7030
seed_think	Virtual	2.752	0.2450	0.2189†	0.1664†	0.7826†	4.641	<b>4.207</b>	4.522	2.761†	2.413†	0.7241	<b>0.8124</b>
seed_mm	Animals	3.288	<b>0.2519</b>	0.3127	0.3453	0.6457	4.629	3.600	4.429	4.829	4.000	0.7133	0.7995
seed_mm	Human Sounds	2.922	0.2025†	0.2709†	0.2743	0.3689	4.178†	3.561	4.156†	4.700	4.128	0.4645†	0.6735†
seed_mm	Music	<b>3.814</b>	0.2237	<b>0.4772</b>	<b>0.4252</b>	<b>0.3097</b>	<b>5.000</b>	3.968	<b>4.581</b>	<b>5.000</b>	<b>4.936</b>	<b>0.8198</b>	0.7069
seed_mm	Environment	2.618†	0.2378	0.4224	0.3158	0.7699†	4.781	3.726	4.315	4.842	4.438	0.7868	0.7377
seed_mm	Physical	2.940	0.2250	0.3720	0.2722	0.4087	4.652	3.391†	4.348	4.796	4.212	0.7233	0.7436
seed_mm	Complex Scene	2.755	0.2121	0.3354	0.2804	0.4667	4.377	3.885	4.404	4.560	3.984	0.7113	0.7030
seed_mm	Virtual	2.832	0.2450	0.2894	0.2178†	0.6239	4.685	<b>4.207</b>	4.576	2.761†	2.511†	0.7581	<b>0.8124</b>
wan2.2_think	Animals	2.902	<b>0.2542</b>	0.3324	0.2630	0.8171	4.257	3.457	4.371	4.500	3.743	<b>0.7105</b>	<b>0.8322</b>
wan2.2_think	Human Sounds	2.682	0.1963†	0.2227†	0.2074	0.5011	3.922†	3.511	4.172†	4.597	4.089	0.4175†	0.6220
wan2.2_think	Music	<b>4.053</b>	0.2144	<b>0.4134</b>	<b>0.3332</b>	<b>0.2387</b>	<b>4.774</b>	3.839	<b>4.452</b>	4.832	<b>4.677</b>	0.6458	0.5422†
wan2.2_think	Environment	2.524†	0.2287	0.3297	0.1975	0.8603	4.726	3.671	4.315	<b>4.884</b>	4.315	0.6375	0.6974
wan2.2_think	Physical	3.033	0.2133	0.2879	0.2019	0.4533	4.370	3.332†	4.239	4.736	4.005	0.5323	0.6284
wan2.2_think	Complex Scene	2.689	0.2074	0.2730	0.2102	0.6372	4.186	3.858	4.432	4.530	3.973	0.6395	0.5870
wan2.2_think	Virtual	2.749	0.2258	0.2306	0.1707†	0.8870†	4.467	<b>3.967</b>	4.402	3.103†	2.717†	0.6280	0.6872
wan2.2_mm	Animals	3.121	<b>0.2542</b>	0.3152	0.3452	0.7257	4.429	3.486	4.486	4.500	3.800	<b>0.6905</b>	<b>0.8322</b>
wan2.2_mm	Human Sounds	2.848	0.1963†	0.2887†	0.2820	0.4800	4.011†	3.472	4.161†	4.597	4.072	0.4653†	0.6220
wan2.2_mm	Music	<b>3.743</b>	0.2144	<b>0.4505</b>	<b>0.3903</b>	<b>0.2903</b>	<b>4.806</b>	3.839	<b>4.581</b>	4.823	<b>4.677</b>	0.5988	0.5422†
wan2.2_mm	Environment	2.570†	0.2287	0.4167	0.2916	0.7068	4.575	3.712	4.288	<b>4.884</b>	4.411	0.6894	0.6974
wan2.2_mm	Physical	2.948	0.2133	0.3740	0.2651	0.946	4.522	3.288†	4.261	4.736	4.141	0.5794	0.6284
wan2.2_mm	Complex Scene	2.674	0.2074	0.3259	0.2865	0.5770	4.333	3.869	4.410	4.530	3.956	0.6446	0.5870
wan2.2_mm	Virtual	2.733	0.2258	0.2989	0.2008†	0.7804†	4.565	<b>3.891</b>	4.424	3.103†	2.717†	0.5937	0.6872
klings_think	Animals	2.963	<b>0.2555</b>	0.3122	0.2670	0.5600	4.743	3.800	4.486	4.857	4.114	0.8286	<b>0.8828</b>
klings_think	Human Sounds	2.701†	0.2103†	0.2344†	0.2026†	0.5900	4.250†	3.694	4.244†	4.708	4.172	0.4937†	0.7171†
klings_think	Music	<b>4.101</b>	0.2236	<b>0.4374</b>	<b>0.3502</b>	<b>0.2258</b>	<b>4.968</b>	4.032	<b>4.710</b>	<b>4.935</b>	<b>4.871</b>	<b>0.9012</b>	0.7408
klings_think	Environment	2.813	0.2363	0.2911	0.2046	0.6904	4.863	3.781	4.397	4.884	4.384	0.7807	0.7980
klings_think	Physical	3.135	0.2339	0.2730	0.2114	0.3989	4.380	3.636†	4.332	4.812	4.217	0.7035	0.7950
klings_think	Complex Scene	2.672	0.2304	0.2658	0.2239	0.7410	4.322	4.000	4.470	4.508	3.973	0.7519	0.7436
klings_think	Virtual	2.920	0.2505	0.2456	0.2166	0.7652†	4.663	<b>4.250</b>	4.641	2.989†	2.772†	0.7548	0.8662
klings_mm	Animals	3.139	<b>0.2555</b>	0.3446	0.3491	0.6917	4.543	3.800	4.514	4.857	4.057	0.7691	<b>0.8828</b>
klings_mm	Human Sounds	3.073	0.2103†	0.3037	0.2827	0.5300	4.144†	3.661	4.244†	4.708	4.183	0.5203†	0.7171†
klings_mm	Music	<b>4.011</b>	0.2236	<b>0.4937</b>	<b>0.4082</b>	<b>0.2581</b>	<b>4.968</b>	4.032	4.548	<b>4.935</b>	<b>4.903</b>	<b>0.8991</b>	0.7408
klings_mm	Environment	2.582†	0.2362	0.4279	0.2988	0.6575	4.836	3.767	4.370	4.884	4.384	0.8347	0.7980
klings_mm	Physical	3.049	0.2339	0.3766	0.2952	0.4120	4.489	3.658†	4.370	4.812	4.207	0.7910	0.7950
klings_mm	Complex Scene	2.722	0.2304	0.3386	0.2920	0.6033	4.339	4.016	4.475	4.508	4.033	0.7420	0.7436
klings_mm	Virtual	2.859	0.2505	0.3019†	0.2453†	0.8152†	4.587	<b>4.250</b>	<b>4.609</b>	2.989†	2.707†	0.8398	0.8662

Table 6. Results of different models on the I2AV task across seven audio categories, excluding Virtual. Same notation as Tab. 5.

Model	Category	Audio-Aes	T-V Align	T-A Align	A-V Align	Desync $\downarrow$	Alignment	Artistry	Expressiveness	Visual Realism	Audio Realism	Audio QA	Visual QA
sora2	Animals	3.164	<b>0.2321</b>	0.4556	<b>0.2952</b>	0.9277	4.830	3.921	<b>4.525</b>	4.949	<b>4.627</b>	0.8210	0.7708
sora2	Human Sounds	2.536 $\dagger$	0.2233	0.3870	0.1997 $\dagger$	1.140 $\dagger$	4.600 $\dagger$	<b>4.000</b>	4.200 $\dagger$	4.800 $\dagger$	4.500	0.8667	0.8405
sora2	Music	<b>3.880</b>	0.1934 $\dagger$	<b>0.4709</b>	0.2618	<b>0.5180</b>	<b>5.000</b>	<b>4.000</b>	4.420	4.980	<b>4.860</b>	<b>0.9100</b>	<b>0.8474</b>
sora2	Environment	2.642	0.2154	0.3420 $\dagger$	0.2585	1.095	4.888	3.950	4.287	<b>5.000</b>	4.487 $\dagger$	0.7855 $\dagger$	0.7540
sora2	Physical	2.850	0.2142	0.3782	0.2356	0.7346	4.923	3.875 $\dagger$	4.279	4.981	4.548	0.8636	0.7596
sora2	Complex Scene	2.622	0.2150	0.3646	0.2412	1.123	4.910	3.950	4.350	4.950	4.560	0.7960	0.6998 $\dagger$
veo3	Animals	3.573	<b>0.2498</b>	<b>0.4884</b>	<b>0.3665</b>	0.4785	4.830	3.966	<b>4.780</b>	4.927	4.678	0.8904	0.8083
veo3	Human Sounds	3.245 $\dagger$	0.2347	0.3661	0.2689 $\dagger$	0.6200	4.700 $\dagger$	4.000	4.300 $\dagger$	4.700 $\dagger$	4.400 $\dagger$	0.8650	0.8467
veo3	Music	<b>4.651</b>	0.2022 $\dagger$	0.4683	0.3612	<b>0.3660</b>	<b>5.000</b>	4.000	4.680	<b>4.980</b>	<b>4.860</b>	<b>0.9263</b>	<b>0.8696</b>
veo3	Environment	3.461	0.2318	0.3030 $\dagger$	0.2905	0.9025 $\dagger$	4.925	3.987	4.537	4.963	4.638	0.7653 $\dagger$	0.7841
veo3	Physical	3.255	0.2265	0.4131	0.2961	0.4933	4.971	3.923 $\dagger$	4.510	4.952	4.615	0.8709	0.7926
veo3	Complex Scene	3.492	0.2284	0.3448	0.2785	0.8700	4.930	<b>4.060</b>	4.580	4.840	4.620	0.8288	0.7568 $\dagger$
wan2.5	Animals	3.759	<b>0.2515</b>	<b>0.3229</b>	<b>0.2555</b>	0.2768	4.746 $\dagger$	3.904	<b>4.667</b>	4.706	4.378	0.7877	0.7827
wan2.5	Human Sounds	3.120	0.2491	0.2349	0.1400 $\dagger$	<b>0.1400</b>	4.900	3.900 $\dagger$	4.200 $\dagger$	4.500 $\dagger$	4.300	0.6850 $\dagger$	0.8610
wan2.5	Music	<b>3.877</b>	0.2183 $\dagger$	0.2802	0.2053	0.7040 $\dagger$	<b>4.980</b>	<b>4.000</b>	4.420	<b>4.880</b>	<b>4.700</b>	<b>0.8622</b>	<b>0.8873</b>
wan2.5	Environment	3.320	0.2321	0.2278 $\dagger$	0.1717	0.3375	4.763	3.925	4.463	4.825	4.275 $\dagger$	0.7918	0.8005
wan2.5	Physical	3.297	0.2296	0.2922	0.2012	0.3846	4.913	3.913	4.356	4.846	4.423	0.8546	0.7630
wan2.5	Complex Scene	3.013 $\dagger$	0.2332	0.2714	0.1846	0.3180	4.770	3.960	4.430	4.710	4.350	0.7958	0.7609 $\dagger$
seed_think	Animals	2.890	<b>0.2426</b>	<b>0.3789</b>	0.2590	0.6712	4.831	3.932	4.650	4.915	4.508	0.7895	0.7485
seed_think	Human Sounds	2.356 $\dagger$	0.2342	0.2102 $\dagger$	0.1802 $\dagger$	1.100 $\dagger$	4.400 $\dagger$	3.900 $\dagger$	4.100 $\dagger$	4.650 $\dagger$	4.300 $\dagger$	0.7100 $\dagger$	0.8014
seed_think	Music	<b>3.981</b>	0.2035 $\dagger$	0.3691	<b>0.2992</b>	<b>0.2680</b>	4.940	3.980	<b>4.680</b>	<b>4.950</b>	<b>4.780</b>	<b>0.8363</b>	<b>0.8316</b>
seed_think	Environment	2.496	0.2218	0.2271	0.1927	0.7225	4.888	3.938	4.463	4.906	4.375	0.7321	0.7251
seed_think	Physical	2.841	0.2231	0.2935	0.2133	0.7077	<b>4.962</b>	3.904	4.433	4.947	4.433	0.8111	0.7405
seed_think	Complex Scene	2.469	0.2216	0.2954	0.2179	0.8200	4.890	<b>4.060</b>	4.430	4.825	4.490	0.7651	0.7118 $\dagger$
seed_mm	Animals	3.117	<b>0.2426</b>	<b>0.4690</b>	<b>0.3572</b>	0.5401	4.876	3.949	<b>4.734</b>	4.915	4.599	0.8145	0.7485
seed_mm	Human Sounds	2.962	0.2342	0.2940 $\dagger$	0.2690 $\dagger$	0.7800 $\dagger$	4.700 $\dagger$	4.000	4.300 $\dagger$	4.650 $\dagger$	4.200 $\dagger$	0.7833	0.8014
seed_mm	Music	<b>4.013</b>	0.2035 $\dagger$	0.3608	0.3257	<b>0.4880</b>	<b>5.000</b>	3.980	4.660	<b>4.950</b>	<b>4.820</b>	<b>0.8405</b>	<b>0.8316</b>
seed_mm	Environment	2.614 $\dagger$	0.2218	0.3610	0.2951	0.6400	4.938	4.013	4.513	4.906	4.450	0.7789	0.7251
seed_mm	Physical	2.822	0.2231	0.4185	0.2888	0.5173	4.952	3.923 $\dagger$	4.462	4.947	4.413	0.8257	0.7405
seed_mm	Complex Scene	2.648	0.2216	0.3584	0.3007	0.7380	4.920	<b>4.020</b>	4.500	4.825	4.460	0.7561 $\dagger$	0.7118 $\dagger$
wan2.2.think	Animals	2.837	<b>0.2459</b>	<b>0.3884</b>	0.2410	0.8102	4.808	3.949	<b>4.638</b>	4.895	4.435	0.7696	0.7440
wan2.2.think	Human Sounds	2.641	0.2413	0.1911 $\dagger$	0.1693 $\dagger$	1.200 $\dagger$	4.400 $\dagger$	3.900	4.000 $\dagger$	4.650 $\dagger$	4.100 $\dagger$	0.6667	<b>0.8310</b>
wan2.2.think	Music	<b>4.010</b>	0.2008 $\dagger$	0.3629	<b>0.2757</b>	<b>0.2960</b>	<b>4.980</b>	<b>3.980</b>	4.620	4.940	<b>4.680</b>	<b>0.8091</b>	0.7977
wan2.2.think	Environment	2.440 $\dagger$	0.2226	0.2377	0.1764	0.9250	4.850	3.938	4.400	<b>4.944</b>	4.250	0.6289 $\dagger$	0.7248
wan2.2.think	Physical	2.946	0.2226	0.3006	0.2082	0.6654	4.913	3.894 $\dagger$	4.394	4.942	4.337	0.7599	0.7062 $\dagger$
wan2.2.think	Complex Scene	2.452	0.2245	0.2846	0.1785	1.022	4.820	3.970	4.450	4.900	4.370	0.6941	0.7097
wan2.2.mm	Animals	3.015	<b>0.2459</b>	<b>0.4739</b>	<b>0.3462</b>	0.7028	4.831 $\dagger$	3.932	<b>4.678</b>	4.895	4.458	0.7600	0.7440
wan2.2.mm	Human Sounds	2.899	0.2413	0.3111 $\dagger$	0.2729	0.8000	4.900	3.900 $\dagger$	4.300 $\dagger$	4.650 $\dagger$	4.200 $\dagger$	<b>0.7967</b>	<b>0.8310</b>
wan2.2.mm	Music	<b>3.995</b>	0.2008 $\dagger$	0.3552	0.3377	<b>0.4160</b>	<b>4.960</b>	3.980	4.580	4.940	<b>4.660</b>	0.7842	0.7977
wan2.2.mm	Environment	2.550 $\dagger$	0.2226	0.3759	0.2676 $\dagger$	0.7200	4.938	3.975	4.425	<b>4.944</b>	4.400	0.7583	0.7248
wan2.2.mm	Physical	2.919	0.2226	0.4156	0.2797	0.6808	4.952	3.952	4.375	4.942	4.317	0.7476	0.7062 $\dagger$
wan2.2.mm	Complex Scene	2.582	0.2245	0.3665	0.2766	0.9480 $\dagger$	4.900	<b>4.010</b>	4.440	4.900	4.380	0.7339 $\dagger$	0.7097
kling_think	Animals	3.009	<b>0.2413</b>	<b>0.3867</b>	0.2707	0.7311	4.791	3.915	<b>4.627</b>	4.904	4.367	0.7776	0.7503
kling_think	Human Sounds	2.431	0.2346	0.2206	0.1351 $\dagger$	1.120 $\dagger$	4.700 $\dagger$	4.000	4.100 $\dagger$	4.650 $\dagger$	4.400	0.6933	0.8240
kling_think	Music	<b>3.980</b>	0.2021 $\dagger$	0.3785	<b>0.3128</b>	<b>0.3040</b>	<b>5.000</b>	<b>4.020</b>	4.580	4.910	<b>4.680</b>	<b>0.8093</b>	<b>0.8313</b>
kling_think	Environment	2.439	0.2234	0.2147 $\dagger$	0.1661	0.8150	4.862	3.987	4.425	<b>4.981</b>	4.225 $\dagger$	0.6425 $\dagger$	0.7213
kling_think	Physical	2.952	0.2215	0.3020	0.2241	0.6019	4.942	3.894 $\dagger$	4.413	4.870	4.356	0.7675	0.7287
kling_think	Complex Scene	2.433 $\dagger$	0.2216	0.2694	0.1884	0.8640	4.840	4.000	4.460	4.840	4.340	0.7058	0.7039 $\dagger$
kling_mm	Animals	3.054	<b>0.2413</b>	<b>0.4660</b>	<b>0.3636</b>	0.6079	4.825	3.932	<b>4.706</b>	4.904	4.503	0.8113	0.7503
kling_mm	Human Sounds	3.196	0.2346	0.3308 $\dagger$	0.2959	1.120 $\dagger$	4.800 $\dagger$	4.000	4.500	4.650 $\dagger$	4.500	0.7567	0.8240
kling_mm	Music	<b>3.867</b>	0.2021 $\dagger$	0.3701	0.3462	<b>0.3400</b>	<b>4.980</b>	3.980	4.520	4.910	<b>4.660</b>	<b>0.8171</b>	<b>0.8313</b>
kling_mm	Environment	2.547 $\dagger$	0.2234	0.3628	0.2719 $\dagger$	0.8325	4.925	3.975	4.412	<b>4.981</b>	4.275 $\dagger$	0.7610	0.7213
kling_mm	Physical	2.895	0.2215	0.4035	0.2739	0.6077	4.913	3.923 $\dagger$	4.375 $\dagger$	4.870	4.365	0.8136	0.7287
kling_mm	Complex Scene	2.651	0.2216	0.3582	0.2809	0.8720	4.860	<b>4.010</b>	4.440	4.840	4.340	0.7499 $\dagger$	0.7039 $\dagger$