Can Omni-models Imagine the Future? A Diagnostic Benchmark for Multimodal World Modeling

Junyeob Kim[1], Jihwan Kim[2], Minha Jhang[1], Sungwook Jeon[3], Hwiyeol Jo[3], Jeonghoon Kim[2,3], Sang-goo Lee[1], Joosung Lee[3]

[1]Seoul National University [2]KAIST AI [3]NAVER Cloud

{juny116,sglee}@europa.snu.ac.kr jihvvan.kim@kaist.ac.kr jminha2014@snu.ac.kr

{lucas.jeon,hwiyeol.jo,jeonghoon.samuel,rung.joo}@navercorp.com

Abstract

Recent progress in Omni-models has yielded strong performance on benchmarks focused on static observations. However, their ability to maintain spatiotemporal continuity when predicting coherent future states from previous context remains underexplored. We address this by formulating Next Scene Prediction (NSP) as a diagnostic task for assessing multimodal near-future state prediction, requiring models to anticipate the immediate future from a given context. We introduce OmniNSP, a benchmark evaluating causal and cross-modal coherence across physical, logical, and social dynamics. Unlike prior tasks that often rely on textual or discriminative proxies, OmniNSP asks models to express predicted future states through text-mediated visual and speech synthesis. Our analysis shows that current Omni-models can produce plausible textual forecasts while still failing to preserve future-state consistency in generated outputs. A frozen-generator control substantially narrows cross-model visual-consistency gaps, attributing much of the original system-level difference to model-specific realization stacks rather than textual foresight alone. A modality-input ablation further shows that audio provides complementary predictive cues beyond vision, with the largest gains for logical transitions and contexts with informative speech. These results position OmniNSP as a diagnostic of both near-future prediction and the preservation of predicted state through multimodal realization.

1 Introduction

Recent years have witnessed steady progress in Large Multimodal Models (LMMs), with the emergence of Omni-models (Xu et al., 2025; Li et al., 2025; Comanici et al., 2025) capable of processing and generating text, images, and audio. Beyond cross-modal alignment, a central goal for such systems is to support world-model-like reasoning:

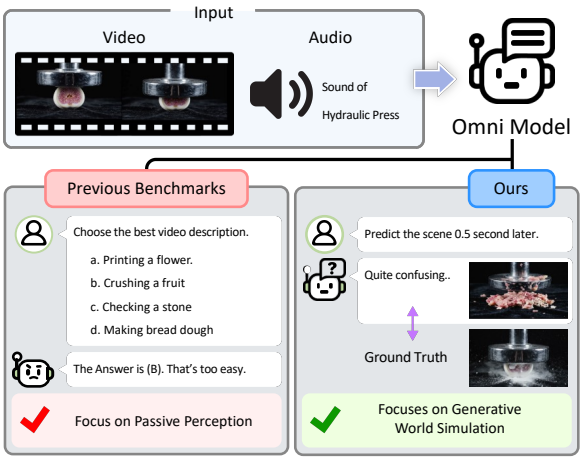


Figure 1: Rethinking evaluation of Omni-models. Existing Omni-models are typically evaluated on static benchmarks focusing on passive perception, which provides limited evidence about their ability to anticipate future states. We propose Next Scene Prediction as a diagnostic benchmark for probing causal and temporal continuity.

maintaining a representation of how scenes evolve under physical, logical, and social constraints (Ha and Schmidhuber, 2018; LeCun et al., 2022). This capability is important for embodied assistants, interactive agents, and safety-critical systems that must anticipate near-future outcomes rather than only describe the current observation (Hafner et al., 2019; Brooks et al., 2024).

However, it remains unclear to what extent current Omni-models can maintain such dynamics when asked to predict what happens next. Existing evaluation protocols (Zhou et al., 2025; Li et al., 2026a,b) predominantly focus on passive perception, where models interpret static or already-observed content. These benchmarks offer limited visibility into prospective reasoning: a model may accurately describe a scene while still failing to anticipate the next physical state, the next step in a symbolic process, or the likely continuation of a

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 47099–47115
October 24-29, 2026 ©2026 Association for Computational Linguistics

social interaction.

To address this, we propose Next Scene Prediction (NSP) as a task formulation for evaluating multimodal near-future state prediction (Figure 1). By shifting the focus from current observation to future continuation, NSP probes whether a model can infer the latent factors that drive a scene's continuation. Under this framework, we introduce OmniNSP, a curated diagnostic benchmark constructed via a semi-automated, human-in-the-loop pipeline to evaluate future-state prediction across physical, logical, and social dynamics. Distinguishing itself from forecasting benchmarks that rely primarily on textual proxies such as MCQA (Chen et al., 2026), OmniNSP requires models to express predicted future states through generated visual and speech outputs in addition to text.

Using OmniNSP, we systematically evaluate current Omni-models under a practical *multi-stage generation protocol*: the model first predicts a textual future state, and that prediction is then used to condition visual and speech synthesis. This protocol is not the ideal end-to-end realization of NSP; rather, it is a controlled instantiation that reflects current model capabilities and exposes intermediate predictions alongside downstream realization. Because the final output remains affected by each model's generation stack, we additionally use a frozen-generator control to distinguish system-level realization effects from textual foresight. Our results show that plausible textual predictions do not always survive model-specific visual and speech realization.

In summary, the contributions of this work are as follows: (i) we formalize NSP as a diagnostic framework for evaluating the transition from passive perception to prospective multimodal prediction; (ii) we introduce OmniNSP, a curated benchmark spanning three operational domains of future-state dynamics; and (iii) we provide an empirical analysis of how textual future-state predictions are preserved or degraded during multimodal realization, including a controlled generator analysis that localizes much of the visual gap to model-specific synthesis stacks.

2 Next Scene Prediction

The NSP task frames world modeling as conditional generation over temporal state transitions: rather than interpreting observed content, a model must anticipate how a scene will progress beyond the provided context. Because the relevant latent state is not directly observable, its continuation must be inferred through mutually consistent visual, auditory, and linguistic evidence. NSP therefore evaluates whether a predicted future scene jointly forms a plausible continuation of the environment's dynamics, rather than a collection of independently reasonable modality outputs.

2.1 Formalization

At time t, we define a multimodal context segment as $C = \{V_{:t}, A_{:t}, I\}$, where $V_{:t}$ denotes the observed visual sequence, $A_{:t}$ represents the synchronized audio stream, and I denotes the textual instruction or task-specific prompt. The objective of the NSP task is for a model $\mathcal{M}$ to generate a subsequent target segment T corresponding to the scene after a temporal offset Δ:

$$T = \{v_{t+\Delta}, a_{t+\Delta}\},$$

where Δ represents a variable temporal interval determined by the duration of the underlying event. Here, $v_{t+\Delta}$ is a representative visual state (e.g., a keyframe), and $a_{t+\Delta}$ is the corresponding auditory output.

An ideal world model approximates the underlying transition distribution $P(T \mid C)$, ensuring the predicted $\hat{T}$ aligns with the ground-truth future state T_{gt} across modalities. In our current evaluation, we additionally elicit a linguistic prediction $x_{t+\Delta}$ as an auxiliary intermediate representation. This text output is not a requirement of NSP itself; rather, it makes the evaluation comparable across present Omni-models whose end-to-end future video-and-audio generation capabilities remain limited.

2.2 Practical Instantiation

While the most direct realization of this task would involve an end-to-end generative process (e.g., video-to-video), current architectural constraints often limit the stability and comparability of such direct mappings across different models. To enable a rigorous and scalable evaluation, we instantiate NSP through a decoupled synthesis approach.

In this setting, the linguistic component $x_{t+\Delta}$ functions as a bridge modality; the model first describes the predicted future state in text, which then serves as a conditioning signal for the generation of visual and auditory outputs. This multi-stage realization preserves the core objective of future-

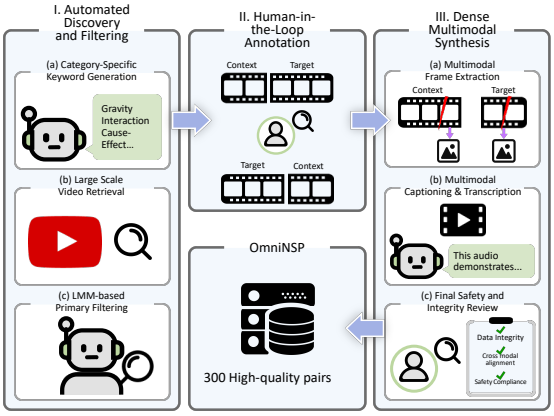


Figure 2: Semi-automated curation pipeline for the OmniNSP dataset. The workflow integrates automated retrieval and filtering (Phase I) with human-expert temporal grounding (Phase II) and dense multimodal synthesis (Phase III) to ensure both scalability and physical authenticity in causal scene transitions.

state prediction while addressing the practical limitations of contemporary multimodal generative systems. It makes stage-wise failures observable, while a shared frozen generator enables targeted attribution of downstream realization effects.

3 OmniNSP

3.1 Data Collection and Curation

To capture diverse future-state dynamics, we developed a semi-automated three-phase pipeline that combines scalable retrieval with human temporal grounding and quality control. Throughout the pipeline, whenever LMM-based processing was required, we used Gemini-3-Flash (Google DeepMind, 2026).

Phase I: Automated Discovery and Filtering. We first generated category-specific search phrases (e.g., "glass breaking in slow motion," "traffic signal change," and "person reacting to surprise") for physical, logical, and social dynamics using an LMM. These phrases describe concrete observable events likely to contain valid future-state transitions, such as collisions, deformation, optical effects, symbolic transformations, or social interactions. We used the YouTube Data API to retrieve candidate videos and then applied LMM-based filtering to identify clips that contained a visually and/or acoustically observable transition. This automated stage was used only to propose candidates; it did not determine the final benchmark instances.

Phase II: Human-in-the-Loop Temporal Grounding. Human annotators manually verified whether each candidate was usable as an NSP instance. A valid instance must contain a non-trivial transition whose future state cannot be read off from a single static frame alone, have a clear context segment preceding the transition, and have a target segment that captures the continuation of the same event. Annotators inspected and corrected the automatically proposed timestamps, selecting a context segment that provides sufficient evidence for prediction and a target segment that represents the immediate future. We preserve these segment-level timestamps even though the current evaluation uses representative frames, so that future versions can support direct video-audio evaluation.

Phase III: Multimodal Annotation and Review. For the current instantiation, we extract the final frame of the context as the visual reference and a representative mid-frame from the target segment as the ground-truth future state ($v_{t+\Delta}$). We generate dense visual captions with Gemini and speech transcriptions with Whisper (Radford et al., 2023). Annotators then review caption quality, transcript quality, context-target alignment, and safety. Instances are removed if the transition is ambiguous, the target does not follow from the context, the media quality is insufficient for evaluation, or the content raises privacy or safety concerns.

The final dataset size reflects this high-precision filtering process. Many retrieved videos are topically relevant but unusable because the transition is unclear, the automatically suggested timestamps are misaligned, or the audio-visual evidence is insufficiently consistent. We therefore frame OmniNSP as a curated diagnostic benchmark rather than an exhaustive large-scale benchmark. The complete annotation protocol is detailed in Appendix A.

3.2 Taxonomy of World Dynamics

We categorize OmniNSP into three diagnostic pillars of world modeling: *Physical World Modeling*, *Logical & Structural Reasoning*, and *Social Intelligence*. We do not claim that these pillars exhaustively define world modeling. Rather, OmniNSP uses them as a practical taxonomy for evaluating observable future-state transitions in multimodal contexts (LeCun et al., 2022).

Categories were included if they were grounded in prior work, produced observable future-state changes in visual or audio channels, and could be annotated reliably at benchmark scale. Ambiguous examples are assigned according to the primary

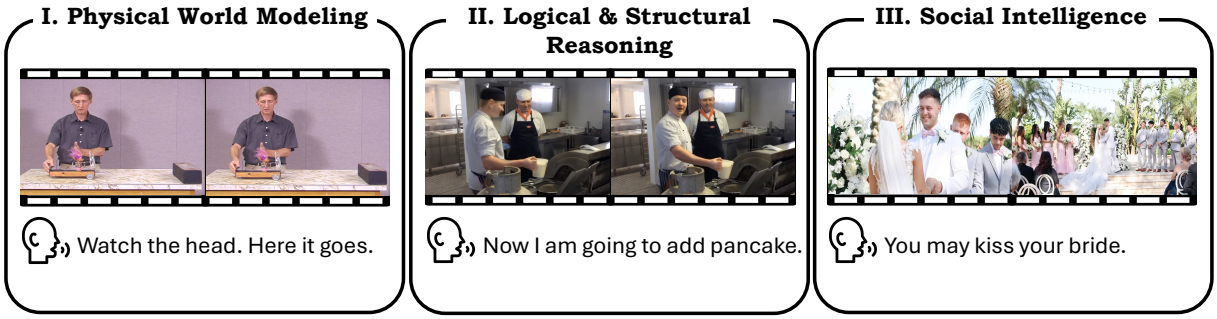


Figure 3: The three diagnostic pillars of world modeling in OmniNSP. (I) *Physical World Modeling* evaluates continuous dynamics such as intuitive physics and environmental effects. (II) *Logical & Structural Reasoning* assesses discrete transformations based on causal, symbolic, or mechanistic logic. (III) *Social Intelligence* tests agent dynamics and adherence to social norms. NSP uses these axes to diagnose observable future-state prediction in multimodal context.

Pillar	Sub-domains	n
Physical	Rigid Body; Deformable Body; Environmental Effects; Thermodynamics & Optics	100
Logical	Causal Reasoning; Mathematical / Symbolic Reasoning	100
Social	Emotion & Intent Perception; Social Norms & Cognition	100
Total		**300**

Table 1: Distribution of OmniNSP. Physical sub-domains contain 25 samples each; logical and social sub-domains contain 50 samples each.

causal driver behind the target transition, while categories that were too rare, subjective, or difficult to observe in short clips were merged or excluded.

Physical World Modeling covers continuous state changes governed by intuitive physical regularities (Spelke and Kinzler, 2007; Battaglia et al., 2013; Smith et al., 2019), including Rigid Body Mechanics, Deformable Body Mechanics, Environmental Effects, and Thermodynamics & Optics. **Logical and Structural Reasoning** covers future states determined by discrete causal or symbolic rules (Pearl, 2009), including Causal Reasoning and Mathematical / Symbolic Reasoning. **Social Intelligence** covers agent-centered continuations involving affect, intention, or conventions (Frith and Frith, 2005; Bicchieri, 2006), including Emotion & Intent Perception and Social Norms & Cognition.

3.3 Statistics and Diversity

The OmniNSP benchmark consists of 300 human-verified multimodal samples, stratified by category to support comparison across different future-state dynamics (Table 1). Each sample contains context and target video timestamps, a context reference frame, a target future frame, the original audio segment where available, speech transcription, and dense visual captions. This design prioritizes temporal grounding and annotation reliability over scale.

4 Evaluation Framework and Experimental Setup

4.1 Multi-stage Generation Protocol

The ideal NSP evaluation would ask a model to generate a future audio-visual segment directly from the observed multimodal context. Current Omni-models, however, do not uniformly support stable and comparable end-to-end future video-and-audio generation. We therefore instantiate NSP with a multi-stage protocol that records textual future-state prediction separately from visual and speech synthesis, while evaluating the resulting pipeline as an integrated system.

Stage 1: Future-State Reasoning. Given the context segment C and a fixed prompt, the model predicts the scene at a temporal offset Δ after the context ends. We require a structured JSON output with two fields: (i) a *Scene Description* describing the expected entities, actions, and state changes, and (ii) an *Audio Script* describing the expected spoken content when speech is relevant. This stage evaluates semantic and causal foresight before any downstream synthesis step.

Stage 2: Future-State Synthesis. The predicted textual state is then used as a conditioning signal for generation. For *Visual Generation*, the model synthesizes a single future image using the predicted scene description and the last frame of the

Component	Evaluation role
Scene description	LMM-judge score for causal plausibility and alignment with the target future state.
Generated image	LMM-judge scores for fidelity, observed-scene consistency, and text-image alignment.
Generated speech	Automatic diagnostics for script preservation, speech quality, and speaker similarity.

Table 2: Evaluation components in the multi-stage NSP protocol. LMM-judge scores are the primary future-state evaluation; speech metrics are auxiliary synthesis diagnostics.

context as reference. For *Speech Generation*, the model synthesizes the predicted audio script using its native text-to-speech capability. If a model supports zero-shot voice cloning, we provide a short context-audio reference to preserve speaker identity; otherwise, the system uses its default or neutral voice.

Scope of Audio Evaluation. The dataset preserves full audio where available, including speech and non-speech sounds. The current evaluation focuses on speech-centered synthesis because the Omni-model baselines evaluated in this study do not provide stable and comparable generation of arbitrary environmental sound effects. We therefore treat non-speech audio as part of the context and annotation, while speech synthesis is evaluated as the comparable audio-generation component. Audio-script plausibility and speech-synthesis diagnostics are computed only for target segments with evaluable speech, whereas the modality ablation partitions all 300 examples by the informativeness of the input context audio.

4.2 Evaluation Metrics and Protocol

We employ a hybrid framework that separates primary NSP evaluation from auxiliary speech-synthesis diagnostics. Gemini-3-Pro is used as the LMM judge for the original rubric-based 0–2 scores of textual future-state reasoning and visual synthesis. Automated speech metrics are computed separately and are not provided as inputs to the LMM judge.

Textual Future-State Metrics. Prospective reasoning is assessed through: (i) *Causal Coherence*, which verifies whether the predicted scene logically continues the context using the ground-truth target frame as a visual anchor; and (ii) *Audio Event Plausibility*, which measures whether the predicted speech content is situationally plausible with respect to the context and target audio annotations. These scores evaluate future-state plausibility rather than downstream speech quality.

Visual Synthesis Metrics. Synthesized images are evaluated via: (i) *Generative Fidelity*, measuring perceptual quality and artifact severity; (ii) *Observed Scene Consistency*, measuring preservation of identities, layout, and environmental tone from the context; and (iii) *Text–Image Semantic Alignment*, measuring whether the generated image instantiates the predicted scene description.

Speech Synthesis Diagnostics. Quantitative analysis of synthesized speech includes: (i) *Word Error Rate (WER)* computed with Whisper-large-v3 (Radford et al., 2023), measuring whether the generated speech preserves the predicted script; (ii) *UTMOS* (Saeki et al., 2022), estimating speech naturalness on a MOS-like 1–5 scale where higher is better; and (iii) *Speaker Similarity*, measured by MFCC-based cosine similarity between the context audio and output speech. These metrics diagnose the speech realization stage and are reported separately from the LMM-judge scores.

4.3 Baseline Models and Implementation

We evaluate four Omni-modal architectures: (i) Uni-MoE-2.0-Omni (Li et al., 2025), (ii) Ming-flash-omni-Preview (AI et al., 2025), (iii) Qwen-3-Omni (Xu et al., 2025), and (iv) Gemini-2.5-flash (Comanici et al., 2025). We selected models that were accessible at evaluation time and could process multimodal context, produce a structured future-state prediction, and support visual and speech realization under a unified zero-shot protocol; this set is representative rather than exhaustive. Uni-MoE and Ming are standalone models, while Qwen-3 and Gemini operate as composite pipelines. Specifically, Qwen-3 pairs its internal audio with Qwen-3-image-edit, while Gemini integrates its reasoning engine with Nano Banana and a TTS suite.

Implementation Details. All models used unified zero-shot prompts and the same structured output fields whenever supported. Uni-MoE, Ming, and Qwen-3 ran on four NVIDIA A100 GPUs, while Gemini used its official API. We used single-step textual reasoning when available; Uni-MoE and Ming used separate description and script calls to preserve the common output schema. Voice cloning was enabled for Uni-MoE and Ming when supported, and Gemini-3-Pro judged outputs with

fixed prompts, rubrics, and JSON fields. Appendix F provides a representative prompt and the model-specific prompt configuration.

5 Results and Analysis

5.1 Overall Performance and Multi-modal Consistency

Benchmarking results (Table 3) reveal a distinct performance hierarchy, specifically in the bridging of linguistic reasoning and generative consistency.

Visual Consistency and the Generation Bottleneck. Gemini-2.5-flash obtains the strongest overall profile, especially for preserving observed-scene consistency and text–image alignment. Qwen-3 is closer on low-level image fidelity, suggesting that the main separation lies less in perceptual quality than in temporally grounded scene consistency. Conversely, Uni-MoE-2.0 shows reasonable textual foresight but weak visual grounding. This disparity highlights a generation bottleneck, where conceptual understanding fails to ground effectively into physical rendering.

Speech Synthesis: Identity vs Quality. Speech profiles vary by strategy. Gemini-2.5-flash maintains the best linguistic accuracy and strong speech quality, while Qwen-3-Omni is competitive in perceived naturalness but occasionally exhibits "linguistic interference" in prosody. Voice-cloning models preserve speaker identity more strongly, but this often incorporates background noise and reduces overall naturalness.

Statistical Robustness. Paired bootstrap tests show that Gemini significantly outperforms Ming-flash-omni and Uni-MoE-2.0 across the main scene and image metrics. Against Qwen-3, Gemini's advantages remain significant for scene description, observed-scene consistency, and text–image alignment, while generative fidelity is statistically comparable. Across 1,000 bootstrap resamples, the full consistency and alignment rankings are retained in 100% of resamples, remain unchanged under 50% subsampling, and are invariant when any one pillar is removed. All six model pairs differ in consistency under paired sign-flip tests ($p \leq .001$). The scene-score ordering is less stable (86.7% full-rank retention), although Gemini remains top-ranked in every resample.

Temporal-Offset Sensitivity. The target offset is concentrated on immediate continuations (median 2.0 seconds, IQR 1–3; 75% are at most 3 seconds). After controlling for pillar and model, $\log \Delta$ has only small, non-significant associations with scene, consistency, and alignment ($\beta = -0.022/-0.045/-0.047$, $p = .40/.07/.09$). The model ranking is unchanged for short (≤ 1 s), medium (1–3 s), and longer (> 3 s) offsets on all three dimensions.

5.2 Attributing the Prediction-to-Realization Gap

Stage-wise scores first establish that text and image evaluation capture different behavior. Pooled across models, the sample-level correlation of scene score with consistency is weak (Spearman $\rho = 0.08$, $p = .006$); correlations with alignment and fidelity are likewise small ($\rho = 0.12$, $p < .001$; $\rho = -0.04$, n.s.). Even among predictions receiving the maximum scene score, consistency is at most 1 for 87.8% of Uni-MoE and 85.4% of Ming outputs, compared with 34.6% for Qwen and 8.6% for Gemini. Complete consistency failure (score 0) occurs in 57.7%, 31.7%, 6.3%, and 0% of these cases, respectively. These are system-level degradation rates, however, and do not alone identify whether the text representation or the downstream generator causes the loss.

We therefore pass every model's original Stage-1 prediction and the same context frame to a *frozen* generator, Gemini-2.5-flash-image, which is also Gemini's original Stage-2 generator. We re-score both each model's original image and its frozen-generator image under the same rubric. Table 4 shows that the consistency gaps relative to Gemini shrink from 0.27–0.64 under the original stacks to 0.02–0.08 with the generator fixed. Gemini is unchanged ($\Delta = 0.00$), providing an internal control, whereas Qwen, Uni-MoE, and Ming improve by 0.25, 0.55, and 0.56. Within the frozen condition, only Gemini–Ming differs significantly in consistency ($p = .026$); Gemini–Qwen does not ($p = .58$).

Because Gemini-3-Pro was no longer available, both conditions in this control are scored with Gemini-3.1-Pro-preview. Overall, much of the visual gap is attributable to model-specific realization stacks, while textual-foresight differences remain visible in Stage 1. Full dimensions are in Appendix C, and prompt configuration is in Appendix F.

By holding the Stage-1 description and context frame fixed, the shared-generator design provides a controlled estimate of downstream realization effects. The gains for all three non-Gemini sys-

Model	Reasoning	Synthesis (Image)			Acoustic (Audio)		
	Scene	*Fid.*	*Consist.*	*Align.*	*WER*↓	*UTMOS*↑	*Sim.*↑
Uni-MoE-2.0	$1.43_{[1.35,1.51]}$	$1.05_{[0.97,1.14]}$	$0.55_{[0.46,0.63]}$	$0.83_{[0.74,0.92]}$	35.00	2.72	**0.981**
Ming-flash-omni	$1.29_{[1.21,1.37]}$	$1.55_{[1.48,1.63]}$	$0.78_{[0.70,0.85]}$	$1.36_{[1.27,1.45]}$	39.79	3.31	0.961
w/ VC	–	–	–	–	42.00	2.78	**0.981**
Qwen-3	$1.33_{[1.25,1.40]}$	$1.61_{[1.54,1.68]}$	$1.59_{[1.51,1.65]}$	$1.60_{[1.53,1.68]}$	20.21	**3.50**	0.965
Gemini-2.5-flash	$\mathbf{1.60}_{[1.54,1.66]}$	$\mathbf{1.68}_{[1.63,1.74]}$	$\mathbf{1.88}_{[1.84,1.92]}$	$\mathbf{1.78}_{[1.73,1.84]}$	**11.69**	**3.50**	0.968

Table 3: Main results of the OmniNSP benchmark across four Omni-modal baselines. Brackets indicate 95% bootstrap confidence intervals. *Scene* denotes scene-description causal coherence; *Consist.* is evaluated against ground-truth future frames. *Sim.* indicates speaker similarity, and VC denotes Voice Cloning. Audio-script plausibility is treated as an auxiliary diagnostic and reported in Table 9.

Model	Original	Frozen	Δ
Gemini-2.5-flash	1.12	1.12	+0.00
Qwen-3	0.85	1.10	$+0.25^{*}$
Uni-MoE-2.0	0.53	1.07	$+0.55^{*}$
Ming-flash-omni	0.47	1.04	$+0.56^{*}$

Table 4: Observed-scene consistency under each model's original realization stack and a shared frozen generator. Both conditions are scored with Gemini-3.1-Pro-preview. *The paired-bootstrap 95% CI for the change excludes zero.

Model	Mech.		Logic		Social	
	Text	Img	Text	Img	Text	Img
Uni-MoE-2.0	1.29	0.92	1.49	0.55	1.51	0.95
Ming-flash-omni	1.16	1.44	1.34	0.96	1.36	1.31
Qwen-3	1.20	1.61	1.39	1.42	1.39	1.76
Gemini-2.5-flash	**1.57**	**1.82**	**1.64**	**1.69**	**1.60**	**1.85**

Table 5: Performance breakdown across the three cognitive pillars (0–2 scale). *Text* denotes Scene Description; *Img* is the mean of Fidelity, Consistency, and Alignment. Full 95% bootstrap confidence intervals are reported in Appendix B.

tems, together with Gemini's null change, support a substantial generation bottleneck rather than only weaker forecasts. Across the full dimensions, fidelity converges, whereas alignment changes less uniformly because it depends on forecast content and rendering.

5.3 Domain-Specific Insight: The Three Pillars

Table 5 illustrates model performance across the three pillars: *Physical*, *Logical*, and *Social*. The results reveal distinct behavioral patterns in how reasoning translates to visual instantiation.

Physical: Constraints on Continuity. Predicting complex trajectories remains a universal bottleneck. Gemini-2.5-flash demonstrates superior robustness in maintaining physical states, whereas Uni-MoE-2.0 struggles significantly with visual grounding. At the system level, Gemini's evaluated pipeline better preserves spatiotemporal continuity under rigid physical constraints.

Logical: The Grounding Gap. The logical pillar exhibits a severe grounding gap. While Uni-MoE-2.0 and Ming-flash-omni show competitive textual reasoning, their visual synthesis scores drop sharply. This is most evident in symbolic tasks (e.g., mathematical notation), where these models often produce blank or distorted frames. This highlights a critical disconnect between linguistic foresight and the spatial instantiation of symbolic logic.

Social: Priors and Consistency. Both Gemini-2.5-flash and Qwen-3-Omni are strongest in the social domain. This may reflect the predictable, human-centric patterns of social interaction, which align well with the models' internal priors. Qwen-3-Omni, in particular, effectively preserves character identity and environmental context, suggesting that common-sense priors facilitate higher generative consistency compared to the rigid requirements of mechanics or logic.

5.4 Qualitative Analysis

Representative qualitative comparisons are shown in Figure 4.

Future Scene Synthesis and Stylistic Integrity. The examples highlight clear differences in *Observed Consistency*. Gemini-2.5-flash best preserves tone, lighting, and scene identity, while Qwen-3-Omni sometimes drifts in atmospheric context despite strong perceptual quality. Uni-MoE-2.0 and Ming-flash-omni often keep salient

Group	n	Full	Video -only	Audio -only	Δ F−V
All	300	1.60	1.52	0.61	0.08
Info. audio	151	1.65	1.52	0.82	**0.13**
Uninfo./empty	149	1.56	1.52	0.39	0.03

Table 6: Gemini-2.5-flash modality ablation by context-audio informativeness. Δ denotes Full − Video-only scene-description score. Full confidence intervals and the informative speech/non-speech split are reported in Appendix D.

Dimension	Item ρ	Rank τ	Pairwise
Consistency	0.85	0.91	95.0%
Text-Image Align.	0.64	1.00	91.7%
Generative Fidelity	0.65	1.00	90.7%
Scene Description	0.42	0.33	80.6%

Table 7: Human–LMM agreement. Item ρ is Spearman correlation over 60 model–sample items; rank τ compares the four model means; Pairwise is same-sample model-preference agreement, excluding ties.

objects but reset background structure, indicating weaker temporal world-state tracking.

Justifying Quantitative Trends. These examples explain the trends in Table 3. Gemini's descriptions specify local state changes and its images tend to preserve them. Uni-MoE-2.0 can produce globally plausible scenes, which helps explain its non-trivial text-reasoning score, but its images often lack fine detail and weakly preserve the observed context. We also observe a brevity bias: short predictions reduce explicit factual errors but omit details needed for faithful synthesis, whereas longer descriptions can still fail if generation misses object identity, scene continuity, or material change. Together with the frozen-generator control, these cases indicate degradation under the complete text-mediated realization stack rather than an intrinsic limitation of text or a pure failure of future-state reasoning. This motivates future NSP systems that preserve richer visual or latent state information alongside text and strengthen downstream state-preserving synthesis.

5.5 Modality Dependence Diagnostic

We run a Gemini-2.5-flash input-modality ablation because it is the strongest baseline, making modality sensitivity less likely to be hidden by low overall performance. Table 6 shows that removing audio matters most when the context contains informative audio, but has little measurable effect when audio is uninformative or empty. This supports the intended design of OmniNSP: the benchmark does not reward audio indiscriminately, but tests whether models use a modality when it carries predictive information. OmniNSP therefore covers both genuinely multimodal and visually dominant prediction regimes.

5.6 Human-LMM Evaluation Alignment

We compare Gemini-3-Pro judgments with ratings from three human experts on 60 model–sample items. Table 7 reports agreement between the LMM judge and human experts. On the three visual dimensions, the judge reproduces the human model ordering ($\tau = 0.91$–1.00) and agrees with 90.7–95.0% of decided pairwise preferences; bootstrap lower bounds are 79.4–90.3%. Scene-description pairwise agreement is also 80.6% [57.1,96.4]. Overall, these results support the LMM judge as a reliable and scalable evaluator, with particularly strong validation on visual dimensions. Detailed analyses are in Appendix E.

Audio-script plausibility is less constrained than visual future-state consistency: several utterances can be reasonable continuations of the same context. Reference-based scoring therefore yields lower absolute scores when a script is pragmatically plausible but lexically different from the target, and humans were slightly stricter than the LMM judge on this dimension. We use audio-script scores as an auxiliary diagnostic, while treating scene and image metrics as the primary NSP correctness measures.

6 Related Works

Omni-models and any-to-any generation. Recent LMMs have evolved from static text-image systems toward Omni-models that process and generate interleaved text, image, audio, and video. Any-to-any frameworks established unified modality generation (Wu et al., 2024; Tang et al., 2023, 2024), and recent systems now support broader streaming or multimodal outputs (Li et al., 2025; Luo et al., 2026; AI et al., 2025; Xu et al., 2025; Team, 2026). Yet their ability to maintain a coherent evolving world state across generated modalities remains underexplored.

Multimodal benchmarks and future reasoning. Existing multimodal benchmarks mainly evaluate static perception, temporal recognition, or holistic description (Li et al., 2024; Yue et al., 2024; Cai et al., 2024; Li et al., 2026a,b). Future-oriented

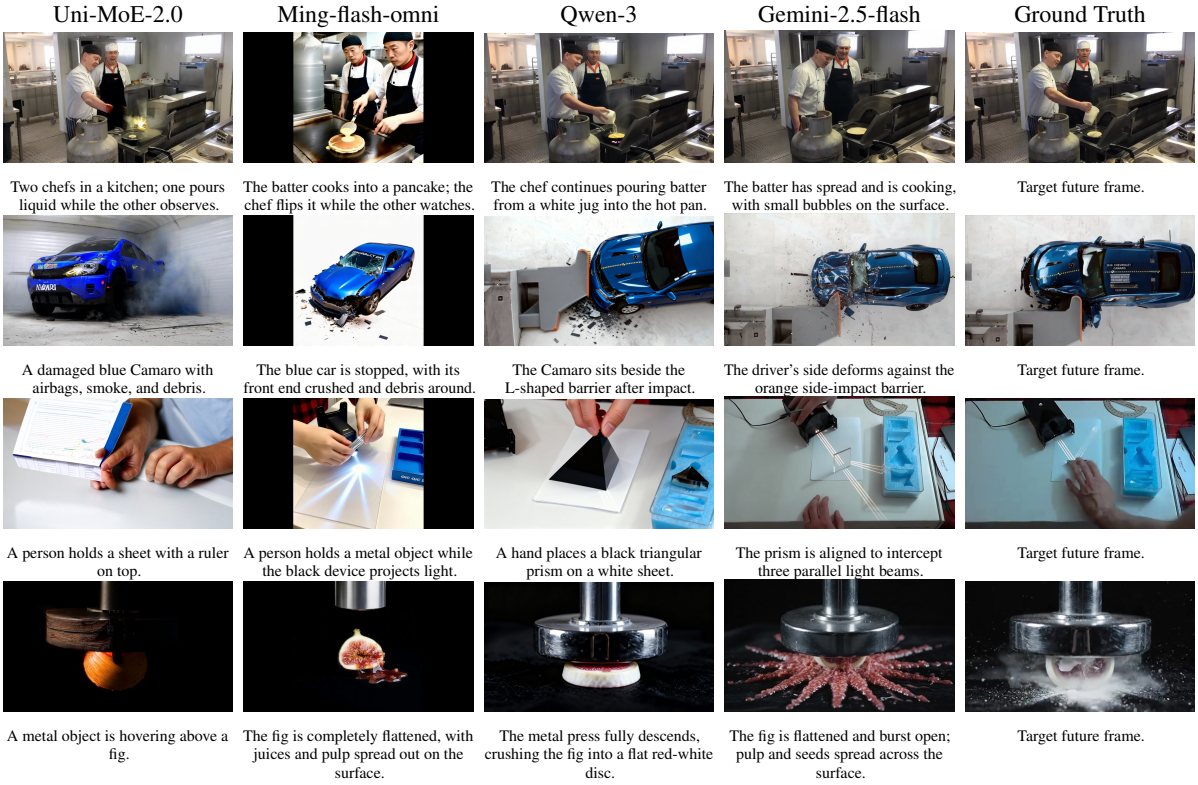


Figure 4: Representative qualitative comparisons with compact, truncated model-generated scene descriptions under each output. Rows show cooking, crash-test, optics, and deformation examples. Gemini better preserves future-state continuity and scene identity, whereas weaker baselines may produce descriptions that are plausible but temporally weak, overly terse, or insufficiently grounded in the generated image.

evaluation appears in both dedicated forecasting benchmarks and subtasks of broader omni-modal suites (Chen et al., 2026; Jiang et al., 2025); WorldSense instead evaluates broad audio-visual understanding via MCQA (Hong et al., 2026). The forecasting tasks are valuable for measuring whether a model can identify plausible future events, but they do not require the model to express that prediction as generated visual and speech outputs whose cross-modal coherence can be inspected.

World models and generative future prediction. World modeling has long supported predictive agents (Ha and Schmidhuber, 2018; LeCun et al., 2022; Hafner et al., 2019), and WorldGPT/WorldNet directly studies multimodal state-transition prediction with a dedicated world-model architecture (Ge et al., 2024). Video generation and next-frame or next-event prediction also emphasize prospective visual dynamics (Brooks et al., 2024; Simon et al., 2026; Wang et al., 2026), but often optimize perceptual plausibility or learned transition modeling rather than benchmarked causal coherence across modalities in existing Omni-models. OmniNSP addresses this gap by requiring synthesized visual and vocal future states, testing whether linguistic foresight becomes coherent multimodal output.

7 Conclusion

We proposed OmniNSP, a curated diagnostic benchmark for evaluating multimodal near-future prediction in Omni-models. Our results show that plausible textual predictions can be degraded during visual and acoustic realization. A frozen-generator control attributes much of the original cross-model visual gap to model-specific realization stacks rather than textual foresight alone, clarifying that OmniNSP diagnoses the integrated path from prediction to output rather than a pure reasoning capability. Modality ablations further demonstrate measurable gains from audio cues, particularly for logical transitions and contexts with informative speech. This stage-wise view enables targeted improvements to future Omni-models. These findings motivate cross-modal state preservation and future video-and-audio generation.

Limitations

Despite our efforts to establish a rigorous evaluation framework, several limitations remain.

Dataset Scale and Coverage. OmniNSP contains 300 human-verified instances. This scale is sufficient for the comparative diagnostic analyses reported here but not for exhaustive claims about all forms of world modeling or all real-world future-state dynamics. Our four baselines were selected for accessible multimodal input and output capabilities and do not cover the full or rapidly changing Omni-model landscape. We prioritize temporally grounded, manually verified examples over larger but noisier coverage. Future work should expand the benchmark and report robustness across broader models, domains, and languages.

Evaluation Protocol and Scope. Our evaluation uses a decoupled two-stage generation protocol because current Omni-models differ in their end-to-end output capabilities. This design exposes stage-wise outputs, but the final score still combines textual prediction with each model's realization stack. Our frozen-generator control strengthens attribution by evaluating all Stage 1 predictions under a common realization stack, while its use of Gemini's generator remains a scope consideration. Human evaluation supports the reliability of LMM judgments for model comparison, particularly on visual dimensions. OmniNSP also focuses on a representative future frame and speech-centered audio synthesis rather than continuous video generation or arbitrary non-speech sound events. Although recent high-end video generators can produce videos from text or still images, many lack a uniform interface for conditioning on the observed context video and generating synchronized video-and-audio continuations. A two-stage variant could pass our textual future prediction to a video generator, but this would require a distinct video-output protocol with temporal metrics and generator-specific constraints; we leave it to future work.

Ethical Statements

Data Privacy and Licensing. Our dataset is constructed from publicly available video content. To respect copyright restrictions and platform terms, we do not redistribute raw video files. The planned release provides metadata and annotations, including source identifiers, timestamps, prompts, model outputs, evaluation rubrics, and reconstruction scripts subject to the availability and license terms of the original sources. We mark license status where available and distinguish redistributable metadata from non-redistributed media. We did not obtain direct consent from individuals who may appear in source videos; accordingly, we avoid redistributing raw media or identifying personal information. For privacy protection, we performed manual and automated filtering to exclude videos containing sensitive personal information or non-consensual imagery.

Human-in-the-Loop Annotation. The data curation and validation process involved human annotators who were tasked with quality control and filtering inappropriate content. All participants were provided with fair compensation and clear guidelines to ensure their well-being. We also implemented a multi-stage review process to minimize exposure to potentially distressing or harmful material during the filtering phase.

Cultural and Social Bias. We acknowledge that the *Social* dynamics pillar involves interpretations of norms that may vary across cultures. While we prioritized globally recognizable and neutral social interactions to mitigate bias, the dataset is provided with the understanding that social behaviors are context-dependent. We have taken care to ensure that the dataset does not reinforce harmful stereotypes or include discriminatory content.

AI Usage Disclosure

The authors acknowledge the use of large language models to assist in the preparation of this work. Specifically, Claude (Anthropic) was utilized to support the development and optimization of the experimental codebase and evaluation scripts. Additionally, ChatGPT (OpenAI) was employed to assist with manuscript writing, including structural refinement and linguistic polishing. Throughout this process, the authors maintained full control over the research design, data analysis, and final interpretation of results, and remain solely responsible for the content and integrity of the work.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-00562784).

References

Inclusion AI, Bowen Ma, Cheng Zou, ChengKun Du, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu

Lian, Chengxiang Fan, Dandan Zheng, et al. 2025. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332.

Cristina Bicchieri. 2006. The grammar of society. *The nature and dynamics of social norms*.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8):1.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Qian Chen, Jinlan Fu, Changsong Li, Min Zhang, See-Kiong Ng, and Xipeng Qiu. 2026. Futureomni: Evaluating future forecasting from omnimodal context for multimodal llms. *arXiv preprint arXiv:2601.13836*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Chris Frith and Uta Frith. 2005. Theory of mind. *Current biology*, 15(17):R644–R645.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355.

Google DeepMind. 2026. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Accessed: 2026-03-17.

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2026. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pages 52423–52443.

Yue Jiang, Dingkang Yang, Minghao Han, Jinghang Han, Zizhi Chen, Yizhou Liu, Mingcheng Li, Peng Zhai, and Lihua Zhang. 2025. Fysicsworld: A unified full-modality benchmark for any-to-any understanding, generation, and reasoning. *arXiv preprint arXiv:2512.12756*.

Yann LeCun et al. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. 2026a. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pages 138214–138236.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22195–22206.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. 2026b. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, et al. 2025. Unimoe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. 2026. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. In *International Conference on Learning Representations*, volume 2026, pages 147298–147334.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Marcel Simon, Eunbin Lee, Tae-Ho Kim, and Seul-Ki Yeom. 2026. Next-frame prediction as a reliability-aware training paradigm for robust vision encoders. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*.

Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. 2019. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32.

Elizabeth S Spelke and Katherine D Kinzler. 2007. Core knowledge. *Developmental science*, 10(1):89–96.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2024. Codi-2: In-context, interleaved, and interactive any-to-any generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27415–27424. IEEE.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. 2023. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099.

NAVER Cloud HyperCLOVA X Team. 2026. Hyperclova x 8b omni. *arXiv preprint arXiv:2601.01792*.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2026. Fostering video reasoning via next-event prediction. In *International Conference on Learning Representations*, volume 2026, pages 31524–31570.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53366–53397. PMLR.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

A Detailed Data Curation Workflow

To ensure the transparency and reproducibility of the OmniNSP dataset construction, we provide a granular description of the seven-stage semi-automated pipeline illustrated in Figure 2.

1. **Category-Specific Keyword Generation**: We used Gemini-3-Flash to generate search keywords tailored to the target dynamics (physical, social, and logical). Keywords were designed to describe concrete observable situations rather than abstract category names, increasing the likelihood that retrieved videos would contain visible future-state transitions.

2. **Large-scale Video Retrieval**: Using the YouTube Data API, we collected candidate videos based on the generated keywords. We prioritized videos with sufficient resolution and audio quality to support multimodal evaluation.

3. **LMM-based Primary Filtering**: To manage the candidate pool, we employed automated screening with Gemini. This stage identified candidate segments containing clear causal transitions or event-driven shifts. The automated stage served as a proposal mechanism; all final instances were human verified.

4. **Human-in-the-loop Temporal Grounding**: Human annotators manually verified the validity of each candidate event. They inspected or corrected the context and target timestamps, ensuring that the context provides sufficient evidence for prediction and that the target segment captures the immediate continuation of the same event. Candidates were rejected when the transition was ambiguous, too trivial, visually unclear, or not temporally aligned.

5. **Multimodal Frame Extraction**: For the current benchmark instantiation, we extracted the last frame of the context segment as a reference image (v_t) and a representative frame from the target segment as the ground-truth future state ($v_{t+\Delta}$). The target frame was selected to capture the salient state change rather than being chosen mechanically from the segment boundary.

6. **Multimodal Captioning and Transcription**: We leveraged Gemini and Whisper to gener-

ate dense visual captions and speech transcriptions. These annotations provide semantic targets for evaluation and support future extensions beyond the current speech-centered generation setting.

7. **Final Safety and Integrity Review**: We conducted a rigorous manual review process to ensure data integrity and compliance with safety standards. This step specifically focused on identifying and **strictly excluding** any content that named or uniquely identified individual people. Furthermore, we implemented a zero-tolerance filtering policy for harmful, sensitive, or biased material; any data points flagged for containing offensive content or potential privacy risks were **entirely removed** from the dataset rather than being anonymized, ensuring the highest level of safety and ethical alignment.

Temporal Offset Selection The temporal offset Δ is event-dependent rather than fixed in seconds. Annotators selected context and target timestamps so that the context contains sufficient evidence for prediction and the target captures the immediate continuation where the salient state change becomes observable. We rejected candidates when the target was too close to the context and therefore trivial, too far from the context and therefore ambiguous, or not temporally aligned with the same underlying event. The representative target frame was then chosen from within the accepted target segment to reflect this state change rather than to maximize visual similarity to the context. Across the final benchmark, Δ has median 2.0 seconds and IQR 1–3 seconds (75% are ≤ 3 seconds). Consistency rankings are identical in short ($\leq$ 1 s), medium (1–3 s), and longer ($>$ 3 s) bins. In an OLS analysis controlling for model and pillar, the coefficient of $\log \Delta$ is small and non-significant for scene ($\beta = -0.022$, $p = .40$), consistency ($\beta = -0.045$, $p = .07$), and alignment ($\beta = -0.047$, $p = .09$).

B Detailed Confidence Intervals

We compute confidence intervals using nonparametric bootstrap resampling over benchmark instances. For mean scores in Tables 3 and 10, each bootstrap sample resamples instances with replacement and recomputes the model-level mean. For pairwise model comparisons, we use paired bootstrap tests that resample the same instance indices across the two compared models, preserving the item-level pairing induced by evaluating all models on the same benchmark instances. The reported p-values are used to support directional comparisons between models, while the confidence intervals describe uncertainty in each model's mean score. As an additional robustness diagnostic, we recompute model rankings under random 50–80% subsamples of the benchmark; this subsampling analysis checks whether the observed ordering is sensitive to a small subset of examples and is not used as a separate null-hypothesis test.

C Frozen-Generator Control Details

For this control, each model's saved Stage-1 scene description is passed to Gemini-2.5-flash-image together with the same context reference frame. This generator is also the image-realization component of the original Gemini pipeline, so Gemini provides an internal control for the generator swap. Both original and frozen outputs are scored under the same rubric with Gemini-3.1-Pro-preview because the original Gemini-3-Pro judge was no longer available.

The consistency gain is significant by paired bootstrap for Qwen, Uni-MoE, and Ming, but not for Gemini. Within the frozen condition, Gemini exceeds Ming in consistency ($p = .026$), while Gemini and Qwen are not statistically distinguishable ($p = .58$). Fidelity convergence provides a complementary check that generator-specific image quality has been equalized. Alignment changes are less uniform because the same generator interacts differently with the models' descriptions and prompt templates.

D Modality Ablation Details

Table 12 reports the full confidence intervals and audio-informativeness breakdown for the main modality ablation. Table 13 provides the corresponding per-pillar diagnostic.

Audio Informativeness Labels The modality-ablation split is based on the input context audio, not on whether the target contains speech. We label a context as *informative speech* when spoken content provides predictive cues about the upcoming event, such as an instruction, warning, reaction, or conversational setup. We label it as *informative non-speech* when non-linguistic audio, such as an impact, alarm, countdown, or environmental sound, helps infer the future state. Contexts with

Model	Scene Description	Audio Script	Generative Fidelity	Consistency	Alignment
Gemini	1.26 ± 0.34	0.88 ± 0.35	1.93 ± 0.14	1.52 ± 0.24	1.71 ± 0.33
Qwen	1.27 ± 0.44	0.76 ± 0.37	1.47 ± 0.38	1.00 ± 0.54	1.40 ± 0.37
Ming	1.13 ± 0.42	0.56 ± 0.31	1.13 ± 0.38	0.29 ± 0.27	1.11 ± 0.53
Uni	1.02 ± 0.45	0.53 ± 0.38	0.44 ± 0.48	0.24 ± 0.39	0.67 ± 0.37

Table 8: Human-evaluation mean scores ($\mu \pm \sigma$) on the 60-item expert-rated subset. These values are distinct from the full-benchmark LMM-judge scores in Table 3.

Model	Audio Script	95% CI
Uni-MoE-2.0	0.64	[0.57, 0.70]
Ming-flash-omni	0.70	[0.63, 0.77]
Qwen-3	0.84	[0.76, 0.93]
Gemini-2.5-flash	**0.96**	[0.86, 1.07]

Table 9: Auxiliary audio-script plausibility results ($n = 196$ target-speech samples). Scores are interpreted as diagnostic because future speech allows multiple plausible continuations and humans applied stricter reference matching than the LMM judge.

silence, background music, generic ambient noise, or audio that merely duplicates visible information are grouped as *uninformative / empty*. Ambiguous cases are assigned to the uninformative group unless the audio provides predictive evidence beyond the visual context.

E Detailed Human Evaluation Setup and Results

This section provides the full human-evaluation protocol and additional statistical results for examining alignment between human ratings and the LMM-as-a-judge framework.

Reference Usage in Evaluation OmniNSP uses real future states as anchors rather than exact reference matching. For the primary scene and image metrics, the judge is given the observed context, the target future frame, and the model output, and scores whether the output is a plausible and semantically consistent continuation. Audio-script plausibility is evaluated only for target-speech samples and is treated as auxiliary because multiple utterances can be reasonable continuations of the same context. Speech synthesis metrics are reported separately as realization diagnostics rather than primary NSP correctness measures.

E.1 Experimental Setup and Protocol

Annotator Recruitment To ensure high-quality qualitative assessment, we recruited three experts currently pursuing degrees in AI-related fields, specializing in multimodal learning and generative models. Crucially, a blinded evaluation setting was maintained; annotators were not informed of the model identities or the scores generated by the LMM judge during the entire process.

Dataset Sampling Given the diverse domains in OmniNSP, we employed a stratified random sampling strategy. We randomly selected 5 samples from each of the three major domains per model, resulting in 15 samples per model. With 4 models evaluated, a total of 60 multimodal samples (video-audio-text triplets) were used for human assessment. This scale was chosen to balance statistical power with the high cognitive load required for expert analysis.

Evaluation Procedure and Compensation Each annotator spent an average of 1 to 2 hours completing the suite. All participants received fair compensation consistent with professional research consulting rates and at or above local research-assistant hourly norms for the expected evaluation time. The evaluation was performed through a dedicated interface allowing simultaneous review of input context and model-generated future scenes.

Instruction and Criteria Annotators received the same task content used in the LMM judge prompts: the observed context, target future frame or audio reference where applicable, model output, and dimension-specific scoring criteria in Table 15. We removed only LMM-specific formatting requirements (e.g., JSON schemas, chain-of-thought triggers) to provide a clean, human-centric evaluation interface. Annotators were instructed to score outputs on the 0–2 rubric, ignore model identity, and skip any item they found unclear or uncomfortable to evaluate.

E.2 Per-Model Performance Breakdown

Table 8 presents the mean scores (μ) and standard deviations (σ) for the four models. Gemini received the highest mean scores for *generative fidelity* and

Model	Mech.		Logic		Social	
	Text	Img	Text	Img	Text	Img
Uni-MoE-2.0	1.29 [1.14,1.43]	0.92 [0.81,1.03]	1.49 [1.36,1.63]	0.55 [0.45,0.66]	1.51 [1.39,1.62]	0.95 [0.84,1.06]
Ming-flash-omni	1.16 [1.02,1.30]	1.44 [1.36,1.52]	1.34 [1.20,1.48]	0.96 [0.83,1.08]	1.36 [1.25,1.47]	1.31 [1.20,1.41]
Qwen-3	1.20 [1.06,1.34]	1.61 [1.54,1.68]	1.39 [1.26,1.51]	1.42 [1.30,1.54]	1.39 [1.28,1.50]	1.76 [1.70,1.82]
Gemini-2.5-flash	**1.57** [1.46,1.68]	**1.82** [1.76,1.87]	**1.64** [1.53,1.74]	**1.69** [1.60,1.76]	**1.60** [1.49,1.70]	**1.85** [1.80,1.89]

Table 10: Detailed pillar performance with 95% bootstrap confidence intervals. This table expands Table 5.

Model	Fidelity		Consistency		Alignment	
	Original	Frozen	Original	Frozen	Original	Frozen
Gemini-2.5-flash	1.23	1.26	1.12	1.12	1.21	1.27
Qwen-3	1.19	1.26	0.85	1.10	1.20	1.22
Uni-MoE-2.0	0.92	1.29	0.53	1.07	0.83	1.34
Ming-flash-omni	1.16	1.29	0.47	1.04	1.16	1.32

Table 11: Full frozen-generator results scored with Gemini-3.1-Pro-preview. Each condition contains 300 examples per model, except Uni-MoE original images ($n = 299$).

alignment, while the other models received lower *consistency* scores.

E.3 Inter-Rater Reliability and Alignment Analysis

We analyze consensus using pooled Fleiss' κ, exact pairwise agreement, Krippendorff's α for ordinal labels, and pairwise quadratic-weighted κ. Fleiss' κ is computed at the pooled item level over model–sample pairs to match the human-evaluation design. As shown in Table 14, ordinal-aware metrics indicate moderate agreement for visual fidelity and consistency, while agreement is weaker for audio-script judgments.

For human–LMM validation, we average the available human ratings for each of the 60 model–sample items. In addition to item-level Spearman correlation, we compare the four model means using Kendall's τ and compute same-sample pairwise model-preference agreement after excluding ties. Consistency yields item-level $\rho = 0.85$, model-rank $\tau = 0.91$, and 95.0% pairwise agreement. Alignment and fidelity yield $\rho = 0.64/0.65$, $\tau = 1.00/1.00$, and 91.7%/90.7% pairwise agreement. Scene description yields $\rho = 0.42$, $\tau = 0.33$, and 80.6% pairwise agreement, while the visual dimensions show consistently stronger alignment. Together, these results support the LMM judge as a scalable evaluator for model comparison, with especially strong alignment on visual dimensions.

Evaluator Bias Pairwise Spearman correlation analysis confirmed that while individual scoring thresholds varied—specifically, one evaluator was consistently more stringent (mean score 0.36–0.96)—the relative ranking of models remained stable across all annotators. This robust ranking alignment supports the conclusion that performance gaps identified by OmniNSP are perceptible and consistent across different experts.

F Example Prompt and Configuration

The following example illustrates the Stage 1 instructions used to elicit structured future-state predictions. The placeholders Δ and d are replaced per instance with the temporal offset and target-segment duration.

Example Stage 1 prompt

> You are a multimodal reasoning model performing a Next Scene Prediction task. Given a context video, predict the scene Δ seconds after it ends. Return (1) a 1–3 sentence *Scene Description* specifying the people or entities, actions, objects, environment, and visible state changes while maintaining continuity with the context, and (2) an *Audio Script* containing only the spoken dialogue expected at that moment, corresponding to approximately d seconds of speech. Use an empty string when no one is speaking. Return only valid JSON:
>
> ```
> {"scene_description": "...",
>  "audio_script": "..."}
> ```

Generation and evaluation configuration All models followed the same semantic instructions and common output schema. Uni-MoE and Ming produced the scene-description and audio-script fields in separate calls; Ming used `[NO SPEECH]` when no speech was predicted. Stage 2 generated

Group	n	Full	Video-only	Audio-only	Δ F−V
All	300	1.60 [1.54,1.66]	1.52 [1.46,1.59]	0.61 [0.52,0.70]	**0.08 [0.01,0.15]**
Informative speech	141	1.66 [1.58,1.74]	1.52 [1.42,1.61]	0.82 [0.68,0.96]	**0.14 [0.05,0.23]**
Informative non-speech	10	1.50 [1.20,1.80]	1.60 [1.20,1.90]	0.90 [0.50,1.30]	-0.10 [-0.40,0.20]
Uninformative / empty	149	1.56 [1.46,1.64]	1.52 [1.44,1.62]	0.39 [0.28,0.50]	0.03 [-0.08,0.15]

Table 12: Detailed Gemini-2.5-flash modality ablation by audio informativeness. Brackets are 95% bootstrap confidence intervals; Δ denotes Full − Video-only.

Pillar	n	Full	Video-only	Audio-only	Δ F−V
Physical	100	1.57 [1.46,1.68]	1.57 [1.45,1.68]	0.44 [0.31,0.58]	0.00 [-0.12,0.13]
Logical	100	1.64 [1.53,1.74]	1.48 [1.36,1.60]	0.73 [0.57,0.91]	**0.16 [0.02,0.31]**
Social	100	1.60 [1.50,1.70]	1.52 [1.41,1.62]	0.65 [0.51,0.80]	0.08 [-0.02,0.19]

Table 13: Modality ablation by pillar. The logical pillar benefits most from audio; the physical pillar shows no measurable audio effect.

one future image from the context frame and saved Stage 1 description. The frozen control used the same inputs with Gemini-2.5-flash-image. Outputs were scored under the criteria in Table 15; the original results used Gemini-3-Pro, and the frozen control used Gemini-3.1-Pro-preview with the same rubric and input ordering.

Dimension	Pairwise Agree.	Fleiss' κ	Kripp. α	Weighted κ
Scene Description	58.2%	0.215	0.277	0.278
Audio Script	54.8%	0.112	0.228	0.269
Generative Fidelity	61.0%	0.390	**0.583**	**0.605**
Consistency	54.2%	0.280	**0.573**	**0.562**
Text-Image Align.	53.1%	0.257	0.386	0.401

Table 14: Inter-rater reliability among human experts using pooled model-sample items. Pairwise agreement denotes mean exact-match agreement across annotator pairs. Fleiss' κ is computed on pooled items; Krippendorff's α and weighted κ account for the ordinal 0–2 scale.

Metric	Score 0 (Poor)	Score 1 (Fair)	Score 2 (Good)
Scene Description	Predicted scene is causally inconsistent with the context video.	Could plausibly follow context, but does not match the target image.	Plausibly follows context AND aligns with the target image.
Audio Script	Inconsistent with both the context and the target audio caption.	Could plausibly occur, but does not match the target audio.	Closely matches the target audio caption.
Generative Fidelity	Severe artifacts present or the predicted scene is unclear.	Scene is recognizable but contains visible imperfections.	Clear, high-quality, and visually plausible generation.
Consistency	Inconsistent with the observed scene and its dynamics.	Partially consistent with the context but lacks precise continuity.	Visually consistent and logical continuation of the context.
Text-Image Align.	Image content and text descriptions are semantically inconsistent.	Partially consistent alignment between visual and textual elements.	Well-aligned; text perfectly describes the generated visual.

Table 15: Detailed human evaluation criteria for OmniNSP. Each dimension is rated on a 3-point Likert scale (0–2) to assess the reasoning and generative quality of Omni-models.

Metric	Evaluated output	Reference / anchor	Role
Scene Description	Predicted future-scene text	Context segment and target future frame	Primary reasoning score
Generative Fidelity	Generated future image	Generated image quality, with the target future frame as task anchor	Primary visual synthesis score
Consistency	Generated future image	Context reference frame and target future frame	Primary future-state consistency score
Text-Image Align.	Generated image and predicted text	Predicted text, generated image, and target future frame	Primary cross-modal alignment score
Audio Script	Predicted future speech content	Target speech/audio annotation; target-speech samples only	Auxiliary diagnostic
WER / UTMOS / Speaker Similarity	Generated speech	Predicted script and context speaker audio where applicable	Auxiliary speech realization diagnostic

Table 16: Reference and anchor usage across OmniNSP evaluation metrics.