HyperCLOVA X 8B Omni

NAVER Cloud
HyperCLOVA X Team*
🤗 Huggingface Model Card

Abstract

In this report, we present HyperCLOVA X 8B Omni, the first any-to-any omni-imodal model in the HyperCLOVA X family that supports text, audio, and vision as both inputs and outputs. By consolidating multimodal understanding and generation into a single model rather than separate modality-specific pipelines, HyperCLOVA X 8B Omni serves as an 8B-scale *omni-pathfinding* point toward practical any-to-any omni assistants. At a high level, the model unifies modalities through a shared next-token prediction interface over an interleaved multimodal sequence, while vision and audio encoders inject continuous embeddings for fine-grained understanding and grounding. Empirical evaluations demonstrate competitive performance against comparably sized models across diverse input–output combinations spanning text, audio, and vision, in both Korean and English. We anticipate that the open-weight release of HyperCLOVA X 8B Omni will support a wide range of research and deployment scenarios.

1 Introduction

The tight integration of AI systems into real-world contexts necessitates their ability to understand and generate across multiple modalities, such as text, audio, and vision. This requirement arises in part because specific applications inherently involve multimodal inputs and outputs. Moreover, human-generated text is projected to accumulate at a rate that cannot keep up with the rapid scaling of large language models (LLMs; Villalobos et al. 2024). Even if it were the case, text alone cannot capture the full spectrum of multimodal dimensions of reality (Huh et al., 2024; Chen et al., 2025a).

One strategy for developing multimodal models extends existing LLMs by sequentially incorporating encoders and decoders for various modalities. While such modality extension enables a cost- and time-efficient transformation of a text-based model into a multimodal one, multimodal training often incurs catastrophic forgetting of knowledge within the LLM backbone (Zhai et al., 2023; Driess et al., 2023; Lee et al., 2025; Liu et al., 2025). This challenge calls for a joint training across multiple modalities in a unified framework.

In response, we introduce HyperCLOVA X 8B Omni (OMNI), an omnimodal model that supports text, audio, and vision modalities as both inputs and outputs, as shown in Figure 1. OMNI is a decoder-only Transformer jointly modeling an interleaved multimodal sequence of tokens and embeddings. Modality-specific tokens and embeddings share a common next-token prediction interface, thereby facilitating semantic composition across modalities.

We compare the performance of OMNI against that of comparably sized models on benchmarks spanning diverse combinations of input and output modalities, including text-to-text, vision-to-text, text-to-vision, speech-to-text, audio-to-text, and speech-to-speech. In addition, we present a human preference study on text-to-speech conversion. For most modality combinations, evaluations are
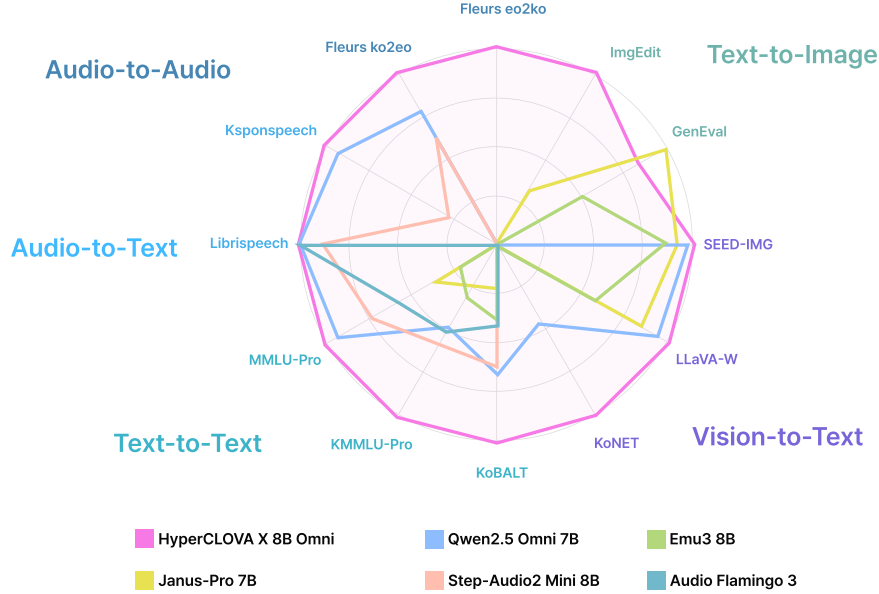

Figure 1: Comparison of multimodal capabilities across text, vision, and audio for both generation and understanding tasks. The results highlight the unified end-to-end design of HyperCLOVA X 8B Omni, which supports any-to-any multimodal understanding and generation within a single model.

carried out in both Korean and English to assess bilingual ability. The results demonstrate the competitive performance of OMNI across the board, despite it being the only model capable of handling all combinations of input and output modalities.

OMNI is released as an open-weight model under a custom license that permits commercial use subject to specified conditions. Given its compact size and competitive performance across diverse input and output modalities, we present OMNI as a valuable resource for academic and industry partners in both the Korean and global research community.

2 HyperCLOVA X 8B Omni

2.1 Design Motivation and Pathfinding

Recent multimodal systems span a wide design space, ranging from late-fusion integration to modality-specific generation pipelines (Team et al., 2025; Chen et al., 2025b; AI et al., 2025; Xu et al., 2025; Chu et al., 2024). In our approach, the guiding hypothesis is that multimodal capabilities can be effectively realized when modality-specific tokens and embeddings share a common next-token prediction interface, enabling semantic composition across modalities. As illustrated in Figure 2, text is represented as discrete tokens, while vision and audio are represented with both discrete tokens and continuous embeddings; these representations are interleaved and jointly processed by a single decoder-only Transformer backbone.

We instantiate the backbone as a 36-layer auto-regressive Transformer with a hidden size of 4,096, closely following the architectural and implementation choices of HyperCLOVA X 32B Think (THINK, HyperCLOVA X Team (2025)). Following THINK, the text tokenization pipeline combines a morphology-preserving pretokenizer and a subword tokenizer and applies low-probability StoChasTok to mitigate token-boundary bias while preserving token efficiency. For subword tokenization, we adapt an English-centric tokenizer via a three-stage vocabulary modification, which significantly improves Korean token efficiency without degrading performance on English, code, or math tasks.

Operationally, we unify multimodal generation by treating each modality tokenizer's discrete code-book entries as additional vocabulary items of the language model, thereby extending next-token
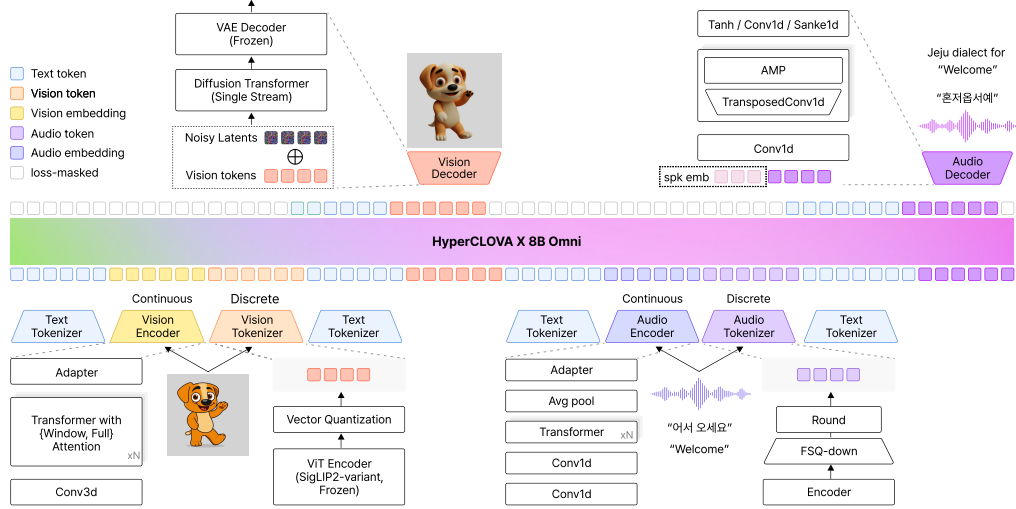
Figure 2: Overall architecture of HyperCLOVA X 8B Omni. Text, vision, and audio inputs are encoded into continuous embeddings and discrete tokens via modality-specific encoders and tokenizers, which are interleaved and jointly processed by a single decoder-only Transformer backbone. Modality-specific decoders reconstruct visual and auditory outputs from the shared sequence representations, enabling end-to-end any-to-any multimodal generation.

prediction from text to a shared multimodal token space. For understanding and fine-grained grounding, we additionally attach modality encoders that produce continuous embeddings projected into the backbone embedding space. Modality decoders subsequently convert predicted non-text tokens into their native signal domains (pixels and waveforms).

The following sections provide detailed specifications of the vision and audio tokenizers/encoders and the associated decoders.

2.2 Vision Modality

OMNI processes visual information through a synergistic integration of three components: a continuous vision encoder for perceptual understanding, a discrete semantic tokenizer for generative representation, and a diffusion-based decoder for pixel synthesis. This tripartite architecture is designed to natively handle interleaved multimodal sequences within a unified framework, in which each component plays a functional role.

First, a continuous vision encoder extracts dense features that are directly aligned with the LLM backbone to support overall vision understanding. Second, to support vision generation, OMNI incorporates a vision tokenizer that quantizes visual features into discrete semantic tokens. This choice is closely tied to the auto-regressive (AR) nature of our Transformer backbone, which is inherently well-suited to modeling discrete tokens (Li et al., 2024b; Deng et al., 2024). Unlike models such as Janus-Pro (Chen et al., 2025b) or Emu 3 (Wang et al., 2024) that rely on low-level VAE-style tokenizers, our tokenizer operates at the semantic level to maximize cross-modal synergy with text embeddings (Zheng et al., 2025)—a critical advantage for our compact 8B backbone, where efficient and semantically aligned features are essential.

Finally, vision generation proceeds by decoding these discrete tokens into pixels using a diffusion-based vision decoder. Because semantic tokenization introduces unavoidable information loss by discarding fine-grained visual details, the diffusion model acts as a complementary component that stochastically recovers missing details. It synthesizes high-frequency textures and fine structures through a channel-concatenation-based architecture, which enables significantly faster convergence and supports near-native aspect ratios.

Encoder. Architecturally, the vision-understanding component of OMNI follows THINK, adopting the Vision Transformer (ViT) architecture from Qwen2.5-VL (Bai et al., 2025) for unified image and video modeling. For architectural stability, we utilize a streamlined linear adapter to align visual
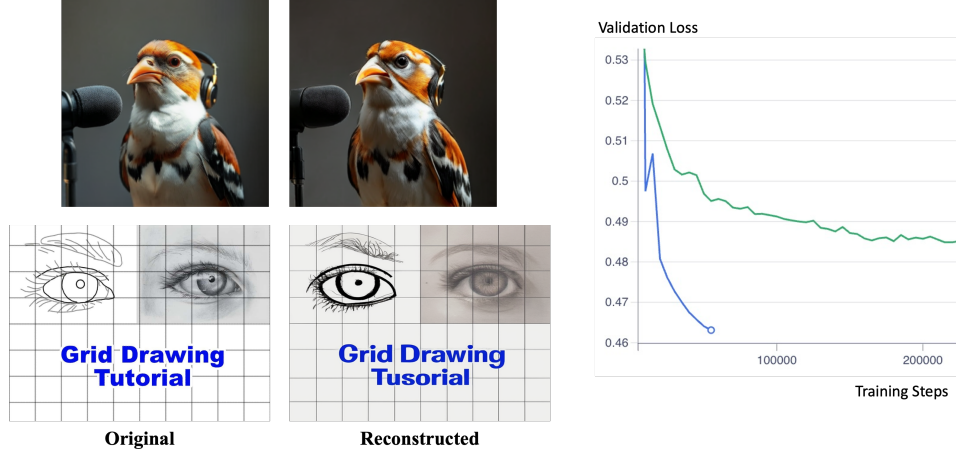
3

Figure 3: (**Left**) Reconstruction test of TA-Tok (Han et al., 2025) using its accompanying decoder. The reconstruction is imperfect due to unavoidable information loss from semantic abstraction and quantization (see the eye and the feather pattern of the bird, and the tonal difference in both cases). A non-square image shown at the bottom is "resized to square → tokenized → decoded as a square image → resized back to the original aspect ratio." We observe that the distortion level is not critical and presume that it could be compensated for by training a new decoder reflecting this process. (**Right**) Convergence of validation loss for the classic attention-based architecture (green) and our channel-concatenation-based architecture (blue).

features with the LLM backbone (Liu et al., 2023a). A primary design objective is computational efficiency; by optimizing the visual token allocation, we reduce training costs by approximately 53% in GPU-hours compared to standard settings. Static images and 120-frame videos are compressed into efficient budgets of 3K and 11K tokens, respectively. Notably, the encoder remains unfrozen throughout training to establish Korean-centric multimodal capabilities, essential for internalizing Korean-specific visual contexts, cultural landmarks, and high-density OCR.

Tokenizer. We reuse a pretrained text-aligned tokenizer, TA-Tok (Han et al., 2025), and keep it fully frozen during training. TA-Tok fine-tunes SigLIP 2 (Tschannen et al., 2025) to quantize its output—patch-wise visual features—into discrete tokens and reconstruct the original visual features from these tokens. One practical limitation of TA-Tok is its fixed input resolution of 384×384. While the loss of resolution is largely compensated by the diffusion-based vision decoder, non-square images must be resized, which may introduce geometric distortion. We empirically evaluated this issue in advance and found that it does not lead to severe degradation in practice (see Figure 3). The degradation is further mitigated by training our own decoder from scratch, in which such resizing scheme is directly integrated.

Decoder. Our vision decoder is similar to the decoders released alongside the TA-Tok model (Han et al., 2025), but it differs in two key aspects. First, it uses a channel-concatenation-based conditioning architecture that enables significantly faster convergence than attention-based. Second, it supports near-native aspect ratios, avoiding the strict square-image constraint imposed by TA-Tok decoders.

The model adopts a diffusion transformer composed exclusively of single-stream blocks of MMDiT (Labs et al., 2025), 2B parameters in total. It operates on the latent space of FLUX.1 VAE (Labs et al., 2025) with patch size 1. Importantly, our model does not use any text-conditioning; the only conditioning signal is vision tokens, which are injected via channel-wise concatenation with the noisy latents. Concretely, discrete vision tokens produced by TA-Tok have a fixed spatial resolution of 27×27; these tokens are first reconstructed into continuous feature vectors and then resized to match the shape of the latents (e.g., 116×78 for a 928×624 image) before concatenation. We empirically observe that this design significantly improves the convergence speed (see Figure 3). Moreover, by avoiding attention–based conditioning, the overall computational cost of the model is reduced significantly. Detailed descriptions of the decoder model training and inference are provided in Appendix A.

4

2.3 Audio Modality

OMNI is designed to support both audio understanding and generation within a unified language modeling framework. The audio module consists of a continuous audio encoder, a discrete audio tokenizer, and a neural audio decoder. Continuous acoustic embeddings and discrete audio tokens are provided as separate input streams to the language model, enabling joint processing of audio and text within a single Transformer backbone. For speech synthesis, discrete audio tokens predicted by the language model are passed to the audio decoder, which reconstructs the time-domain waveform.

Encoder. For continuous audio representation, we adopt a pretrained audio encoder (Chu et al., 2024), which is initialized from the Whisper-large-v3 model (Radford et al., 2023). The input audio is resampled to 16 kHz and transformed into a 128-channel log-mel spectrogram using a 25 ms window size and a 10 ms hop size. A pooling layer with a stride of two is applied to reduce the temporal resolution, such that each output frame approximately corresponds to a 40 ms segment of the original audio. As a result, the encoder produces continuous audio embeddings at an effective frame rate of 25 Hz. Subsequently, the encoder outputs are mapped to the dimension of the language model embeddings via a two-layer MLP adapter consisting of a Linear-GELU-Linear structure. To handle audio within video sequences efficiently, we implement an additional token compression mechanism following Kim and Seo (2025a). Specifically, we incorporate a single-layer MambaMia module (Kim and Seo, 2025b) after the MLP adapter to further downsample the audio representations from 25 Hz to 1 Hz. This architectural refinement significantly enhances token efficiency, allowing the model to process long-form video-interleaved audio while maintaining a manageable context budget. Throughout the training process, the audio encoder remains frozen to fully leverage the robust acoustic representations learned during large-scale pretraining.

Tokenizer. In addition to continuous embeddings, we employ a pretrained audio tokenizer (Du et al., 2024) to represent speech as discrete units. This tokenizer inserts a finite scalar quantization (FSQ) module (Mentzer et al., 2024) into the encoder of a pretrained SenseVoice-Large Automatic Speech Recognition (ASR) model (An et al., 2024). The input speech is first processed by a stack of Transformer blocks to obtain intermediate representations, which are then projected into a low-rank space and quantized using bounded rounding in the FSQ module. The quantized representations are subsequently projected back to the original dimensionality, and discrete audio tokens are obtained by indexing the quantized low-rank vectors in a (2K+1)-ary system. This process yields a codebook of size 6,561 tokens. The resulting audio tokens are generated at a fixed rate of 25 tokens per second, perfectly aligning with the temporal resolution of the continuous audio embeddings.

This dual-encoding design allows the model to exploit the complementary advantages of both representations. Continuous audio embeddings preserve fine-grained acoustic information and rich prosodic details, while discrete audio tokens provide a compact and generation-friendly representation that is well-suited for autoregressive modeling and waveform synthesis.

Decoder. To reconstruct time-domain waveforms from discrete audio tokens, we propose an audio decoder named *Unit-BigVGAN*. The decoder is architecturally derived from BigVGAN-v2 (gil Lee et al., 2023), but is adapted to consume discrete audio tokens generated by the LLM rather than continuous mel-spectrogram features. As the decoder directly operates on symbolic unit sequences, which encode limited speaker identity information, the model conditions the generator on an explicit speaker embedding. A reference speech signal is processed by an ECAPA-TDNN (Desplanques et al., 2020) to extract a fixed-dimensional speaker embedding that captures speaker-specific characteristics. The embedded discrete tokens are concatenated with the speaker embedding along the channel dimension and used as input to the generator.

Given this combined representation, the decoder follows a BigVGAN-style upsampling and residual processing pipeline to progressively increase temporal resolution and generate the final waveform. The generator consists of an initial convolution layer followed by multiple upsampling stages with residual dilated convolution blocks. Within these residual blocks, a filtered Snake nonlinearity is applied to obtain an anti-aliased representation of discrete-time one-dimensional signals, an approach known as anti-aliased multi-periodicity composition (AMP). Under this decoding pipeline, each discrete token represents a fixed temporal span of 40 ms, corresponding to a token rate of 25 Hz.
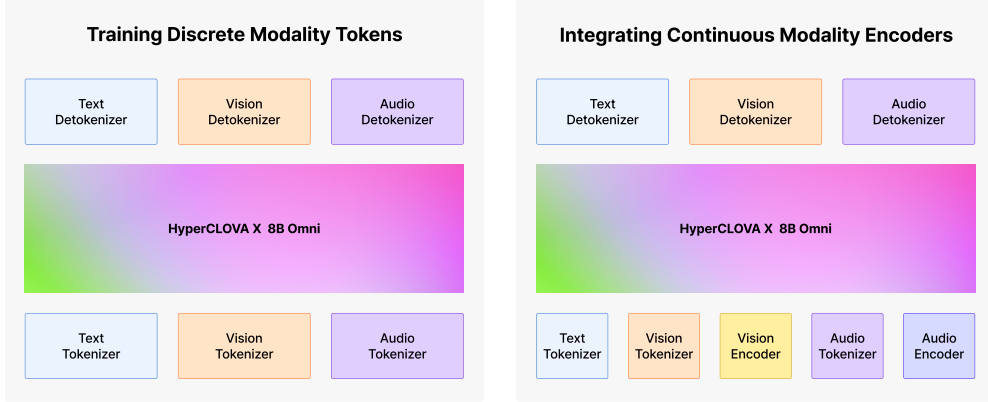
Figure 4: Overview of the training process. The model is first trained using discrete modality tokens for text, vision, and audio, establishing a unified symbolic token interface across modalities. Continuous vision and audio encoders are then integrated and jointly trained alongside the discrete tokens, enabling richer multimodal perception within the same Transformer backbone.

3 Pre-Training

The pre-training consists of two phases to train a text-centric foundation and progressively instill multimodal capabilities. First, the model learns discrete modality tokens for text, vision, and audio, establishing a unified symbolic token interface that enables joint sequence modeling across modalities. Next, continuous modality encoders for vision and audio are integrated and jointly optimized with the existing discrete representations, allowing the model to incorporate rich perceptual signals while preserving a unified token-based processing paradigm.

As shown in Figure 4, the training recipe is designed to progressively build up the omnimodal capabilities of the model. Training begins with text-only pre-training to establish a strong sequence-modeling foundation, followed by omnimodal pre-training that extends the backbone to vision and audio through discrete token learning and the integration of continuous-modality encoders.

3.1 Text Pre-training

Data. The data preparation pipeline follows the scalable preprocessing framework used in THINK, implemented with a hybrid processing stack based on Datatrove and NeMo-Curator. Raw data are collected and normalized into a unified schema, annotated with document-level quality signals (including PII masking), filtered using a combination of heuristic and model-based criteria, expanded with synthetic data, and finally serialized into sharded files for efficient streaming-based training. Data filtering combines heuristic rules and model-based signals to remove low-quality samples while minimizing unnecessary data loss. This consideration is crucial for Korean corpora, where overall data availability is relatively limited. Filtering policies are pre-defined and applied consistently within each training run. Synthetic data generation leverages two complementary approaches: seed-based generation and document rewriting. Reasoning-oriented synthetic data for STEM, code, and mathematics domains are incorporated during mid-training rather than post-training. The proportion of synthetic data is controlled to preserve training stability while maintaining sufficient diversity.

Backbone. The text backbone is pretrained using a multi-stage curriculum with progressively increasing context lengths of 4K, 8K, and 32K tokens. Later stages place greater emphasis on long-form, high-quality, and reasoning-oriented data, with batch sizes adjusted accordingly to accommodate longer contexts. To improve training efficiency under the limited parameter capacity of the 8B-scale backbone, we employ multi-token prediction (Gloeckle et al., 2024). Specifically, an auxiliary prediction head with a single additional layer is introduced and weighted by a scaling factor of 0.2. This design increases supervision density per token while preserving the original next-token prediction objective, resulting in more effective utilization of training signals without altering the primary optimization target.

3.2 Training Discrete Modality Tokens

Stage 1: Multimodal Vocabulary Expansion. In this stage, the discrete codebooks produced by modality-specific tokenizers (vision & audio) are incorporated into the text vocabulary, effectively expanding the model's token space to support multimodal symbolic representations. To prevent degradation of text capabilities during the initial multimodal expansion, the text-token portions of the token embedding matrix and the LM head, along with the decoder layers, are frozen, whereas the embeddings for the newly introduced non-text modality tokens are trained for alignment. This stage is trained for 36K steps, corresponding to 302B tokens, using primarily image–text and audio–text paired data, with text-only data kept at a minimal ratio. To account for modality token scale differences, the input mixture is controlled with a fixed Image:Audio ratio of 3:1.

Stage 2: Full-Parameter Multimodal Pre-training. All parameters are made trainable to enable cross-modal fusion and multimodal reasoning. This stage performs large-scale end-to-end multimodal pre-training over 2.3T tokens, with modality ratios and loss masking carefully controlled to mitigate text degradation caused by the large vision token budget. A curriculum-based loss masking strategy is applied to stabilize training. Specifically, during the initial phase spanning the first 1T tokens, the modality mixture is set to Text:Image:Audio = 2:6.5:1.5, with a vision-token loss masking factor of 0.5. For the second phase, spanning from 1T to 2.3T tokens, the same mixture ratio is maintained while vision loss masking is restored to 1.0.

Stage 3: Long-Context Adaptation for Multimodal. A short and focused long-context adaptation is performed to support downstream vision-interleaved and high-difficulty reasoning data under an extended context. This stage continues from the Stage 2 checkpoint with a 32K context length and a reduced global batch size to improve stability on long sequences, and is trained on approximately 20B tokens.

3.3 Integrating Continuous Modality Encoders

In this phase, we integrate continuous modality encoders for both vision and audio to strengthen perceptual modeling and to align continuous and discrete modality representations within a unified sequence modeling framework. While both encoders are incorporated into the architecture, only the vision encoder is actively optimized to enhance visual perception and to align its representations with those produced by the vision tokenizer.

Stage 1: Vision Encoder Alignment. In the first stage, we align visual features with the language model's embedding space. We keep both the language model backbone and the vision encoder frozen, and we train only a lightweight linear adapter. The training data predominantly consists of image–caption pairs (75.0%), basic OCR tasks (20.0%), and VQA samples (5.0%), establishing a foundational mapping between visual tokens and linguistic representations.

Stage 2: Vision-Centric Full-Parameter Pre-training. In the second stage, we train all model parameters to enhance Korean-specific visual perception, including cultural entities, local landmarks, and high-density Korean-script OCR. To preserve previously acquired capabilities across other modalities, we train on a total of 1.5T tokens spanning text (12.1%), visual understanding (38.5%), visual generation (34.4%), and audio (15.0%). The visual understanding data primarily comprises interleaved in-house and public datasets that capture general visual knowledge, along with text-rich OCR data. The visual generation data includes not only text-to-image samples but also image-editing data. Since image editing utilizes both the vision encoder and the vision tokenizer, this stage further promotes representation alignment between the two components.

Stage 3: Audio Encoder Alignment. In the final step of architectural integration, we incorporate the continuous audio encoder to complete the omnimodal framework. While the preceding stages leveraged discrete tokens to accelerate training, this stage introduces a continuous encoder in parallel to process dense acoustic information. To achieve this, we focus exclusively on ASR tasks, training only a lightweight adapter to bridge the audio encoder and the language model backbone. This stage finalizes the unified architecture, establishing a stable foundation for the subsequent post-training pipeline.
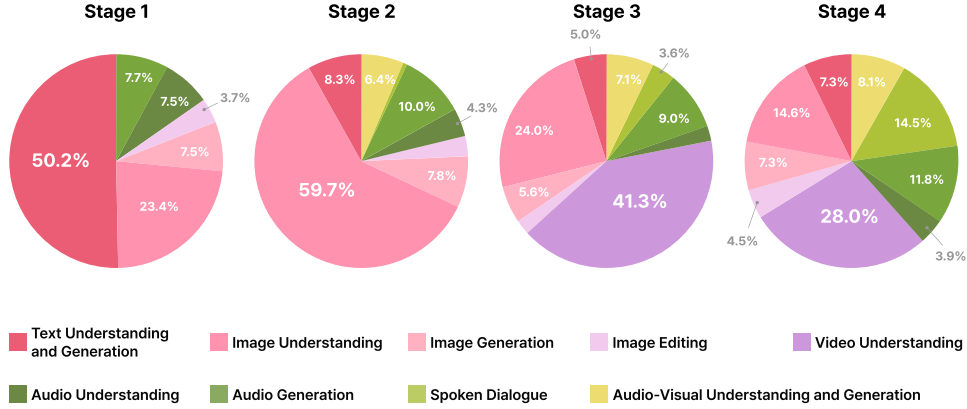
Figure 5: Distribution of the post-training datasets across four training stages. Stage 1 focuses on foundational conversational alignment; Stage 2 expands to task-oriented multimodal instructions; Stage 3 introduces temporal and long-context understanding; and Stage 4 refines user-intent reasoning through integrated reasoning paths.

4 Post-Training

The post-training is designed to transform the pre-trained omnimodal backbone into a Korean-centric AI assistant capable of seamless, instruction-following interaction across text, audio, and vision modalities. Our primary philosophy centers on a *staged curriculum* that progressively transitions from foundational conversational alignment to complex, intent-aware reasoning. While reinforcement learning is recognized as a potent tool for preference alignment, this work primarily focuses on the architectural and data-driven strengths of a supervised fine-tuning (SFT) framework. By establishing this high-fidelity functional baseline, we provide a stable platform capable of seamless cross-modal reasoning, which we identify as the primary milestone for our omnimodal assistant.

4.1 Data Composition and Strategy

The post-training data is meticulously curated to balance general linguistic intelligence with specialized multimodal capabilities. We utilize a mix of high-quality human-annotated dialogues and synthetic reasoning traces designed to enhance logical depth.

A central pillar of our strategy is the prevention of catastrophic forgetting; hence, the curriculum starts with a high concentration of text-based SFT to maintain a stable linguistic foundation before transitioning to complex omnimodal tasks. Specifically, in the initial stage, text-only data constitutes the majority of the training volume (50.2%), ensuring that OMNI preserves its core reasoning and multilingual abilities.

For vision and audio modalities, our strategy emphasizes a progressive increase in task-oriented instructions. This phased, omnimodal-aware approach ensures that OMNI handles diverse any-to-any scenarios without degrading its fundamental capabilities. The detailed distribution of these datasets across the four training stages is visualized in Figure 5.

4.2 Training Recipe

The SFT process for OMNI is organized into four sequential stages, each targeting specific functional milestones.

Stage 1: Foundational Omni Alignment. The initial stage is dedicated to establishing foundational instruction-following capabilities by adapting the pre-trained backbone to a conversational framework. As this stage serves as the primary transition point toward a dialogue-centric assistant, it required the most substantial investment of computational resources and training time within the

post-training pipeline. Central to this phase is the prioritization of text-based SFT, which constitutes the majority of the training volume (50.2%). This high concentration of linguistic data is critical for ensuring that OMNI maintains a stable linguistic foundation and robust reasoning capabilities before the introduction of more complex omnimodal tasks. This core text-only data is complemented by foundational omnimodal tasks, such as image captioning, ASR, Text-to-Speech (TTS), image generation, and image editing tasks. By allocating the largest portion of the total computational budget to this stage, we ensure a reliable alignment between the model's perceptual outputs and the nuances of human dialogue.

Stage 2: Task-Oriented Omni Specialization. The second stage represents a pivotal expansion of the model's functional repertoire, characterized by an exponential increase in both the volume and diversity of instructional tasks. With the linguistic foundation firmly established in Stage 1, the training focus strategically shifts toward large-scale multimodal instruction tuning. During this phase, the proportion of text-only SFT is significantly reduced (8.3%) to accommodate a vast ecosystem of task-oriented omnimodal data, with a primary emphasis on complex image understanding. The objective is to cultivate omnimodal synergy by exposing OMNI to heterogeneous scenarios. By navigating these interleaved tasks, the model learns to synthesize cross-modal evidence, enabling it to solve complex queries that require simultaneous processing of text, audio, and vision.

Stage 3: Long-Context and Video SFT. Stage 3 is primarily dedicated to temporal modeling and long-context management, with a central focus on video understanding. We integrate a substantial volume of video understanding data (41.3%) and incorporate instruction samples that feature extensive internal reasoning traces to bolster model's logical depth. This recipe instills OMNI with the ability to maintain semantic coherence over long multimodal sequences and to perform reasoning over temporal events. To handle the high-resolution audio-visual streams within video efficiently, we introduce a dedicated audio token compressor for video inputs. By preceding this stage with a brief alignment phase trained exclusively on video data, with only the compressor module set as trainable, we ensure stable integration of the compressor module into the omnimodal pipeline.

Stage 4: Intent-Aware Multistep Reasoning. The final stage of our post-training curriculum is designed to instill OMNI with the capacity for both high-level intent parsing and sustained, multistep logical reasoning. We internalize a structured reasoning mechanism—the `<think>` block—to serve as the model's cognitive workspace. In this framework, the reasoning trace typically incorporates an initial intent classification step, where OMNI identifies the task category and orchestrates the necessary modality-specific modules. For high-complexity tasks, such as STEM problem-solving or cross-modal integration, this initial mapping can naturally transition into an extended deductive reasoning process. By systematically breaking down instructions into intermediate logical steps within the latent space, the model can navigate complex problem domains while maintaining strict adherence to user-defined constraints. This versatile reasoning paradigm ensures that OMNI not only selects the correct functional path but also executes deep, context-aware analysis when the task demands it. Examples illustrating the foundational intent parsing and task orchestration are provided in Appendix C.

5 Evaluation

To evaluate the performance of OMNI[2], we select a set of open-source multimodal models as baselines, covering text, vision, and audio modalities. The baseline models are chosen based on similarity in model scale, reproducibility of reported evaluation results, and representativeness within each modality. All comparisons are conducted only on the modalities that each model explicitly supports.

Qwen2.5-Omni-7B (Xu et al., 2025) is an Omni model that jointly supports text, vision, and audio, and serves as the most directly comparable baseline to OMNI across all modalities. Both models are based on a general-purpose Omni architecture, enabling consistent comparisons on text understanding and generation, vision–language reasoning, and speech recognition and generation tasks.

---

[2]In these experiments, we use the checkpoint released under the `TAG-2025-12-31`, tag: `https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Omni-8B/tree/TAG-2025-12-31`.

Language	Dataset	HyperCLOVA X 8B Omni	Qwen2.5 Omni 7B	Emu3 8B	Janus-Pro 7B	X-Omni 7B	Step-Audio2 Mini 8B	Qwen2-Audio 7B Instruct	Audio Flamingo3 7B
		Text-to-Text ↑							
Korean	KMMLU-pro	**64.9**	31.1	18.7	16.4	17.7	38.6	23.8	31.8
	HAERAE	**75.3**	51.0	20.0	14.9	21.3	59.9	40.8	46.0
	KoBALT	**27.7**	17.7	10.3	5.9	9.6	17.1	8.9	12.9
	Flores+ (En→Ko)	**29.2**	23.7	0.5	5.7	–	22.7	19.2	10.5
English	MMLU	**75.7**	71.6	32.4	48.2	40.0	64.9	42.9	58.7
	MMLU-pro	**54.2**	50.5	10.8	20.0	19.3	40.2	17.5	30.8
	GSM8K	**87.3**	87.0	3.0	43.5	61.7	75.3	39.5	62.2
	Flores+ (Ko→En)	27.8	**28.6**	0.9	12.1	–	28.2	21.6	18.4
		Vision-to-Text ↑							
Korean	KoNET	**33.0**	14.7	0.6	0.3	11.3	–	–	–
	K-MMBench	**80.2**	76.5	15.7	36.3	38.9	–	–	–
	K-DTCBench	78.8	**88.8**	31.7	29.2	48.3	–	–	–
English	SEED-IMG	**80.3**	77.0	69.0	72.4	74.0	–	–	–
	LLaVA-W	**93.8**	88.5	51.0	78.2	74.2	–	–	–
	TextVQA	80.3	**84.4**	62.9	58.7	77.5	–	–	–
	DocVQA	90.7	**94.9**	74.1	43.4	88.1	–	–	–
		Text-to-Vision ↑							
English	GenEval	0.64	–	0.39	**0.78**	0.67	–	–	–
	ImgEdit	**3.83**	–	–	1.28	1.30	–	–	–
		Speech-to-Text (WER ↓)							
Korean	KsponSpeech-c	**28.74**	34.96	–	–	–	73.20	54.30	–
	KsponSpeech-o	**33.09**	36.76	–	–	–	83.92	52.59	–
	Fleurs-ko	**15.33**	16.23	–	–	–	46.20	36.86	–
English	LibriSpeech-c	1.93	4.13	–	–	–	11.34	3.56	**1.41**
	LibriSpeech-o	4.47	5.67	–	–	–	15.57	6.15	**3.02**
	Fleurs-en	7.00	5.53	–	–	–	15.20	6.42	**3.99**
		Audio-to-Text (SPIDEr ↑)							
English	Clotho-v1	0.259	0.051	–	–	–	0.238	0.138	**0.296**
		Speech-to-Speech (ASR-BLEU ↑)							
En → Ko	Fleurs-en2ko	**24.70**	0.00	–	–	–	0.09	–	–
Ko → En	Fleurs-ko2en	**22.91**	17.76	–	–	–	14.96	–	–

Table 1: Unified Benchmark Results across Text, Vision, and Audio Modalities. For speech-to-text benchmarks, both clean (c) and other (o) splits are used. The symbol '-' indicates that the model does not support the corresponding modality or task.

Accordingly, Qwen2.5 Omni is used as the primary comprehensive baseline throughout the evaluation.

Models such as Emu3 8B (Wang et al., 2024), Janus-Pro 7B (Chen et al., 2025b), and X-Omni 7B (Geng et al., 2025) are text–vision–centric multimodal models. While they support vision–language understanding and generation, they do not provide audio-related capabilities. Therefore, these models are included as baselines only for text and vision benchmarks.

For audio-related tasks, we include Step-Audio2-Mini-8B (Wu et al., 2025) and Qwen2-Audio-7B Instruct (Chu et al., 2024) models as baselines. These models can be evaluated on all audio benchmarks. In addition, the Audio Flamingo3 7B (Ghosh et al., 2025) is selected due to its reported strengths in English-centric audio and speech-to-text tasks.

Overall, we carefully align each baseline model with the modalities it supports, and apply this principle consistently across both vision and audio evaluations. This evaluation setup allows us to analyze the performance of OMNI as a general-purpose Omni model under realistic and well-defined comparison settings, without relying on unsupported assumptions

5.1 Text-only Results

We evaluate the text-to-text performance of OMNI on both Korean and English benchmarks, and the results are reported in Table 1.

Korean Text-to-Text benchmarks. Across Korean benchmarks, OMNI shows a clear performance advantage over all comparison models. On KMMLU-Pro (Hong et al., 2025), HAERAE-1.0 (Son et al., 2024), and KoBALT (Shin et al., 2025), OMNI outperforms other large-scale models by a large margin. These results indicate that OMNI effectively learns Korean-based multi-domain knowledge and reasoning abilities, and that the Korean-focused data composition and training strategy directly contributes to performance gains.

English Text-to-Text benchmarks. OMNI also achieves strong results on English benchmarks, outperforming comparison models across all evaluation metrics. The MMLU benchmark (Hendrycks et al., 2021) series, including MMLU and MMLU-Pro, evaluates broad English multi-task knowledge and reasoning ability. OMNI consistently achieves top-level performance on all benchmarks in this series. In addition, OMNI attains the highest score on the GSM8K (Cobbe et al., 2021) math reasoning benchmark, showing that numerical reasoning and step-by-step problem-solving abilities are also well learned.

Translation benchmarks. We evaluate 1-shot translation performance on the Flores+ benchmark (NLLB Team et al., 2024) using the BLEU metric for the English-Korean pair. For the English-to-Korean direction, we apply Ko-Mecab pre-tokenization to the generated text to compute the scores. As shown in Table 1, OMNI achieves the best performance in English-to-Korean translation and delivers performance comparable to the top-performing models in Korean-to-English translation. These results demonstrate that OMNI has strong cross-lingual capabilities between Korean and English compared to other baseline models.

Overall, OMNI demonstrates strong text-to-text performance on both Korean and English benchmarks compared to existing models. Notably, the performance gap is larger on Korean benchmarks than on English benchmarks, which shows that OMNI provides stable and robust text understanding and reasoning performance in the Korean language setting.

5.2 Vision & Text Results

We evaluate the visual–language understanding and visual generation capabilities of OMNI using a diverse set of public vision benchmarks. The evaluation covers Vision-to-Text tasks, which include image-based question answering and visual reasoning, as well as Text-to-Vision tasks, which focus on text-conditioned image generation and editing. The quantitative results are summarized in Table 1. Furthermore, we extend our evaluation to the temporal domain by assessing the model's video understanding performance on both public benchmarks and specialized internal datasets. These assessments, detailed in the following paragraph, underscore OMNI's capacity for high-fidelity reasoning across the full visual spectrum.

Vision-to-Text Benchmarks. On Korean Vision-to-Text benchmarks, OMNI achieves the best performance on KoNET (Park and Kim, 2025) and K-MMBench (Ju et al., 2024), and the second-best performance on K-DTCBench (Ju et al., 2024). KoNET is a multimodal visual–language reasoning benchmark constructed from a broad range of Korean national educational tests, spanning elementary, middle, high school, and college-level curricula. It evaluates not only the integration of visual and textual information but also diverse forms of Korean linguistic knowledge and educational reasoning across subjects and difficulty levels. The large performance margin observed on KoNET suggests that the Korean-focused data composition and training strategy of OMNI are effective for robust visual–language understanding in the Korean educational context.

OMNI also shows strong performance on English Vision-to-Text benchmarks. It achieves the best results on SEED-IMG (Li et al., 2024a) and LLaVA-W (Liu et al., 2023b), and the second-best results on TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021). SEED-IMG and LLaVA-W evaluate general visual question answering and holistic multimodal reasoning, and the strong performance on these benchmarks demonstrates that OMNI effectively aligns visual representations with textual semantics. On benchmarks that emphasize diagram- and document-level understanding, such as AI2D, DocVQA, and ChartQA, OMNI performs slightly below the top model but consistently remains among the top-performing systems, indicating stable and robust visual–language understanding.

Language	Benchmark	HyperCLOVA X 8B Omni	GPT-4V	Qwen2.5 Omni 7B	LLaVA-OneVision 7B	LLaVA-NeXT 7B
English	Video-MME	58.2	59.9	**64.3**	57.6	33.8
Korean	NAVER TV Content	**69.7**	50.0	–	–	–

Table 2: Performance comparison on video understanding benchmarks. Video-MME results are reported without subtitles. NAVER TV Content is an internal benchmark consisting of real-world Korean video content to evaluate temporal and cultural context understanding.

Text-to-Vision Benchmarks. For Text-to-Vision tasks, we evaluate image generation quality and the accuracy of the text-conditioning capability. OMNI achieves the best performance on ImgEdit (Ye et al., 2025), which focuses on image editing under text constraints, and ranks third on GenEval (Ghosh et al., 2023), which evaluates general text-to-image generation quality. These results suggest that OMNI is particularly strong in preserving semantic intent while performing localized image edits.

An important observation is that OMNI delivers stable performance across both Vision-to-Text and Text-to-Vision tasks while consistently supporting bidirectional text and vision inputs and outputs. In contrast, Qwen2.5 Omni 7B, which performs well on Vision-to-Text benchmarks, does not support Text-to-Vision generation. Janus-Pro and X-Omni support bidirectional text–vision interaction similar to OMNI, but show noticeably lower performance across benchmarks.

In addition to quantitative evaluations, we provide qualitative examples to illustrate the text-to-vision capabilities of OMNI. Figure 6 demonstrates that the model generates semantically consistent images from prompts expressed in different languages (English and Korean), indicating robust cross-lingual alignment in text-to-image generation. Figure 7 further highlights the model's ability to incorporate Korean cultural attributes into generated images, reflecting effective grounding in culturally specific visual concepts. Finally, Figure 8 showcases the model's image editing abilities, including style change, object removal, and background replacement, which qualitatively supports the strong performance observed on the ImgEdit benchmark.

Furthermore, OMNI supports not only text and vision but also audio inputs and outputs within a single unified model. This broader multimodal capability provides an additional advantage beyond visual–language tasks and enables more complex multimodal application scenarios. The evaluation of audio-related tasks is presented in the following Section 5.3.

Video Benchmarks. As an omnimodal model built on a unified framework, OMNI natively supports video understanding through its integrated vision–language processing pipeline. We evaluate our model on two distinct benchmarks: Video-MME, a comprehensive multi-modal video evaluation suite, and an internal benchmark assessing NAVER TV [3] comprehension (NAVER Cloud, 2025), designed to assess the comprehension of real-world Korean video contents.

The quantitative results are summarized in Table 2. In evaluations on Video-MME (Fu et al., 2025) (conducted without subtitles), OMNI achieves a score of 58.2. While this is lower than the 64.3 of Qwen2.5 Omni 7B, it remains highly competitive for an 8B-scale model, comparable to GPT-4V (OpenAI, 2023) and exceeding LLaVA-NeXT (Zhang et al., 2024) and LLaVA-OneVision(Li et al., 2025). Furthermore, on the NAVER TV benchmark NAVER Cloud (2025), OMNI scores 69.7, which is a substantial improvement over GPT-4V's 50.0. These results indicate that the model's joint training on interleaved Korean-centric multimodal data effectively internalizes the complex temporal and cultural nuances required for specialized video understanding applications.

5.3 Audio & Text Results

We evaluate the audio-related performance of OMNI across automatic speech recognition, speech translation, audio captioning, and text-to-speech tasks. Quantitative evaluations based on public benchmarks are reported in Table 1, while human evaluations of speech synthesis quality under real-world commercial settings are presented in Table 3.

---

[3] https://tv.naver.com/

Generate a photo of a tall tree in the background with a yellow car parked directly in front of it

배경에 우뚝 선 큰 나무와 그 앞에 바로 세워진 노란색 자동차를 담은 사진을 생성해줘

Generate a photo of a dog sitting on the right side and a cat sitting on the left side

오른쪽에 강아지, 왼쪽에 고양이가 앉아있는 이미지를 만들어줘

A close-up of an elderly man with deep wrinkles and kind eyes

깊은 주름과 따뜻한 눈빛을 가진 노인의 클로즈업 사진을 생성해줘

Figure 6: Consistency of generated images from the same semantics, different languages (English and Korean).

Figure 7: Generated images incorporating Korean cultural attributes.

Figure 8: Image editing ability of our model. Input images are marked with green borders. Style change, object removal, background replacement are shown, respectively.

13

Language	HyperCLOVA X 8B Omni	Qwen3-Omni-30B-A3B	Gemini-2.5 flash	ElevenLabs v3	GPT-4o-mini-tts
Text-to-Speech (MOS ↑)					
English	3.94 (± 0.14)	3.96 (± 0.15)	**4.44** (± 0.14)	4.11 (± 0.15)	4.08 (± 0.14)
Korean	**4.22** (± 0.11)	3.40 (± 0.12)	4.20 (± 0.12)	4.05 (± 0.12)	3.43 (± 0.12)

Table 3: Human Evaluation Results on Text-to-Speech Benchmark. A total of 30 participants evaluated 20 samples per model across 5 models, resulting in 100 evaluations per listener. Values in parentheses represent the 95% confidence intervals.

Speech/Audio-to-Text Benchmarks. On ASR benchmarks, OMNI shows competitive performance in both English and Korean. On Korean ASR benchmarks, including KsponSpeech (Bang et al., 2020) and Fleurs-ko (Conneau et al., 2022), OMNI achieves state-of-the-art word error rates (WER), establishing strong recognition performance for Korean speech. On English datasets such as LibriSpeech (Panayotov et al., 2015) and Fleurs (Conneau et al., 2022), OMNI performs slightly below Audio Flamingo3 7B, but remains competitive with other English speech models.

For audio-to-text tasks, OMNI also achieves the second-highest SPIDEr score on the Clotho-v1 (Drossos et al., 2019) audio captioning benchmark, following only Audio Flamingo3 7B, which indicates its ability to effectively summarize acoustic events and semantic information in text form. We attribute part of these gains to the unified multimodal architecture, which supports generalization across ASR and audio-to-text generation tasks.

Speech-to-Speech Benchmarks. We conduct a Speech-to-Speech (S2S) evaluation using a speech-to-speech translation (S2ST) benchmark to evaluate the model's ability to directly convert spoken language from a source tongue into spoken language in a target tongue while maintaining semantic integrity and naturalness. Unlike traditional cascaded systems, OMNI aims to streamline this process, thereby reducing latency and potential error propagation. To rigorously assess these capabilities, we conducted cross-lingual translation tasks between English and Korean using a curated dataset of 270 translation pairs for both En → Ko and Ko → En directions.

To quantitatively assess the translation accuracy of the generated audio, we utilize the ASR-BLEU metric. This methodology involves transcribing the model's speech output into text, which is then compared against the ground-truth reference via the BLEU score. To ensure a fair and consistent comparison, we employed `gpt-4o-mini-transcribe` as the unified ASR engine for all candidate models. BLEU scores were evaluated using sacrebleu (Post, 2018). For English-to-Korean translation evaluation, we applied Ko-Mecab pre-tokenization before score computation. Our results demonstrate that OMNI shows superior performance on speech translation tasks; specifically, it achieves the highest performance in both English-to-Korean (En → Ko) and Korean-to-English (Ko → En) translation directions among all evaluated models.

Text-to-Speech Human Evaluation. To evaluate the performance of Text-to-Speech capabilities, we conducted a Mean Opinion Score (MOS) test focusing on the naturalness of the synthesized speech. A total of 30 human listeners participated in the evaluation. The test set comprised 20 distinct utterances, consisting of 10 English sentences and 10 Korean sentences, to assess the model's proficiency across different linguistic contexts.

The evaluation primarily focused on how closely the synthesized speech resembles human-like pronunciation, intonation, and rhythm. Participants were instructed to rate each audio sample on a 5-point Likert scale, where a score of 1 indicates *"Bad: Due to severe artifacts and lack of naturalness, the audio is nearly unintelligible."* and 5 indicates *"Excellent: The speech is nearly indistinguishable from a real human voice, with natural pronunciation, intonation, and rhythm.".* Details of the evaluation protocol, participant setup, and scoring criteria are provided in Appendix B.

Table 3 reports the results of human evaluations (MOS) conducted on an internal dataset, comparing OMNI with widely used commercial text-to-speech systems under real-world service conditions. OMNI achieves competitive MOS scores in both English and Korean in terms of naturalness and pronunciation clarity. In particular, OMNI receives higher scores than comparison models for Korean text-to-speech.

Overall, OMNI achieves balanced performance across a wide range of audio tasks, including speech recognition, audio understanding, speech translation, and text-to-speech. OMNI supports a unified interface in which both inputs and outputs are represented as either audio or text. This design allows the model to be easily extended to multiple audio-related tasks without requiring task-specific model architectures. As a result, heterogeneous tasks such as speech recognition, audio understanding, speech translation, and speech synthesis can be handled in a consistent manner within a single model.

In addition to quantitative evaluations on public benchmarks, human evaluation results conducted under real-world commercial settings show that OMNI achieves quality comparable to existing commercial speech models. In particular, the human evaluation results for text-to-speech indicate that OMNI provides audio quality suitable not only for research settings but also for practical service environments. These results suggest that OMNI, as a general-purpose Omni model, can be effectively applied to a wide range of audio-centric application scenarios.

6 Conclusion

In this work, we presented HyperCLOVA X 8B Omni, the first omnimodal model in the Hyper-CLOVA X family that supports text, audio and vision modalities as both inputs and outputs. HyperCLOVA X 8B Omni is trained with a unified autoregressive objective that extends next-token prediction beyond text by incorporating discrete vision and audio codebook entries as additional vocabulary items in interleaved sequences. On top of this symbolic interface, continuous vision/audio encoders inject richer perceptual embeddings projected into the same backbone space, while modality-specific decoders translate the shared sequence representations back to pixels and waveforms, compensating for information lost in semantic tokenization.

Empirical evaluations show that HyperCLOVA X 8B Omni achieves competitive performance against comparably sized models across diverse combinations and input and output modalities: text-to-text, vision-to-text, text-to-vision, speech-to-text, audio-to-text, speech-to-speech, and text-to-speech. We expect that its open-sourcing will benefit researchers and practitioners seeking a compact yet versatile model.

The 8B-scale HyperCLOVA X 8B Omni model serves as the first *pathfinding* point of the design in which a unified auto-regressive backbone supports both interleaved multimodal understanding and any-to-any generation when paired with modality-specific encoders and decoders. While the performance of HyperCLOVA X 8B Omni is strong relative to its size, we anticipate that increasing its size will yield considerable performance gains. A larger and more advanced variant would be particularly valuable in situations with sufficient computational resources where higher performance is desired. Therefore, scaling up the model represents an important avenue for our future research toward developing a collection of robust omnimodal models that can accommodate the needs and restrictions of diverse scenarios.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10(19).

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025a. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. 2024. Autoregressive Video Generation without Vector Quantization. ArXiv:2412.14169 [cs].

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pages 3830–3834.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *CVPR*.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, Linus, Di Wang, and Jie Jiang. 2025. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *CoRR*, abs/2507.22058.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jiaming Han, Hao Chen, Yang Zhao, Hanyu Wang, Qi Zhao, Ziyan Yang, Hao He, Xiangyu Yue, and Lu Jiang. 2025. Vision as a Dialect: Unifying Visual Understanding and Generation via Text-Aligned Representations. ArXiv:2506.18898 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. 2025. From kmmlu-redux to kmmlu-pro: A professional korean benchmark suite for llm evaluation.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

NAVER Cloud HyperCLOVA X Team. 2025. Hyperclova x think technical report.

Jeongho Ju, Daeyoung Kim, SunYoung Park, and Youngjune Kim. 2024. Varco-vision: Expanding frontiers in korean vision-language models.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. 2024. Guiding a Diffusion Model with a Bad Version of Itself. ArXiv:2406.02507 [cs, stat].

Geewook Kim and Minjoon Seo. 2025a. Does Audio Matter for Modern Video-LLMs and Their Benchmarks?

Geewook Kim and Minjoon Seo. 2025b. MambaMia: A State-Space-Model-Based Compression for Efficient Video Understanding in Large Multimodal Models.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. ArXiv:2506.15742 [cs].

Seongyun Lee, Geewook Kim, Jiyeon Kim, Hyunji Lee, Hoyeon Chang, Sue Hyun Park, and Minjoon Seo. 2025. How Does Vision-Language Adaptation Impact the Safety of Vision Language Models? In *The Thirteenth International Conference on Learning Representations*.

B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. 2025. LLaVA-oneVision: Easy Visual Task Transfer. *TMLR*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2024b. Autoregressive Image Generation without Vector Quantization. ArXiv:2406.11838 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Wenzhuo Liu, Fei Zhu, Haiyang Guo, Longhui Wei, and Cheng-Lin Liu. 2025. Llava-c: Continual improved visual instruction tuning.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2024. Finite scalar quantization: Vq-vae made simple. In *International Conference on Representation Learning*, volume 2024, pages 51772–51783.

NAVER Cloud. 2025. HyperCLOVA X Video: Seeing through motion. `https://clova.ai/en/tech-blog/hyperclova-x-video-seeing-through-motion`. Accessed: 2025-12-31.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OpenAI. 2023. GPT-4V(ision) System Card. System card, OpenAI.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Sanghee Park and Geewook Kim. 2025. Evaluating multimodal generative AI with Korean educational standards. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 671–688, Albuquerque, New Mexico. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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.

Hyopil Shin, Sangah Lee, Dongjun Jang, Wooseok Song, Jaeyoon Kim, Chaeyoung Oh, Hyemi Jo, Youngchae Ahn, Sihyun Oh, Hyohyeong Chang, Sunkyoung Kim, and Jinsik Lee. 2025. Kobalt: Korean benchmark for advanced linguistic tasks.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. ArXiv:2502.14786 [cs].

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Hao Nie, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiancheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, Yuxiang Zhang, Zidong Yang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, Jiansheng Chen, Jing Li, Xiangyu Zhang, and Yibo Zhu. 2025. Step-audio 2 technical report.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025. Imgedit: A unified image editing dataset and benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning (Proceedings Track)*.

Y. Zhang, B. Li, H. Liu, Y. J. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. 2025. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*.

Contributions and Acknowledgments

*Within each role, **names are listed in alphabetical order** by first name, followed by the last name.*

Technical Writing

Cheonbok Park
Dongyoon Han
Geewook Kim
Hwiyeol Jo
Jeonghoon Kim
Jin-Hwa Kim
Jiseob Kim
Joosung Lee
Sangdoo Yun
Sanghyuk Choi
Sungwook Jeon
Taeho Kil
Yoonsik Kim

Model Research and Training

Bado Lee
Cheonbok Park
Daehee Kim
Geewook Kim
Gichang Lee
Hangyeol Yu
Heesu Kim
Hodong Lee
Jeonghoon Kim
Jinbae Im
Jinhyeon Kim
Jiseob Kim
Jungwhan Kim
Ka Yeon Song
Kyeongseok Jeong
Moonbin Yim
Nako Sung
Ohsung Kwon
Sang Hee Park
Sanghyuk Choi
Seongjin Shin
Seunggyu Chang
Soyoon Kim
Suk Min Seo
Taeho Kil
Taehwan Yoo
Yeontaek Oh
Yoonsik Kim

Model Evaluation and Analysis

Daehee Kim
Dongjin Lee
Gayoung Lee
Hagyeong Lee
Hangyeol Yu

Heesu Kim
Hwiyeol Jo
Hyunhoon Jung
Hyunsoo Ha
Jeonghyun Lee
Jieun Lee
Jieun Shin
Jonghak Kim
Joosung Lee
Jungwhan Kim
Ka Yeon Song
Kiyoon Moon
Minkyoung Kim
Munhyong Kim
MyungIn You
Saerim Cho
Soyoon Kim
Suk Min Seo
Taemin Lim
Taeyong Kim
Woobin Choi
Yehbin Lee
Yelim Jeong
Yeonsun Ahn
Yeontaek Oh
Youngjin Kwon
Zoo Hyun Lee

Data

Byoungeul Kim
Byungwook Lee
Chan-Ho Song
Chansong Jo
Chiheon Ham
Donghyeon Ko
Dongjin Lee
Eunwoo Song
Hanbyul Kim
Hoyeon Lee
Hyun-Wook Yoon
Hyunsoo Ha
Injae Lee
Jaehong Lee
Jaemin Han
Jaeuk Byun
Jahyeong Lee
Jeongmin Liu
Jieun Lee
Jin-Seob Kim
Jinbae Im
Jingu Kang
Jisung Wang
Jong-Hwan Kim

Juncheol Kim
Kang Lae Jung
Kiyoon Moon
Kyeongseok Jeong
Min Young Lee
Min-Seok Choi
Minjae Lee
Minkyoung Kim
Minseong Choi
Moonbin Yim
Munhyong Kim
MyungIn You
Ohsung Kwon
Sangkil Lee
Seongjin Shin
Seunggyu Chang
Shinyoung Joo
Soo-Whan Chung
Sookyo In
Soyeon Choe
Suhyeon Oh
Sung Ae Lee
Sungju Kim
Sungjun Choi
Sunmi Rim
Taehong Min
Taehwan Yoo
Taeyong Kim
Yeguk Jin
Yehbin Lee
You Jin Kim
Youna Ji
Youngjun Kim
Youngki Hong

Model Serving and Inference

Hanbae Seo
Hodong Lee
Hyunjoon Jeong
Jaeeun Kil
Jaegwang Lee
Jeongtae Lee
Jinhyeon Kim
Joonghoon Kim
Junhee Yoo
Minjung Jo
Minsub Kim
Sang Hee Park
Sungjae Lee
Sungju Kim
Yeonsun Ahn

Model Planning

Gayoung Lee
Hagyeong Lee
Hyunhoon Jung
Jeonghyun Lee

Jieun Shin
Jonghak Kim
Saerim Cho
Taemin Lim
Woobin Choi
Yelim Jeong
Youngjin Kwon
Zoo Hyun Lee

Business and Brand Strategy

Dukmin Jung
Kyungmin Lee
Hyojin Park
Sujin Roh
Misuk Park

Residency Program

Bumkyu Park
Byung Hyun Lee
Doohyuk Jang
Geeho Kim
Hyewon Jeon
Hyunbin Jin
Hyungwook Choi
Ijun Jang
Inju Ha
Jewon Yeom
Jihwan Kim
Jihwan Kwak
Joonki Min
Juan Yeo
Junbeom Kim
Junyeob Kim
Kunhee Kim
Kyubyung Chae
Kyudan Jung
Minha Jhang
Sangyoon Lee
Sehyun Lee
Seunghee Kim
Song-ha Jo
Suho Ryu
Yokyung Lee

Internships

Dong-Jae Lee
Jihwan Moon
Jinho Heo
Jisu Jeon
Minsik Choi
Seulbi Lee
Singon Kim
Sumin Cho
Woojin Chung

Appendix

A Implementation Details of the Vision Decoder

Training Curriculum. The training of the vision decoder is organized into four sequential phases to stabilize optimization and ensure high-fidelity synthesis across various scales:

1. **Low-resolution crop training**: Initial training at 0.25 Megapixels ($\sim$512$\times$512) focusing on cropped regions.
2. **Full-resolution crop training**: Training at 0.6 Megapixels ($\sim$768$\times$768) using 1:2 cropped image regions.
3. **Full-resolution full training**: Training on the entire image area at 0.6 Megapixels.
4. **Refinement stage**: A final stage with a reduced learning rate to stabilize the model and polish fine visual details.

Crop training is particularly effective in our framework because each cropped region is tightly coupled to its corresponding vision token grid. This allows the diffusion model to unambiguously reference the correct conditioning tokens even when operating on partial image patches.

Inference and Autoguidance. During inference, we adopt autoguidance (Karras et al., 2024), which leads to a substantial improvement in visual quality (see Figure 9). Given that our decoder relies on dense semantic conditioning, the model can occasionally exhibit overly local patterns, resulting in degraded small-scale textures. Autoguidance amplifies the conditioning signal, helping the decoder preserve fine structures—such as typography and intricate patterns—more consistently across the generated output. To apply autoguidance, we train a smaller model (470M) briefly ($\sim$1/20 steps of the main model) and use guidance scale of 1.75.

Original　　　　　Reconstruction　　　　　Reconstruction w/ Autoguidance

Figure 9: Autoguidance (Karras et al., 2024) significantly improves the overall quality of our vision decoder (see the enhancement in hockey helmets, typography, and fingers).

B Text-to-Speech Evaluations

Participants were presented with synthesized speech from five TTS systems, anonymized and randomly shuffled to prevent model identification. For each text script, annotators listened to all five audio samples and assigned Mean Opinion Scores (1–5) based on fluency and pronunciation clarity. The quality of the synthesized speech was evaluated using a 5-point Likert scale, measuring naturalness and intelligibility. Each score corresponds to the following descriptive criteria:

1 (Bad): Due to severe artifacts and lack of naturalness, the audio is nearly unintelligible.

2 (Poor): The audio sounds heavily robotic, and issues with pronunciation or intonation make it uncomfortable to listen to.

3 (Fair): While there is a noticeable awkwardness, there are no major issues in understanding the content.

4 (Good): The speech is generally natural, though slight awkwardness or a robotic feel may be heard occasionally.

5 (Excellent): The speech is nearly indistinguishable from a real human voice, with natural pronunciation, intonation, and rhythm.

Figure 10: Annotation interface used for MOS evaluation.

C Examples of User-Intent Aware Reasoning

As discussed in Section 4.1, OMNI incorporates an internal reasoning process to classify user intent and determine the optimal modality-specific modules for response generation. This section provides three illustrative examples of the internalized reasoning (the `<think>` block), demonstrating the model's ability to parse complex instructions and manage cross-modality constraints.

C.1 Visual Question Answering

```
<|im_start|>assistant
<think>
This is a VQA task requiring high-level visual reasoning.
The input modality consists of an image and a textual query.
I will leverage the continuous vision encoder to extract
perceptual features and map them to linguistic
representations, providing an objective and detailed
factual response.
</think>

[Actual Text Response]
<|im_end|>
```

C.2 Image Editing

```
<|im_start|>assistant
<think>
This is a request to edit specific visual attributes of
the input image. The input modality is image-text
interleaved. I will utilize the synthesis functions of
the diffusion-based vision decoder to modify the requested
elements while maintaining the semantic consistency of
unchanged objects.
</think>

[Generated Image Output]
<|im_end|>
```

C.3 Audio-Visual Understanding with Speech Response

```
<|im_start|>assistant
<think>
This is a complex omnimodal task involving an
audio-based user query regarding visual content.
I will first process the continuous audio
embedding to parse the user's intent, analyze
the image features via the vision encoder,
and finally generate a synchronized speech response
using discrete audio tokens for the neural audio decoder.
</think>

[Actual Multimodal/Audio Response]
<|im_end|>
```