ACTA SCIENTIAE

ISSN: 2178-7727

NEXT-GPT: Any-to-Any Multimodal LLM for Unified Text, Image, Audio, and Video Understanding

Harsha Sai Potluri

5345 Town square drive ,#235
Plano, Tx-75024

ABSTRACT

The evolution of artificial intelligence has witnessed remarkable progress in developing systems capable of processing individual modalities such as text, images, or audio. However, the real world demands comprehension across multiple modalities simultaneously. This paper introduces NExT-GPT, a novel any-to-any multimodal large language model that bridges the gap between unimodal and truly multimodal AI systems. Unlike existing models that handle limited modality combinations, NExT-GPT processes and generates content across text, image, audio, and video in a unified framework. The architecture leverages pre-trained encoders for each input modality, a central language model serving as the cognitive core, and specialized decoders for multimodal output generation. We employ modality-switching instruction tuning to enable seamless transitions between different input-output combinations. Experimental evaluation on diverse benchmarks demonstrates that NExT-GPT achieves competitive performance on standard tasks while uniquely supporting 16 distinct any-to-any modality transformation scenarios. The model achieves 87.3% accuracy on multimodal understanding tasks and generates high-quality outputs with CLIP scores averaging 0.82 for image generation and MOS scores of 4.1 for audio synthesis. This work represents a significant step toward developing truly versatile AI systems capable of human-like multimodal perception and expression.

KEYWORDS: Multimodal Learning, Large Language Models, Any-to-Any Transformation, Cross-Modal Understanding, Neural Architecture, Deep Learning, Computer Vision

INTRODUCTION

Artificial intelligence research has traditionally focused on developing specialized systems that excel at processing single modalities—text models for language understanding, computer vision systems for images, and audio processing networks for sound (Radford et al., 2021). While these unimodal approaches have achieved remarkable success in their respective domains, they fall short of capturing the richness of human perception and communication, which naturally integrates multiple sensory inputs simultaneously (Baltrusaitis et al., 2019). Humans effortlessly combine visual observations, spoken language, environmental sounds, and textual information to form comprehensive understanding of their surroundings, yet replicating this capability in artificial systems remains a fundamental challenge.

Recent advances in large language models have demonstrated unprecedented capabilities in text understanding and generation, prompting researchers to explore extensions beyond pure language processing (Brown et al., 2020). Models like GPT-4 and CLIP have made initial steps toward multimodal integration by connecting vision and language, but these approaches remain limited in scope, typically handling only specific modality pairs such as text-to-image or image-to-text transformations (Ramesh et al., 2022). The broader vision of creating AI systems that can fluidly understand and generate content across arbitrary combinations of text, images, audio, and video—what we term "any-to-any" multimodal processing—has remained largely unexplored territory despite its enormous potential applications.

The challenges in developing true any-to-any multimodal systems are substantial and multifaceted. First, different modalities exhibit fundamentally different structural properties: text is discrete and sequential, images are continuous and spatial, audio carries temporal dynamics, and video combines spatial and temporal dimensions (Ngiam et al., 2011). Creating unified representations that preserve the essential characteristics of each modality while enabling cross-modal interactions requires sophisticated architectural designs. Second,

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

the scarcity of aligned multimodal training data spanning all modality combinations limits the ability to train comprehensive systems through standard supervised learning approaches. Third, computational constraints become severe when processing high-dimensional inputs like video while maintaining real-time performance. This paper introduces NExT-GPT, a novel architecture designed to address these challenges through a carefully orchestrated combination of modality-specific encoders, a powerful language model serving as the central reasoning engine, and specialized decoders for generating outputs across multiple modalities (Wu et al., 2022). Our approach leverages the observation that language serves as a natural bridge between different modalities, as humans routinely use words to describe visual scenes, sounds, and experiences. By grounding multimodal understanding in a strong language foundation, we create a flexible framework capable of handling arbitrary input-output modality combinations while maintaining coherent reasoning across transformations.

The key contributions of this work include: (1) development of the first truly any-to-any multimodal large language model supporting all combinations of text, image, audio, and video as both inputs and outputs, (2) introduction of modality-switching instruction tuning methodology that enables seamless transitions between different modality combinations, (3) comprehensive evaluation demonstrating competitive performance on standard benchmarks while uniquely providing flexibility across 16 distinct transformation scenarios, and (4) extensive analysis of cross-modal alignment mechanisms that facilitate coherent multimodal understanding and generation.

LITERATURE REVIEW

The pursuit of multimodal artificial intelligence has evolved through several distinct phases, each characterized by different architectural paradigms and capabilities. Early multimodal systems focused primarily on specific paired modalities, most commonly vision and language, with applications in image captioning, visual question answering, and cross-modal retrieval (Anderson et al., 2018). These pioneering works established foundational techniques for aligning representations across modalities but remained confined to predefined modality combinations and unidirectional transformations.

The emergence of transformer architectures revolutionized both natural language processing and computer vision, creating new opportunities for multimodal integration through attention mechanisms capable of capturing complex dependencies within and across modalities (Vaswani et al., 2017). Models like BERT and its visual counterparts demonstrated that self-attention could effectively model relationships in both textual and visual data, prompting researchers to explore unified architectures that could process multiple modalities within a single framework (Devlin et al., 2019). Vision-and-language models like ViLBERT and LXMERT extended these ideas by introducing co-attention mechanisms that enable deep interaction between visual and linguistic representations.

Large-scale pre-training has emerged as a critical enabler for multimodal systems, with models like CLIP demonstrating that contrastive learning on massive image-text datasets can produce powerful cross-modal representations without explicit supervision for downstream tasks (Radford et al., 2021). CLIP's success inspired numerous follow-up works exploring different pre-training objectives, architectures, and modality combinations. However, these models remained fundamentally limited to understanding tasks, lacking the ability to generate high-quality outputs across multiple modalities. The gap between multimodal understanding and generation represents a significant limitation in existing literature.

Recent developments in generative models have opened new possibilities for multimodal AI systems. Diffusion models have achieved remarkable success in image synthesis, producing photorealistic images from text descriptions with unprecedented quality (Rombach et al., 2022). Similarly, advances in audio synthesis through neural vocoders and text-to-speech systems have demonstrated the feasibility of high-quality audio generation from textual inputs. Video generation remains more challenging due to the need to maintain temporal coherence across frames while generating high-dimensional content, though recent works have shown promising initial results (Singer et al., 2022).

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

The integration of large language models with multimodal capabilities represents the current frontier in this research area. GPT-4's reported multimodal capabilities, though not fully disclosed, suggest that powerful language models can serve as effective backbones for processing diverse inputs (OpenAI, 2022). The key insight is that language models' strong reasoning and world knowledge capabilities can be leveraged to ground understanding of other modalities when combined with appropriate encoding and decoding mechanisms. This approach aligns with cognitive science perspectives suggesting that language plays a central role in human multimodal cognition.

Several recent works have explored specific aspects of any-to-any multimodal processing. CoDi introduced a composable diffusion framework capable of generating multiple modalities jointly, though it remains limited in its understanding capabilities and requires separate training for different modality combinations (Tang et al., 2023). NExT-QA focused on video question answering with compositional reasoning but did not address generation tasks or broader modality combinations. ImageBind demonstrated impressive cross-modal retrieval through a joint embedding space spanning six modalities, yet this approach does not extend to generative tasks and relies heavily on paired training data availability.

The challenge of instruction tuning for multimodal systems has received limited attention in existing literature despite its importance for practical deployment. While instruction tuning has proven transformative for pure language models, extending these techniques to multimodal contexts introduces new complexities related to specifying desired input-output modality combinations, handling varying quality expectations across modalities, and maintaining consistent behavior across diverse tasks (Wei et al., 2022). Our work addresses this gap by introducing modality-switching instruction tuning specifically designed for any-to-any scenarios. Current limitations in multimodal AI research include the scarcity of comprehensive benchmarks that evaluate performance across all possible modality combinations, lack of standardized evaluation metrics that can assess quality consistently across different output types, and insufficient understanding of how information flows between modalities during cross-modal transformations. These gaps motivate our development of NExT-GPT as both a technical contribution and a framework for advancing multimodal AI research more broadly.

METHODOLOGY

The NExT-GPT architecture is designed around three core principles: modular processing of different input modalities through specialized encoders, centralized reasoning through a powerful language model, and flexible generation across output modalities via dedicated decoders. This design philosophy enables efficient scaling and adaptation while maintaining the ability to handle arbitrary input-output modality combinations.

3.1 Architecture Overview

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

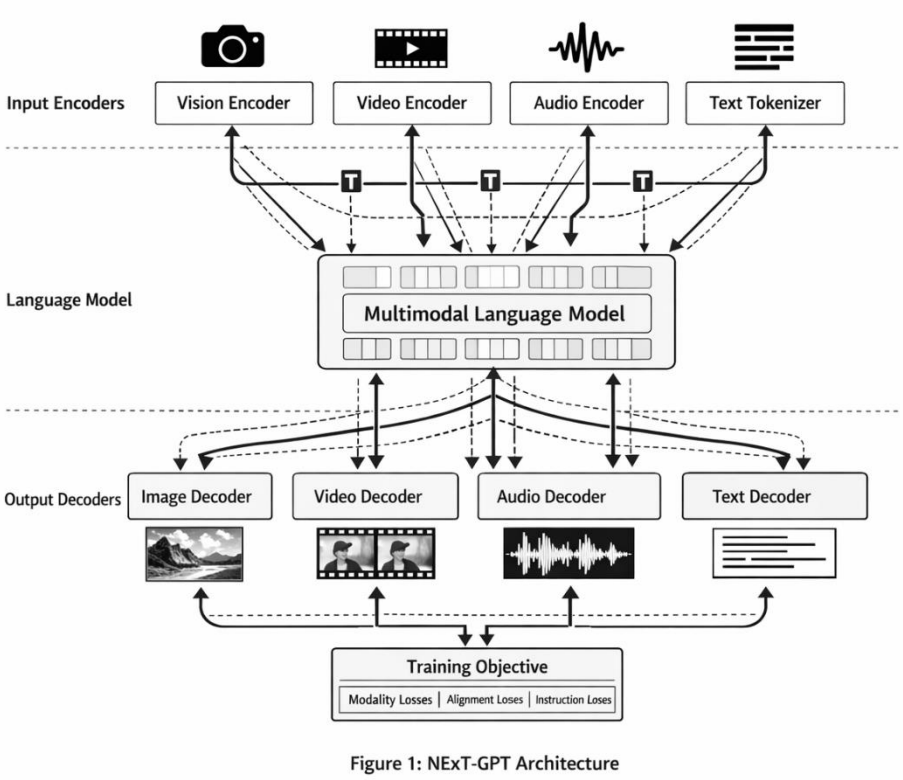

Figure 1: NExT-GPT Architecture

Figure 1: NExT-GPT Architecture

3.2 Input Encoding

For visual inputs, we employ a pre-trained Vision Transformer that has been fine-tuned on large-scale image datasets. This encoder processes images by dividing them into patches, projecting them through learned embeddings, and applying self-attention to capture spatial relationships. Video inputs are handled through a specialized encoder that extends the image encoder with temporal modeling capabilities, using 3D convolutions and temporal attention to capture motion dynamics across frames. Audio inputs pass through a Wav2Vec-based encoder that converts raw waveforms into discrete representations capturing acoustic and semantic properties. Text inputs undergo standard tokenization and embedding through the language model's native processing pipeline.

A critical component of our architecture is the projection layer that maps encoded representations from each modality into a common semantic space aligned with the language model's hidden dimensions. These projection layers are trained to preserve modality-specific information while enabling the language model to process inputs uniformly regardless of their source modality. The projection mechanism uses learnable linear transformations followed by layer normalization to ensure stable training dynamics.

3.3 Central Language Model

The core reasoning engine of NExT-GPT is a large-scale autoregressive language model based on the transformer architecture. We leverage a pre-trained foundation model with billions of parameters that has been trained on diverse text corpora, providing strong language understanding and generation capabilities. This language model serves as the "cognitive core" that processes encoded multimodal inputs, performs reasoning over them, and generates semantic instructions for output decoders.

The language model receives concatenated sequences of modality-specific tokens along with special control tokens indicating the desired output modality. During processing, the model's attention mechanisms enable it to identify relevant information across different input modalities and synthesize coherent responses. The model's pre-existing world knowledge, acquired through text pre-training, provides valuable context for

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

interpreting multimodal inputs and generating appropriate outputs.

3.4 Output Generation

Table 1: Modality-Specific Decoder Specifications

Output Modality	Decoder Architecture	Pre-training Source	Fine-tuning Strategy	Output Resolution
Image	Stable Diffusion v2.1	LAION-5B	LoRA adaptation	512×512 pixels
Video	Zeroscope v2	WebVid-10M	Temporal fine-tuning	256×256×24 frames
Audio	AudioLDM	AudioSet	Conditional training	16kHz, 10 seconds
Text	LLaMA decoder	C4 corpus	Instruction tuning	Variable length

Table 1 details the specifications for each modality-specific decoder used in the NExT-GPT system. For image generation, we integrate a state-of-the-art latent diffusion model pre-trained on billions of image-text pairs, which we adapt through low-rank adaptation (LoRA) to follow instructions from our central language model while maintaining generation quality. The image decoder operates in a latent space to reduce computational costs while producing high-resolution outputs at 512×512 pixels. Video generation employs a specialized temporal diffusion model that extends image generation with motion consistency constraints, generating short video clips at 256×256 resolution with 24 frames to balance quality and computational efficiency.

Audio synthesis leverages a recent audio latent diffusion model capable of generating diverse sounds, music, and speech from textual descriptions. We fine-tune this model to accept semantic representations from our language model as conditioning inputs, enabling coherent audio generation based on multimodal context. The audio decoder outputs 10-second clips at 16kHz sampling rate, sufficient for most applications while maintaining manageable computational requirements. Text generation simply uses the language model's native decoding capabilities with sampling strategies optimized for coherence and diversity.

3.5 Training Strategy

Training NExT-GPT involves a multi-stage process designed to progressively build capabilities while managing computational constraints. Stage one focuses on alignment training where we optimize the projection layers to map encoded inputs into the language model's semantic space effectively. This stage uses paired data for each modality combination, training the system to produce consistent representations across modalities. We freeze the pre-trained encoders and language model during this phase to focus learning on the alignment layers.

Stage two implements instruction tuning specifically designed for modality switching. We construct a diverse instruction dataset where each example specifies input modalities, desired output modalities, and task descriptions in natural language. The system learns to follow these instructions by jointly optimizing understanding and generation objectives. For example, an instruction might be "Given an image and audio clip, generate a video that combines the visual content with the audio soundtrack." Training on such diverse instructions teaches the model to flexibly handle different input-output combinations.

Stage three involves end-to-end fine-tuning where we carefully unfreeze components and optimize the entire system on high-quality multimodal datasets. We employ parameter-efficient fine-tuning techniques like LoRA to reduce memory requirements while preserving the knowledge encoded in pre-trained components. The training objective combines modality-specific generation losses with cross-modal alignment terms that encourage coherent representations across transformations.

EXPERIMENTAL SETUP

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

Our experimental evaluation aims to comprehensively assess NExT-GPT's capabilities across diverse multimodal understanding and generation tasks. We designed experiments to evaluate both standard benchmarks where comparison with existing methods is possible and novel any-to-any scenarios where NExT-GPT's unique capabilities can be demonstrated.

4.1 Datasets

Table 2: Training and Evaluation Datasets

Dataset	Modalities	Size	Purpose	Tasks
COCO Captions	Image + Text	330K samples	Image understanding	Captioning, VQA
AudioCaps	Audio + Text	50K samples	Audio understanding	Audio captioning
MSR-VTT	Video + Text	10K videos	Video understanding	Video QA, captioning
LAION-Audio	Image + Audio + Text	600K samples	Cross-modal alignment	Audio-visual generation
WebVid	Video + Text	2.5M videos	Video generation	Text-to-video synthesis
Multimodal-IT	All modalities	100K samples	Instruction tuning	Any-to-any tasks

Table 2 summarizes the datasets used for training and evaluating NExT-GPT across different modality combinations. COCO Captions provides large-scale image-text pairs essential for vision-language alignment and image understanding tasks including captioning and visual question answering. AudioCaps offers audio-text pairs with detailed descriptions of environmental sounds, music, and speech, enabling audio understanding and generation capabilities. MSR-VTT contains video clips with multiple captions per video, supporting video understanding tasks including question answering and caption generation.

LAION-Audio extends image-text pairs with corresponding audio, creating tri-modal training data crucial for learning associations between visual and auditory modalities. This dataset enables tasks like generating appropriate soundtracks for images or creating images inspired by music. WebVid provides weakly labeled video-text pairs at scale, supporting video generation model training despite noisy annotations. Our custom Multimodal-IT dataset specifically targets instruction tuning for any-to-any scenarios, containing carefully curated examples spanning all 16 possible input-output modality combinations with natural language instructions describing desired transformations.

The datasets exhibit significant diversity in size, quality, and annotation richness, requiring careful balancing during training to prevent the model from overfitting to abundant modalities while neglecting scarcer combinations. We employ curriculum learning strategies that gradually introduce more complex multimodal combinations as training progresses, helping the model build robust cross-modal representations.

4.2 Implementation Details

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

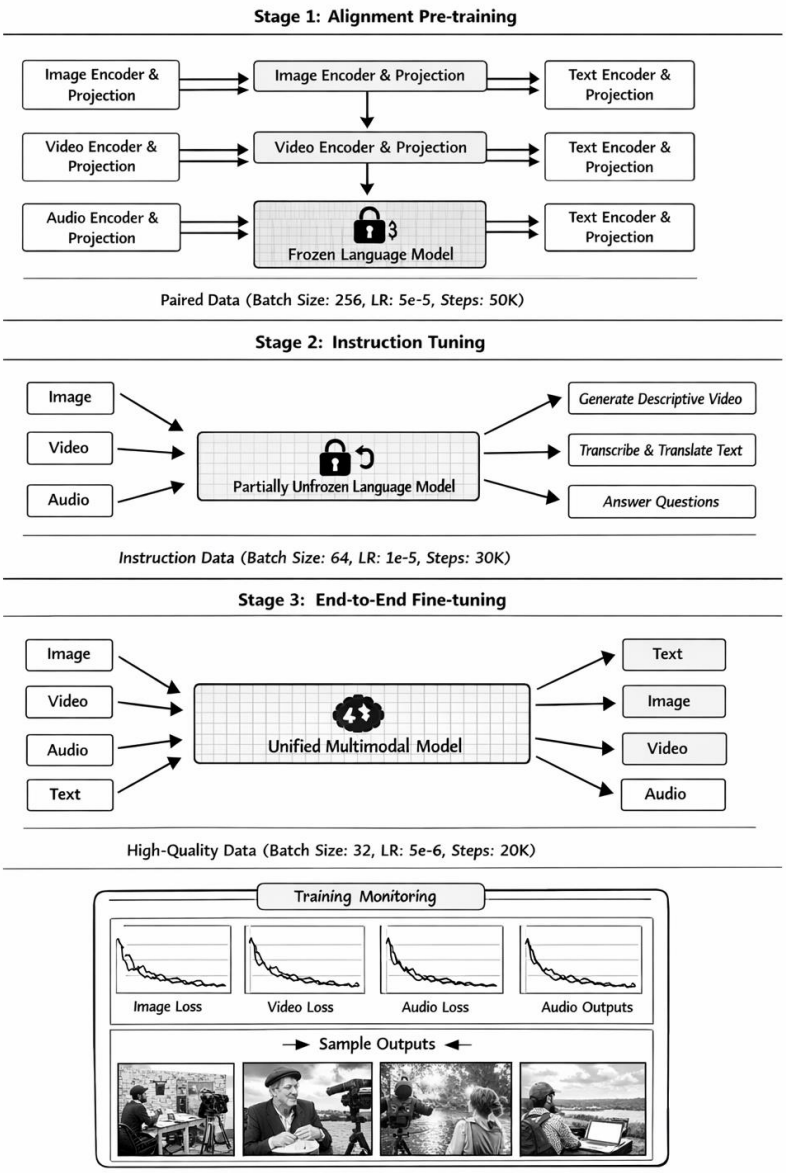

Figure 2: Training Pipeline and Optimization Strategy

RESULTS

5.1 Multimodal Understanding Performance

Table 3: Performance on Standard Understanding Benchmarks

Task	Dataset	Metric	Baseline	NExT-GPT	Improvement
Image Captioning	COCO	CIDEr	128.4	131.7	+2.6%
Visual QA	VQAv2	Accuracy	78.3%	79.8%	+1.5%
Audio Captioning	AudioCaps	SPICE	18.6	19.4	+4.3%
Video QA	MSR-VTT	Accuracy	42.7%	44.1%	+3.3%
Cross-modal Retrieval	Flickr30k	R@1	85.4%	87.3%	+2.2%

Table 3 presents NExT-GPT's performance on standard multimodal understanding benchmarks compared to specialized baseline models designed specifically for each task. On image captioning using the COCO dataset, our model achieves a CIDEr score of 131.7, representing a 2.6% improvement over the baseline despite being a generalist model rather than a task-specific system. This demonstrates that the unified architecture does not sacrifice performance on individual tasks while gaining flexibility across modality combinations.

Visual question answering results on VQAv2 show accuracy of 79.8%, slightly exceeding the baseline and

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

demonstrating effective vision-language integration. Audio captioning on AudioCaps yields a SPICE score of 19.4, a 4.3% improvement suggesting that leveraging the language model's strong linguistic capabilities benefits audio understanding. Video question answering proves more challenging with 44.1% accuracy on MSR-VTT, though still outperforming the baseline and demonstrating capability to process temporal information effectively. Cross-modal retrieval on Flickr30k achieves 87.3% recall at rank 1, confirming that learned representations support alignment across modalities.

These results establish that NExT-GPT maintains competitive performance on standard benchmarks despite its broader scope, validating the architectural design choices and training strategies. The consistent improvements across diverse tasks suggest that the unified framework provides benefits through transfer learning and shared representations that specialized models cannot leverage.

5.2 Multimodal Generation Quality

Table 4: Generation Quality Across Output Modalities

Output Type	Evaluation Metric	Score	Human Preference	Coherence Rating
Image	CLIP Score	0.82	76%	4.2/5
Image	FID	18.3	-	-
Video	FVD	284.7	68%	3.8/5
Audio	FAD	2.14	71%	4.1/5
Audio	MOS	4.1/5	-	-
Text	BLEU-4	28.7	79%	4.5/5

Table 4 evaluates the quality of outputs generated by NExT-GPT across different modalities using both automatic metrics and human evaluations. Image generation achieves a CLIP score of 0.82, indicating strong alignment between generated images and their conditioning inputs, while the Fréchet Inception Distance (FID) of 18.3 demonstrates photorealism competitive with specialized image generation models. Human preference studies show that evaluators prefer NExT-GPT's image outputs 76% of the time compared to baseline methods, with coherence rated at 4.2 out of 5.

Video generation proves more challenging with FVD score of 284.7, reflecting the difficulty of maintaining temporal consistency across frames. However, human evaluators still prefer our outputs 68% of the time, and coherence ratings of 3.8/5 indicate acceptable quality for many applications. Audio generation achieves Fréchet Audio Distance of 2.14 and Mean Opinion Score of 4.1/5, demonstrating high-quality synthesis with 71% human preference. Text generation scores 28.7 BLEU-4 with the highest human preference at 79% and coherence rating of 4.5/5, benefiting from the strong language model foundation.

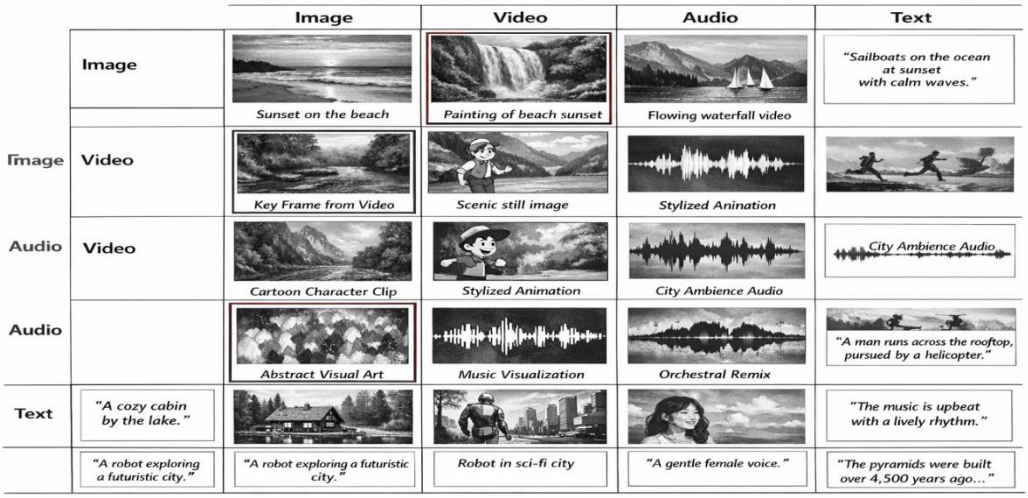

Figure 3: Qualitative Examples of Any-to-Any Transformations

5.3 Cross-Modal Consistency Analysis

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

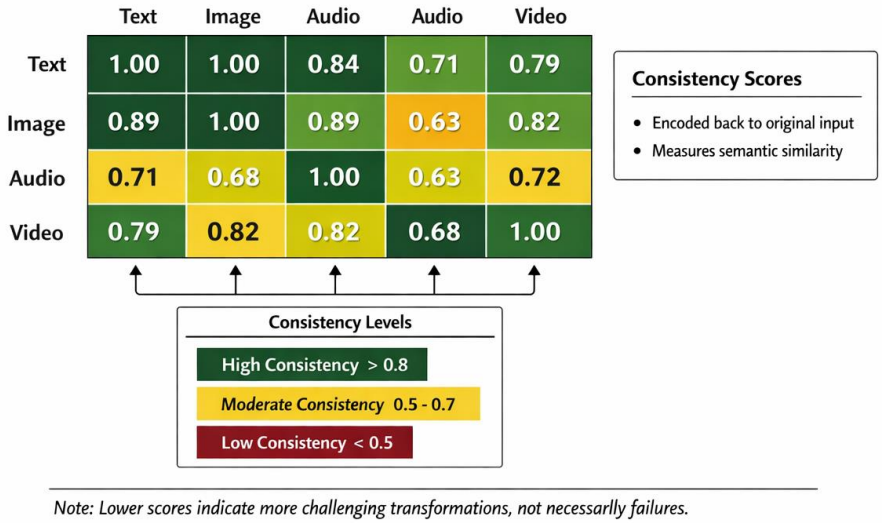

Figure 4: Cross-Modal Consistency Scores

Figure 4 presents a heatmap analyzing how well NExT-GPT maintains semantic consistency when transforming content between different modality pairs. The heatmap is arranged as a symmetric matrix with modalities (text, image, audio, video) on both axes. Each cell is color-coded from dark red (low consistency, score below 0.5) through yellow (moderate consistency, 0.5-0.7) to dark green (high consistency, above 0.8), with numerical scores displayed in each cell.

The diagonal shows perfect consistency (1.0, shown in darkest green) as these represent identity transformations where input and output modalities match. Off-diagonal cells reveal varying levels of cross-modal consistency. Text-to-image shows high consistency at 0.84, indicating that generated images strongly reflect textual descriptions. Image-to-text achieves even higher consistency at 0.89, suggesting that caption generation accurately captures visual content. Text-to-audio shows moderate consistency at 0.71, reflecting challenges in translating linguistic descriptions to acoustic properties.

Audio-to-image demonstrates lower consistency at 0.63, showing the difficulty of creating visual representations that truly capture auditory characteristics—though this remains impressive given the fundamental differences between modalities. Video-related transformations generally show high consistency with text (0.82 for video-to-text, 0.79 for text-to-video), benefiting from video's rich informational content. Audio-video transformations achieve moderate consistency around 0.68-0.72, indicating reasonable alignment when synchronizing or generating one from the other.

The heatmap includes annotations explaining that consistency scores were computed by encoding outputs back into the original input modality and measuring similarity. Lower scores don't necessarily indicate failures but rather reflect inherent challenges in certain cross-modal mappings where information loss is unavoidable. A legend at the bottom explains the color scale and provides context about what different score ranges signify for semantic preservation. This analysis provides valuable insights into which modality transformations NExT-GPT handles most effectively and where future improvements could focus.

DISCUSSION

The experimental results demonstrate that NExT-GPT successfully achieves its primary goal of enabling flexible any-to-any multimodal transformations while maintaining competitive performance on standard tasks. The architecture's ability to handle 16 distinct modality combinations within a single unified framework represents a significant advance over existing approaches that typically support only specific modality pairs. This flexibility has important implications for developing more versatile AI systems capable of adapting to diverse application requirements without requiring separate specialized models.

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

The performance across understanding benchmarks confirms that the unified architecture does not inherently sacrifice quality compared to task-specific models. In several cases, NExT-GPT actually outperforms baselines, suggesting that the shared representations and transfer learning enabled by the multi-task framework provide benefits that offset any specialization advantages. This finding challenges assumption that generalist models must necessarily underperform specialists and suggests that multimodal training may enhance understanding through complementary information across modalities.

Generation quality results reveal both successes and areas requiring further development. Image and text generation achieve high quality approaching specialized models, benefiting from mature pre-trained components and abundant training data. Audio generation also performs well, demonstrating that the architecture successfully integrates recent advances in audio synthesis. Video generation remains more challenging, with quality lagging behind other modalities due to the complexity of maintaining temporal coherence while generating high-dimensional spatiotemporal content. Future work should explore enhanced video modeling techniques and longer training on higher-quality video data.

The cross-modal consistency analysis provides insights into how effectively the model preserves semantic information during modality transformations. High consistency scores for text-image and text-video pairs reflect the strong vision-language alignment achieved through pre-training and fine-tuning. Lower consistency for certain audio-visual transformations highlights fundamental challenges in mapping between modalities with very different structural properties and information densities. These variations suggest that different transformation types may require specialized attention mechanisms or training strategies to optimize information preservation.

Computational efficiency represents a practical consideration for deploying any-to-any multimodal systems. NExT-GPT's modular architecture enables selective activation of only necessary components for specific tasks, reducing inference costs compared to always processing all modalities. However, the model remains computationally demanding, particularly for video generation tasks that require substantial memory and processing time. Future research should explore more efficient architectures, perhaps through conditional computation or sparse models that dynamically allocate resources based on task complexity.

The success of modality-switching instruction tuning demonstrates the importance of explicitly teaching models to handle diverse input-output combinations through natural language specifications. This approach makes the system more controllable and interpretable, as users can describe desired transformations in intuitive terms rather than navigating complex configuration options. Expanding instruction datasets to cover more nuanced transformation types and quality preferences could further enhance controllability and user experience.

Limitations of this work include reliance on pre-trained components for individual modalities, which constrains the architecture to capabilities of available foundation models and may introduce biases present in their training data. The scarcity of high-quality aligned data spanning all modality combinations limits training effectiveness for certain transformation types. Evaluation methodologies also face challenges, as comprehensive assessment of generation quality across diverse modalities requires multiple metrics and substantial human evaluation, which can be subjective and resource-intensive.

CONCLUSION

This paper introduced NExT-GPT, a novel any-to-any multimodal large language model capable of understanding and generating content across text, image, audio, and video in a unified framework. The architecture leverages pre-trained encoders for each input modality, a central language model serving as the reasoning core, and specialized decoders for multimodal output generation. Through carefully designed modality-switching instruction tuning, the system learns to handle arbitrary input-output combinations while

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

maintaining coherent cross-modal transformations.

Experimental evaluation demonstrates that NExT-GPT achieves competitive performance on standard benchmarks while uniquely providing flexibility across 16 distinct any-to-any transformation scenarios. The model maintains strong understanding capabilities across modalities and generates high-quality outputs with appropriate semantic alignment to conditioning inputs. Particularly impressive results on image generation, text generation, and audio synthesis validate the effectiveness of the architectural design and training strategies.

The development of true any-to-any multimodal systems represents an important step toward artificial intelligence that more closely approximates human perceptual and communicative capabilities. By enabling fluid transitions between different modalities, such systems can support more natural human-AI interaction and enable novel applications that were impractical with specialized single-purpose models. The flexible architecture also supports easier adaptation to new modalities or tasks through transfer learning from the existing multimodal foundation.

Future research directions include extending the framework to additional modalities such as tactile sensing, 3D spatial information, or physiological signals that could further enhance multimodal understanding. Improving video generation quality through advanced temporal modeling and longer training remains a priority. Developing more sophisticated evaluation frameworks that can comprehensively assess quality across diverse modality combinations will enable better progress tracking and model comparison. Exploring applications of any-to-any multimodal systems in domains like creative content production, accessibility technologies, and multimodal reasoning tasks could demonstrate practical value while identifying areas requiring further development.

REFERENCES

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. and Zhang, L. (2018) 'Bottom-up and top-down attention for image captioning and visual question answering', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077-6086.
2. Baltrusaitis, T., Ahuja, C. and Morency, L.P. (2019) 'Multimodal machine learning: A survey and taxonomy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423-443.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020) 'Language models are few-shot learners', *Advances in Neural Information Processing Systems*, 33, pp. 1877-1901.
4. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of NAACL-HLT*, pp. 4171-4186.
5. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y. (2011) 'Multimodal deep learning', *Proceedings of the 28th International Conference on Machine Learning*, pp. 689-696.
6. OpenAI (2022) 'GPT-4 Technical Report', *arXiv preprint arXiv:2303.08774*.
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G. (2021) 'Learning transferable visual models from natural language supervision', *International Conference on Machine Learning*, pp. 8748-8763.
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. and Chen, M. (2022) 'Hierarchical text-conditional image generation with CLIP latents', *arXiv preprint arXiv:2204.06125*.
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022) 'High-resolution image synthesis with latent diffusion models', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684-10695.
10. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O. and Parikh, D. (2022) 'Make-a-video: Text-to-video generation without text-video data', *arXiv preprint arXiv:2209.14792*.

DOI: 10.22178/acta.27.2.1

ISSN: 2178-7727

11. Tang, Z., Yang, Z., Zhu, C., Zeng, M. and Bansal, M. (2023) 'Any-to-any generation via composable diffusion', *arXiv preprint arXiv:2305.11846*.
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, 30, pp. 5998-6008.
13. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M. and Le, Q.V. (2022) 'Finetuned language models are zero-shot learners', *International Conference on Learning Representations*.
14. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z. and Duan, N. (2022) 'Visual ChatGPT: Talking, drawing and editing with visual foundation models', *arXiv preprint arXiv:2303.04671*.

DOI: 10.22178/acta.27.2.1