OMNIVIVO: TOWARDS UNIFIED MULTIMODAL GENERATIVE MODELING FOR SIMULTANEOUS LANGUAGE-GUIDED SPEECH AND IMAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent large language models (LLMs) based on autoregressive (AR) next-token prediction have achieved remarkable success in natural language generation and are rapidly expanding to image and speech synthesis. Yet most current approaches still treat these modalities in isolation—training independent models or loosely coupling multiple generators. Even recent omni-models such as UGen and Qwen2.5-Omni mainly address understanding tasks or text–image generation and do not provide a single AR backbone capable of simultaneously producing high-fidelity images and natural speech. Inspired by the human brain's capability to imagine and speak simultaneously, we propose **OmniVIVO**, a unified AR approach for modeling **vi**sual and **vo**ice modalities together, capable of generating high-fidelity images and natural speech in parallel from a single text input. Our OmniVIVO integrates a state-of-the-art AR image generator with a novel lightweight speech decoder, enabling the first unified approach for the concurrent generation of natural speech and high-fidelity images. By sharing representations across modalities within a single transformer backbone, the model learns a rich multimodal space that enables tighter semantic alignment and more efficient joint generation than existing multi-model pipelines. Through a unified backbone, OmniVIVO produces speech with high perceptual quality and naturalness, surpassing comparably sized text-to-speech (TTS) systems and being on par with state-of-the-art systems like Cosyvoice2 and VITS, while maintaining high-fidelity image generation. To quantify contextual understanding across modalities, we propose a multimodal ranking metric spanning text, speech, and images, demonstrating that OmniVIVO's bi-modal outputs are effective in information acquisition. We construct **VIVOGen**, a high-quality tri-modal text–image–speech dataset that leverages OmniVIVO's multimodal outputs, providing a valuable resource for research in multimodal learning and applications in education and language acquisition, which we will publicly release.

1 INTRODUCTION

Recent advancements in artificial intelligence (AI), driven by large language models (LLMs) such as GPT-4 (OpenAI, 2023), LLaMA-3 (Dubey et al., 2024), and Qwen2.5 (Yang et al., 2024), have significantly expanded the boundaries of AI, making powerful tools more accessible to a broader audience. Recently, research has expanded LLM applications to new areas such as visual generation and speech synthesis. Noteworthy achievements have been made in text-to-image (T2I) generation, where LLMs produce high-fidelity images, rivaling the performance of task-specific models such as diffusion-based architectures (Sun et al., 2024; Tian et al.). In the domain of speech synthesis, LLMs have also demonstrated impressive progress in text-to-speech (TTS) systems, generating highly natural-sounding speech (Gao et al., 2025) while showcasing advanced capabilities such as prompt-based control and zero-shot learning (Du et al., 2024). Furthermore, recent studies have explored multimodal LLMs such as Qwen2.5-Omni (Jin Xu, 2025), UGen (Tang et al., 2025), which are capable of processing inputs from diverse modalities—images, text, and speech—and producing text or speech output.

While existing research in multimodal AI has primarily concentrated on processing and understanding multimodal inputs, there remains a significant gap in the generation of multimodal outputs. In particular, producing outputs such as speech and images is a fundamental step toward emulating the human capacity to imagine and vocalize simultaneously. To bridge this gap, we propose a novel approach that enables the concurrent generation of high-fidelity images and natural-sounding speech from a single LLM architecture. This method not only mimics human-like cognition but also enhances the flexibility and scalability of multimodal systems, paving the way for more immersive human-AI interactions, including dynamic storytelling applications.

We present **OmniVIVO**, the first model capable of generating both **vi**sual and **vo**ice outputs within a single LLM backbone. Unlike prior works that treat image and speech generation as independent tasks, OmniVIVO integrates a state-of-the-art (SOTA) pretrained image generator with a novel lightweight yet effective speech generation module, enabling parallel production of both high-quality images and speech. Through this unified design, experiments show that OmniVIVO not only surpasses TTS baselines of comparable scale in speech quality but also demonstrates strong cross-modal coherence, marking a step toward truly multimodal generative intelligence.

To systematically evaluate the multimodal generation, we further introduce a multimodal ranking metric that complements conventional subjective and objective evaluations used in speech generation, providing a more comprehensive assessment of OmniVIVO's outputs. Experimental findings confirm that OmniVIVO produces multimodal outputs that are highly effective for information acquisition, highlighting its promise for diverse applications. In addition, we release **VIVO-Gen**, a high-quality tri-modal dataset consisting of language-specific knowledge inputs paired with OmniVIVO-generated image and speech outputs, fostering progress in multimodal education, language acquisition, and related domains.

The key contributions of this work include: (a) we introduce **OmniVIVO**, the first LLM model capable of simultaneously generating high-quality speech and images; (b) we present a new multimodal ranking metric to evaluate the quality of OmniVIVO's outputs and multimodal generation; and (c) we release **VIVOGen**, a high-quality tri-modal text–image–speech dataset designed to advance research in multimodal applications.

2 Related Work

Autoregressive Image Generation: Early advancements in image generation transitioned from GANs (Goodfellow et al., 2020), which achieved high fidelity but struggled with unstable training, to diffusion models (Ho et al., 2020), which have become the dominant approach due to their scalable architectures and stable likelihood-based training. However, diffusion models are computationally expensive during inference due to the iterative denoising process, which has sparked renewed interest in autoregressive (AR) approaches that directly model discrete image tokens in sequence. Initial AR models, such as VQ-VAE (Van Den Oord et al., 2017) and VQGAN (Esser et al.), employed transformer decoders to predict image tokens in a raster-scan order. While these models demonstrated feasibility, they were inefficient and produced spatially unnatural results. The introduction of VAR (Tian et al.), which employed a coarse-to-fine next-scale prediction strategy, improved image quality but still required thousands of tokens per image, resulting in significant computational costs.

Recent advances have redefined AR generation by leveraging the power of LLMs. For instance, LlamaGen (Sun et al., 2024) demonstrates that scaling a vanilla decoder-only LLM to billions of parameters, combined with carefully curated data, enables AR models to match or even surpass diffusion models on ImageNet (Deng et al., 2009). Similarly, Muse (Chang et al.) employs masked AR training with LLM-based techniques to achieve T2I quality on par with diffusion models.

Autoregressive Text-Speech Language Model: Modern TTS systems have recently made a breakthrough by moving from specialized modules such as Tacotron2 (Shen et al., 2018) and FastSpeech2 (Ren et al.) to architectures based on LLMs (Du et al., 2024). This shift enables models to leverage the power of pretrained LLMs trained on massive datasets, thereby improving contextual modeling and producing speech that is natural, expressive, and high-fidelity. Instead of separating linguistic and acoustic processing, LLM-based TTS unifies the workflow by modeling sequences of discrete units (semantic, prosodic, acoustic) quantized from speech signals, thus providing a smoother bridge between text and speech (Wang et al.).

Pioneering models that applied LLMs to speech generation include VALL-E (Wang et al.), which demonstrated zero-shot TTS from just a few seconds of reference audio. More recently, systems such as CosyVoice2 (Du et al., 2024) and Spark-TTS (Wang et al., 2025) have advanced this direction further, not only achieving high synthesis quality but also supporting advanced controllability features such as instruction prompting, zero-shot. These advances mark a paradigm shift from task-specific pipelines toward general-purpose generative models, where a single backbone can flexibly handle speech generation.

Multimodal Large Language Models: Recent advancements in multimodal large language models (MLLMs) have extended the capabilities of text-based models to process different modalities. Systems like Flamingo (Alayrac et al.), Qwen2.5-Omni (Jin Xu, 2025), UGen (Tang et al., 2025) integrate vision and speech with text, enabling models to process diverse inputs. However, these approaches are primarily focused on multimodal understanding, where inputs come from various sources, but the outputs are typically limited to text or speech. Recently, Team (2024) introduced a multimodal large language model (MLLM) capable of bidirectional generation, enabling both text and image processing. This is achieved through the integration of an image tokenizer and a text tokenizer within a unified autoregressive backbone, allowing flexible support for diverse multimodal tasks. Despite these advancements, Team (2024) and related models remain limited, as they do not extend to the generation of speech outputs.

To address this gap, we propose OmniVIVO, the first unified AR model designed to generate both images and speech simultaneously from a single text input. By leveraging a SOTA image generator with a lightweight speech decoder under a shared LLM backbone, OmniVIVO pushes the boundaries of multimodal generation, enabling groundbreaking advancements in fields such as multimodal education, language acquisition, and interactive AI.

3 METHODOLOGY

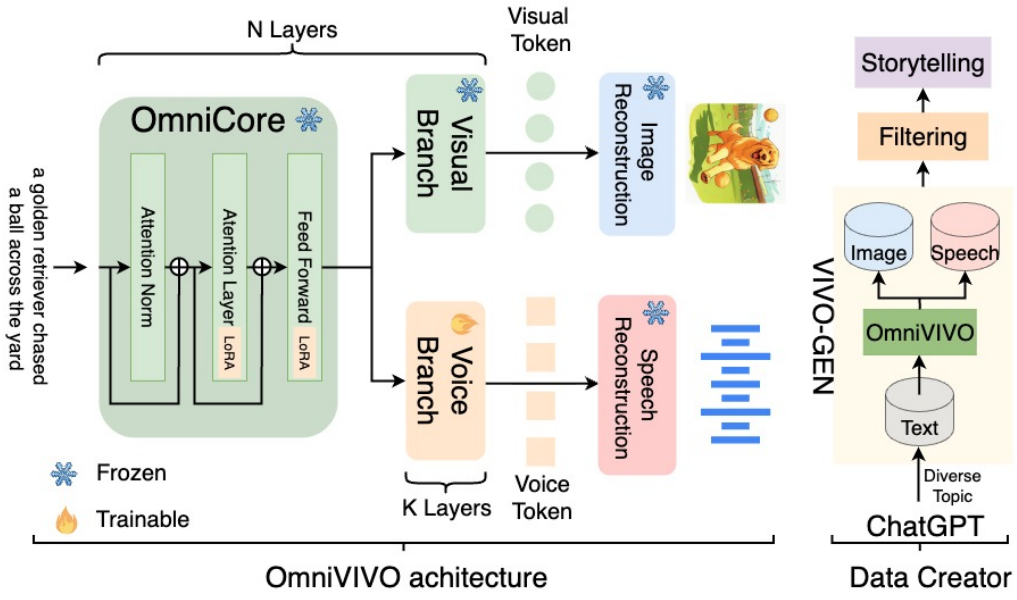

Figure 1: OmniVIVO: the proposed model is capable of generating high-quality images and speech. Additionally, VIVOGen is released to advance multimodal applications like storytelling.

3.1 PROPOSED MODEL: OMNIVIVO

We propose **OmniVIVO**, shown in Figure 1, a multimodal LLM model capable of generating both images and speech from a text input. Our design leverages a pretrained image backbone as a shared

semantic encoder, augmented with lightweight LoRA adapters and an additional speech branch. This approach preserves image generation quality while enabling efficient cross-modal adaptation.

3.1.1 OMNICORE

First, the input text is tokenized by the pretrained Flan-T5 tokenizer (Chung et al.) into a vector x with T token:

$$x = (x_1, \ldots, x_T), \quad x_t \in \mathcal{V}_{\text{text}}.$$

We adopt a pretrained image generator backbone f_θ (LlamaGen (Sun et al., 2024)) as the shared transformer. To enable adaptation without degrading image generation, we freeze all original parameters W_θ and insert low-rank adapters (LoRA Hu et al. (2022)) into each linear projection:

$$W_\beta = W_\theta + \alpha \cdot W_A W_B,$$

where $W_A \in \mathbb{R}^{d_{\text{in}} \times r}$ and $W_B \in \mathbb{R}^{r \times d_{\text{out}}}$ are trainable low-rank matrices, and α is a fixed scaling factor. This yields an adapted backbone f_β that retains the representational capacity of f_θ while gaining flexibility for new modalities.

3.1.2 IMAGE GENERATION

For image generation, the token vector x is embedded and passed through the frozen backbone to obtain the latent feature h_i with d-dimension:

$$h_i = f_\theta(x), \quad h_i \in \mathbb{R}^{T \times d}.$$

The pretrained output head $W_{\bar{\theta}}$ from LlamaGen (Sun et al., 2024) is kept fixed. The conditional distribution over N image tokens $\bar{E}$ is:

$$p_\theta(\bar{e} \mid x) = \prod_{n=1}^{N} \text{Softmax}(W_{\bar{\theta}} h_i[n]) \,.$$

Because neither f_θ nor $W_{\bar{\theta}}$ is updated, OmniVIVO fully preserves the image generation performance of the pretrained model.

3.1.3 SPEECH GENERATION

To extend the backbone for speech generation, we attach a lightweight speech transformer f_ϕ on top of the adapted backbone to obtain the latent feature z_s:

$$h_s = f_\beta(x),$$

$$z_s = f_\phi(h_s), \quad z_s \in \mathbb{R}^{T \times d}.$$

The outputs are projected into M discrete speech tokens $\bar{S}$ through a trainable head $W_{\bar{\phi}}$:

$$p_\phi(\bar{s} \mid x) = \prod_{m=1}^{M} \text{Softmax}\big(W_{\bar{\phi}} z_s[m]\big)$$

Here, $\{W_A, W_B, f_\phi, W_{\bar{\phi}}\}$ are updated during training.

3.1.4 OMNIVIVO TRAINING OBJECTIVE

OmniVIVO is trained solely on speech data using cross-entropy loss:

$$\mathcal{L}_\phi = \mathbb{E}_{(x,\bar{s})} \left[-\sum_{m=1}^{M} \log p_\phi(\bar{s}_m \mid \bar{s}_{<m}, x) \right].$$

3.1.5 OMNIVIVO DESIGN PROPERTIES

The OmniVIVO architecture aims to facilitate three desirable properties:

- **Preservation:** Freezing f_θ and $W_{\bar{\theta}}$ ensures image quality is unaffected.
- **Efficiency:** Only LoRA adapters and the speech branch are trainable, reducing parameter updates.
- **Cross-modal sharing:** The backbone provides a shared semantic space between text, image, and speech.

Together, this lightweight unified design enables OmniVIVO to balance *cross-modal sharing* with *modality-specific specialization*

3.2 SPEECH TOKENIZER AND RECONSTRUCTION

The semantic speech tokenizer (Du et al., 2024) converts the raw speech input S into intermediate feature representations R through a speech encoder, SenseVoice-Large ASR model (An et al., 2024), E_{speech}. The encoder outputs a sequence $R = \{r_1, r_2, \ldots, r_T\}$, where T is the sequence length.

$$R = E_{\text{speech}}(Y)$$

The features R are quantized into discrete values using Finite Scalar Quantization (FSQ) (Mentzer et al., 2024), which maps each feature to a scalar within the range $[-L, L]$, where L is the number of quantization levels:

$$R'_{\text{qt}} = \text{FSQ}(R)$$

Subsequently, discrete speech tokens T_k are computed as:

$$S_k = \sum_{m=0}^{M-1} R'_{\text{qt}[k,m]} \cdot (2L+1)^m$$

Where S_k is the speech token at time step k, which represents the discrete value corresponding to the quantized feature vector at time step k. The summation $\sum_{m=0}^{M-1}$ indicates that the token S_k is generated by summing the quantized feature vector R'_{qt} at time step k and with m dimension.

Finally, **speech reconstruction** is performed in two stages. A flow matching model (Lipman et al.) first maps the quantized representations into a Mel-spectrogram, and a HiFi-GAN vocoder (Kong et al.) subsequently converts this Mel-spectrogram into high-fidelity, natural-sounding waveforms. This process is similarly applied to the speech token from OmniVIVO $\bar{S}$.

3.3 IMAGE TOKENIZER AND RECONSTRUCTION

To transform images into discrete symbols for AR modeling, we employ a VQGAN-based model (Esser et al.) consisting of an encoder, a vector quantizer, and a decoder.

Given an input image $y \in \mathbb{R}^{H\times W\times 3}$, the encoder compresses it into a latent representation

$$f = \text{Encoder}(y) \in \mathbb{R}^{h\times w\times C}, \quad h = H/p, \;\; w = W/p,$$

where p is the downsampling factor and C denotes the feature dimensionality.

The quantizer replaces each latent vector $f(i,j)$ with the closest entry from a learnable codebook $\mathcal{E} = \{e_1, \ldots, e_K\} \subset \mathbb{R}^{K\times C}$, containing K prototype embeddings of dimension C. This assignment is defined as

$$q(i,j) = \arg\min_{k\in\{1,\ldots,K\}} \|f(i,j) - e_k\|_2^2.$$

The resulting discrete index map $q \in \{0, \ldots, K-1\}^{h\times w}$ is subsequently linearized into a sequence of $h \cdot w$ tokens that can be modeled autoregressively.

Finally, **image reconstruction** is performed by mapping the quantized embeddings back into pixel space, similarly to the image token from OmniVIVO $\bar{E}$:

$$\hat{y} = \text{Decoder}\big(e_{q(i,j)}\big).$$

3.4 Data Creation

We introduce VIVOGen, a dataset designed to advance multimodal applications in various domains, including language education and storytelling. The dataset consists of 100 high-fidelity samples of images and speech generated by OmniVIVO. Specifically, we use ChatGPT (OpenAI, 2023) to generate text inputs on diverse topics such as animals, pets, vehicles, nature, and more. Additionally, Whisper-V2 (Radford et al.) is used to remove low-quality speech samples, ensuring that the VIVOGen dataset maintains high intelligibility. Finally, human reviewers are involved in the final filtering stage, retaining only high-quality image and speech pairs.

4 Experimental Setup

Architecture. OmniVIVO unifies visual and speech generation within a single transformer backbone. It leverages a pretrained 36-layer transformer image generator with 20 attention heads and a hidden size of 1280, termed Omni-Core, with an 8-layer speech branch, totaling approximately 1 billion parameters, of which 225 million are trainable.

Adaptation. To enable efficient fine-tuning, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank $r = 16$ and scaling factor $\alpha = 16$, updating task-specific modules in Omni-Core and the speech branch while freezing most Omni-Core weights.

Tokenization: Text inputs are processed using the Flan-T5 tokenizer (Chung et al.) (vocabulary size 32,100). Speech tokens are generated via a pretrained Semantic Speech Tokenizer from CosyVoice2 (Du et al., 2024), and reconstructed using flow-matching models with HiFi-GAN vocoders (Kong et al.). Images are tokenized and reconstructed via the pretrained VQ-VAE from LlamaGen (Sun et al., 2024).

Training: OmniVIVO is fine-tuned on the LibriTTS (Zen et al., 2019) dataset (585 hours, multi-speaker) for 100,000 steps using the AdamW optimizer (Loshchilov & Hutter) ($\beta = (0.9, 0.999)$, weight decay = 0.01, learning rate 1×10^{-4}, constant schedule). Training minimizes a cross-entropy loss with a batch size of 14, applies gradient clipping ($\|g\|_2 \leq 1.0$), and averages the last five checkpoints for stability.

Evaluation To evaluate the proposed OmniVIVO system, we first conduct an ablation study using Word Error Rate (WER) and Character Error Rate (CER) to assess the impact of model depth on intelligibility, using 1,000 test clean samples from LibriTTS (Zen et al., 2019). For this, we use Whisper-Large-V2 (Radford et al.), an Automatic Speech Recognition (ASR) model, to transcribe speech into text. Additionally, we perform subjective evaluations of speech and image quality across four metrics: Naturalness, Intelligibility, Multimodal Coherence, and Multimodal Ranking. Each subjective test involves 15 participants. For speech quality (Naturalness and Intelligibility), we compare OmniVIVO, VITS (Kim et al.), CosyVoice2 (Du et al., 2024), and Ground Truth, with each system contributing 10 samples, resulting in 40 total. For multimodal evaluations (Coherence and Ranking), only OmniVIVO is assessed, with 10 samples per experiment, as no baseline systems are available. Detailed evaluation protocols and criteria are provided in Appendix A.1.

Multimodel Ranking: To address the lack of effective evaluation methods for multimodal generation outputs, we propose a new metric to investigate how different presentation formats influence the ease of acquiring information. We categorize the formats into three levels: **Excellent**, **Acceptable**, and **Less Effective**.

- **Excellent**: Information is conveyed quickly, clearly, and effortlessly.
- **Acceptable**: Information is sufficiently clear, though not optimal.
- **Less Effective**: Information is understandable but lacks clarity and effectiveness.

At each level, the participant will select one of the following formats: (A) Text, (B) Speech, (C) Image, (D) Text + Speech, (E) Text + Image, (F) Speech + Image, or (G) Text + Speech + Image.

Table 1: Comparison of WER and CER for TTS-Baseline and OmniVIVO across Different Model Depths.

Model Depths	TTS-Baseline		OmniVIVO	
	WER↓	CER↓	WER↓	CER↓
2Layer	37.77	26.36	18.72	11.95
4Layer	25.41	17.17	12.50	7.28
6Layer	24.28	16.23	11.51	6.50
8Layer (Proposed)	22.65	15.26	**10.64**	**6.06**
10Layer	22.50	15.42	11.07	6.12

5 RESULTS

5.1 EFFECT OF MODEL DEPTH ON SPEECH QUALITY

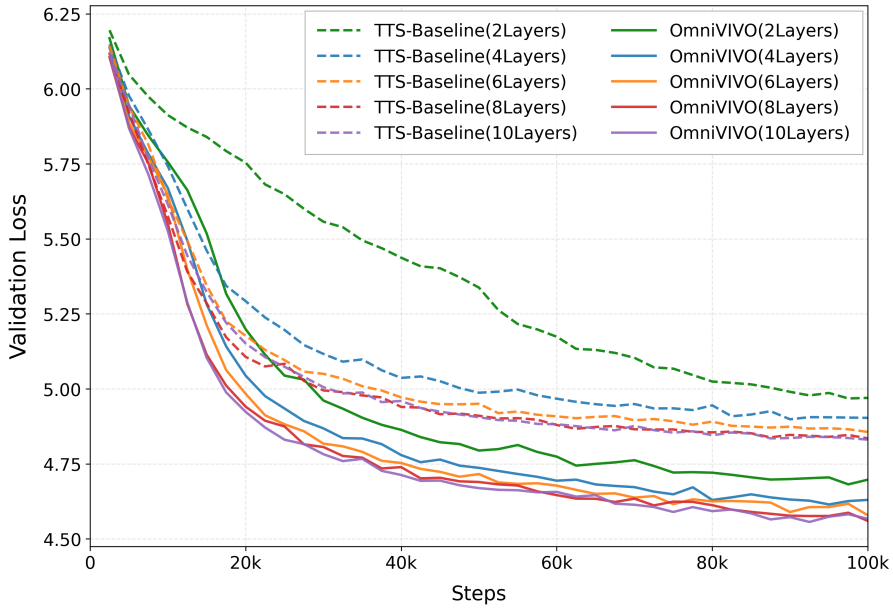

Figure 2: Comparison of validation loss for TTS-Baseline and OmniVIVO on Speech Branch across Different Model Depths.

Table 1 presents the results of a depth ablation study comparing the proposed OmniVIVO architecture against a TTS baseline across five network depths (2, 4, 6, 8, and 10 layers), using the WER and CER metrics. Note that the TTS-Baseline uses the same architecture and parameters as OmniVIVO, but it is trained from scratch independently. We use 1,000 text sentences from the test-clean subset of LibriTTS and employ Whisper V2 to transcribe speech to text. OmniVIVO consistently outperforms the baseline across all depths. For example, at the 8-layer configuration, OmniVIVO achieves a WER of 10.64 and a CER of 6.06, compared to the baseline's WER of 22.65 and CER of 15.26, yielding absolute reductions of 12.01 WER points and 9.20 CER points (relative reductions of 53.0% and 60.3%, respectively). Comparable improvements are observed at other depths, including the 2-layer and 4-layer configurations, highlighting the robustness and efficacy of OmniVIVO's design.

Furthermore, as shown in Figure 2, OmniVIVO consistently achieves lower validation loss (lower is better) compared to the TTS baseline at similar depths, reinforcing the superiority of our proposed approach.

5.2 IMAGE GENERATION QUALITY

Table 2: Inception Score (IS) Comparison for LlamaGen and OmniVIVO

Image Quality	IS Score ↑
LlamaGen	7.78±0.78
OmniVIVO	6.93±0.65

Table 2 presents a comparison of the Inception Score (IS) between OmniVIVO and LlamaGen, evaluated on 250 images generated from text prompts produced by ChatGPT (OpenAI, 2023) on diverse topics. OmniVIVO, which extends the LlamaGen architecture with modifications to support speech generation, achieves an IS of 6.93 ± 0.65, compared to LlamaGen's 7.78 ± 0.78. The comparable scores indicate that OmniVIVO preserves strong image generation quality while additionally enabling speech synthesis, demonstrating its effectiveness for multimodal generation.

5.3 SUBJECTIVE SPEECH QUALITY ASSESSMENT

Table 3: Mean Opinion Scores (MOS) with 95% confidence intervals.

Method	Naturalness↑	Intelligibility↑
GroundTruth	4.25 ± 0.14	4.38 ± 0.13
VITS	3.86 ± 0.15	4.31 ± 0.12
CosyVoice2	3.70 ± 0.17	**4.33 ± 0.12**
OmniVIVO	**3.94 ± 0.16**	4.19 ± 0.14

In this experiment, we evaluate the Mean Opinion Scores (MOS) for speech quality using pretrained VITS, pretrained CosyVoice2, and our proposed OmniVIVO, as shown in Table 3. Among the generative models, OmniVIVO achieves the highest score in Naturalness (3.94 ± 0.16), surpassing VITS (3.86 ± 0.15) and CosyVoice2 (3.70 ± 0.17). For Intelligibility, OmniVIVO obtains 4.19 ± 0.14, which is slightly lower than CosyVoice2 (4.33 ± 0.12) and VITS (4.31 ± 0.12), but remains competitive. These results, based on training OmniVIVO with LibriTTS for speech-unit token generation, demonstrate its ability to produce more natural-sounding speech while preserving robust intelligibility.

Table 4: Subjective test for Multimodal Coherence with 95% confidence intervals

	Multimodal quality↑
OmniVIVO	3.79 ± 0.16

5.4 MULTIMODEL RANKING

As shown in Table 4, OmniVIVO achieves a high subjective multimodal coherence score of 3.79 ± 0.16. Since no prior work provides directly comparable multimodal evaluations, results are reported exclusively for OmniVIVO. The 5-level scale used for evaluation is described in Section A.1.

Table 5 presents the multimodal ranking evaluation. The combination of text input with OmniVIVO's speech and image outputs receives the highest proportion of excellent ratings (56.95%). Dual-modality configurations (e.g., Text+Image) are rated at 27.81%, while single-modality options are less effective for information acquisition. These findings suggest that tri-modality (text+speech+image) plays a crucial role in multimodal applications. Therefore, we release VIVOGen, a high-quality dataset with 100 samples, to advance multimodal applications such as storytelling, interactive interfaces. Representative outputs from VIVOGen are shown in Figure 3. Note that we use WhisperV2 (Radford et al.), an ASR model, to transcribe the speech output of OmniVIVO into text.

Overall, these results demonstrate OmniVIVO's ability to generate coherent multimodal outputs while maintaining strong speech quality (Table 3).

Table 5: Multimodel Ranking Metric, Unit:%

Excellent		**Acceptable**		**Less Effective**	
Text + Speech + Image	**56.95**	Speech + Image	27.81	Text	39.74
Text + Image	17.88	Image	17.22	Speech	27.81
Text + Speech	9.27	Text	15.89	Image	25.83
Other	15.9	Other	39.08	Other	6.62

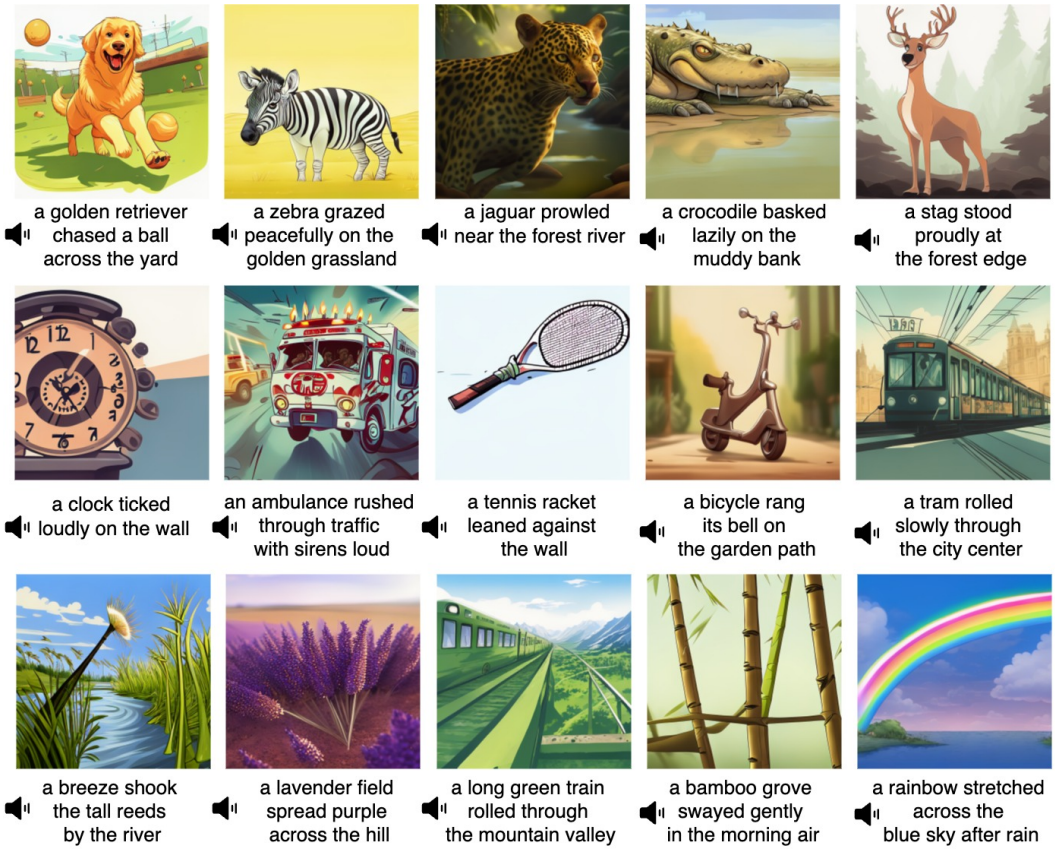

Figure 3: The images and text transcripts are taken from our released dataset, VIVOGen. Note that Whisper-V2, an ASR model, is used to transcribe speech into text.

6 Conclusion

In this work, we present **OmniVIVO**, the first unified autoregressive backbone capable of concurrently generating high-fidelity images and natural speech from a single text input. Unlike prior approaches that isolate modalities or combine separate generators, our OmniVIVO demonstrates the effectiveness of a single neural architecture that jointly models vision and voice within a shared multimodal space. Through extensive evaluation, we demonstrate that OmniVIVO outperforms a TTS baseline model of comparable size, and achieves comparable quality to SOTA models in both image quality (e.g., LlamaGen) and speech quality (e.g., VITS and CosyVoice2), as shown in subjective tests. Furthermore, we propose a new multimodal ranking metric that provides an effective way of assessing performance across modalities. Our experiments demonstrate that integrating text, image, and speech enhances information acquisition and broadens the scope of multimodal applications. In line with these findings, we target to release **VIVOGen**, a high-quality tri-modal dataset containing paired text, image, and speech data, which we expect will serve as a valuable resource for advancing multimodal generation in domains such as dynamic storytelling and education. Both the source code and dataset will be released upon acceptance of the paper.

7 ETHICS STATEMENT

We confirm that we have read and agree to follow the ICLR Code of Ethics. We commit to conducting our research responsibly, adhering to ethical standards throughout our involvement in the conference.

8 REPRODUCIBILITY STATEMENT

We confirm that our work is reproducible. We will release our source code upon acceptance of the paper.

9 THE USE OF LARGE LANGUAGE MODELS (LLMS)

We adhere to the ICLR guidelines regarding the use of LLMs. In this research, we used ChatGPT-5 to improve our writing and conduct surveys of related work.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*.

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. FunAudioLLM: Voice Understanding and Generation Foundation Models for Natural Interaction Between Humans and LLMs. *CoRR*, 2024.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-To-Image Generation via Masked Generative Transformers. In *International Conference on Machine Learning, ICML 2023*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Abhimanyu Dubey, Abhinav Jauhri, and et al. The Llama 3 Herd of Models. *CoRR*, 2024.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

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

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda*. OpenReview.net.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA*. OpenReview.net.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite Scalar Quantization: VQ-VAE Made Simple. In *The Twelfth International Conference on Learning Representations*, 2024.

OpenAI. GPT-4 Technical Report. *CoRR*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, and others. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *CoRR*, 2024.

Hongxuan Tang, Hao Liu, and Xinyan Xiao. UGen: Unified Autoregressive Multimodal Model with Progressive Vocabulary Learning. *CoRR*, 2025.

Chameleon Team. Chameleon: Mixed-Modal Early-Fusion Foundation Models. *CoRR*, abs/2405.09818, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024*.

Aaron Van Den Oord, Oriol Vinyals, and others. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *CoRR*.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *CoRR*, 2024.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. 2019.

A Appendix

A.1 Subjective Evaluation Metric

Participants rate **Naturalness** and **Intelligibility** on a 1–5 scale, with the following criteria:

Naturalness Rating Scale:

- **1 - Very Unnatural**: Speech sounds robotic or synthetic.
- **2 - Unnatural**: Noticeable artificiality in speech.
- **3 - Neutral**: Neither natural nor unnatural.
- **4 - Natural**: Speech is mostly natural with minor artifacts.
- **5 - Very Natural**: Fully natural, indistinguishable from human speech.

Intelligibility Rating Scale:

- **1 - Unintelligible**: Speech is entirely unclear.
- **2 - Poor**: Only a few words are recognizable.
- **3 - Fair**: Some segments are intelligible, but errors persist.
- **4 - Good**: Largely intelligible with minor artifacts.
- **5 - Perfect**: Fully intelligible, no effort required.

For **Multimodal Coherence**, participants evaluate the combined quality of OmniVIVO's image and speech outputs for clarity and coherence on a 1–5 scale:

Multimodal Coherence Rating Scale:

- **1 - Very Poor**: Image and speech are unclear, hard to understand, and not coherent.
- **2 - Poor**: Image and speech are somewhat unclear, difficult to understand, and lack coherence.
- **3 - Neutral**: Image and speech are clear but lack a seamless connection.
- **4 - Good**: Image and speech are clear, easy to understand, and mostly coherent.
- **5 - Excellent**: Image and speech are completely clear, easy to understand, and flow seamlessly.