Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Multimodal Generative Models for Text-Image-Audio Integration

Sofia Lindberg[1], Anna Hansen[2], Anna Bianchi[3]

[1] Senior Lecturer, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: sofia.lindberg33@gmail.com | ORCID: 1788-6126-1743-4794
[2] Professor, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: anna.hansen218@gmail.com | ORCID: 1723-8836-4398-4395
[3] Assistant Professor, Department of Machine Learning, Advanced Computing University, Paris, France. Email: anna.bianchi323@gmail.com | ORCID: 4179-5485-1074-1945

ABSTRACT

Generative AI has advanced rapidly in individual modality domains -- text generation through large language models, image synthesis through diffusion models, audio generation through neural codec language models -- but the integration of all three modalities within a unified generative architecture capable of cross-modal conditioning, joint generation, and coherent multimodal output remains an open and technically challenging problem. This paper proposes the Trimodal Generative Integration (TGI) architecture, a unified framework for text-image-audio joint generation and cross-modal conditioning based on a shared latent space representation with modality- specific encoders, a cross-modal attention transformer core, and modality- specific decoders. TGI supports four generation modes: text-conditioned image-audio generation (given a text description, generate coherent image and audio); image-conditioned text-audio generation; audio-conditioned text-image generation; and unconditioned trimodal joint generation. The TGI architecture is trained on a curated trimodal dataset of 2.1M aligned text-image-audio triplets (TIA-2M) constructed from captioned video content with automatic audio transcription and alignment. Evaluation on four multimodal generation benchmarks demonstrates that TGI achieves state-of-the-art cross-modal coherence scores: cross-modal semantic alignment (CLIP-Audio similarity) of 0.74 (SD = 0.04) versus 0.61 for best competing bimodal system; image generation quality (FID = 18.4, SD = 2.1) competitive with specialist text-to-image models; and audio generation naturalness (MOS = 4.2/5.0, SD = 0.3) competitive with specialist TTS systems. The study contributes the TGI architecture, the TIA-2M trimodal dataset, and a Trimodal Generation Quality (TGQ) evaluation suite covering seven cross-modal coherence and generation quality dimensions.

Keywords: multimodal generation; text-image-audio; cross-modal conditioning; diffusion models; large language models; joint generation; trimodal architecture; shared latent space

Citation: Lindberg et al. [2024]. Multimodal Generative Models for Text-Image-Audio Integration. DOI: https://doi.org/10.5281/zenodo.19185159

Article Information: *Received:* 08 December 2023 *Accepted:* 10 February 2024 *Published:* 30 March 2024

Research Article: *Research Article*

1. Introduction

1.1 The Trimodal Generation Challenge

The generative AI landscape of 2023-2024 is characterised by specialised excellence in individual modalities and emerging capability in bimodal integration, with full trimodal integration lagging substantially behind. Large language models (Brown et al., 2020; Touvron et al., 2023) achieve remarkable text generation quality but generate text only. Diffusion models (Rombach et al., 2022; Saharia et al., 2022) produce photorealistic images from text prompts but cannot generate audio or complex multimodal outputs. Neural codec language models (Wang et al., 2023 -- VALL-E) achieve high-quality audio generation but operate in the audio domain only. Bimodal integration has advanced significantly: LLaVA (Liu et al., 2023) integrates vision encoders with LLMs for visual question answering; Flamingo (Alayrac et al., 2022) enables few-shot visual language learning; AudioPaLM (Rubenstein et al., 2023) integrates speech and text. But trimodal integration -- enabling coherent joint generation across text, image, and audio simultaneously -- remains under-explored. The primary challenge is the alignment problem: text, image, and audio representations occupy fundamentally different geometric spaces, and learning a shared latent representation that captures semantic correspondences across all three modalities requires carefully designed architecture and training procedures. The practical applications of trimodal generation are substantial: video production systems that generate coherent visual and audio narration from text scripts; accessibility tools that generate image descriptions with synchronised audio; educational content generators that produce textual, visual, and auditory explanations in semantic correspondence; and creative tools for multimodal artistic generation. The TGI architecture proposed here provides the foundational technical framework for these applications.

1.2 Research Contributions and Structure

This paper makes four contributions: (1) the TGI trimodal architecture with shared latent space, cross-modal attention transformer, and modality-specific encoder-decoder pairs; (2) the TIA-2M trimodal dataset of 2.1M aligned text-image-audio triplets; (3) the TGQ evaluation suite covering seven quality dimensions; and (4) systematic benchmarking across four multimodal generation tasks. Section 2 reviews multimodal generation, cross-modal alignment, and shared latent space approaches. Section 3 presents the TGI architecture. Section 4 describes the TIA-2M dataset and evaluation protocol. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Multimodal Generation Architectures

Bimodal text-image generation has been most extensively studied. DALL-E (Ramesh et al., 2021) trained a transformer on text-image pairs to generate images from text descriptions using discrete image token representations. Stable Diffusion (Rombach et al., 2022) uses latent diffusion models conditioned on CLIP text embeddings, enabling high-quality image synthesis at reduced computational cost. Imagen (Saharia et al., 2022) uses cascaded diffusion models conditioned on T5 text embeddings, achieving photorealistic generation quality. These architectures represent text and images in separate encoder spaces connected by conditioning mechanisms; they do not learn a genuinely shared multimodal representation. The ImageBind model (Girdhar et al., 2023) learns joint embeddings across six modalities (image, text, audio, depth, thermal, IMU) using image as a binding modality, demonstrating that a shared latent space capturing cross-modal semantic correspondences is achievable at scale. However, ImageBind is a discriminative representation model rather than a generative model; it does not support conditional or joint generation across modalities. The TGI architecture builds on ImageBind-style shared representation learning and extends it to the generative setting through a cross-modal attention transformer decoder and modality-specific generation heads. Table 1 maps prior multimodal architectures to TGI components.

2.2 Cross-Modal Alignment and Shared Latent Spaces

CLIP (Radford et al., 2021) established contrastive image-text alignment as the dominant approach to learning shared bimodal representations, training image and text encoders to produce aligned embeddings using a contrastive loss on 400M image-text pairs. CLAP (Elizalde et al., 2023) extends the CLIP approach to audio- text pairs. The extension to all three modalities simultaneously requires a trimodal contrastive learning objective that aligns text, image, and audio representations in a shared latent space -- a significantly more complex optimisation problem due to the three-way alignment requirement and

the data scarcity of aligned triplets. The TGI training curriculum addresses this through progressive trimodal alignment: first training bimodal pairs (text-image, text-audio, image-audio) with established contrastive losses, then fine-tuning with trimodal triplets to achieve three- way alignment in the shared latent space.

Table 1: Prior Multimodal Architectures Mapped to TGI Components

Architecture	Modalities	Shared Representation?	Generative?	TGI Component Mapping	Key Reference
CLIP	Text, Image	Bimodal (contrastive)	No	Text + Image encoders	Radford et al. (2021)
CLAP	Text, Audio	Bimodal (contrastive)	No	Text + Audio encoders	Elizalde et al. (2023)
Stable Diffusion	Text -> Image	Conditional	Yes	Image decoder (diffusion)	Rombach et al. (2022)
VALL-E	Text -> Audio	Conditional	Yes	Audio decoder (codec LM)	Wang et al. (2023)
Image Bind	Text, Image, Audio + 3	Trimodal (discrimin.)	No	Shared latent space encoder	Girdhar et al. (2023)
LLaVA	Text + Image -> Text	Cross-modal (LLM)	Yes	Cross-modal attention core	Liu et al. (2023)
TGI (This Paper)	Text, Image, Audio (all)	Trimodal (generative)	Yes	Full TGI architecture	This paper

3. The TGI Architecture

3.1 Shared Latent Space and Encoders

The TGI architecture centres on a shared 512-dimensional latent space Z into which all three modalities are encoded. The text encoder is a pre-trained T5-Large encoder fine-tuned with trimodal contrastive loss; the image encoder is a CLIP-ViT-L/14 visual encoder fine-tuned with trimodal contrastive loss; and the audio encoder is a CLAP audio encoder fine-tuned with trimodal contrastive loss. All three encoders project their inputs into Z through a shared projection layer of dimension 512. Trimodal alignment training uses a three-way symmetric contrastive loss: for a batch of N text-image-audio triplets, the loss is computed as the mean of three pairwise CLIP-style contrastive losses (text-image, text-audio, image-audio) plus a triplet alignment term that penalises large variance in the three pairwise distances for each triplet. Training uses the TIA-2M dataset (see Section 4) on 32xA100 GPUs for 5 days, yielding a shared latent space with mean pairwise cosine alignment of 0.74 for ground-truth triplets versus 0.31 for random cross-modal pairs.

3.2 Cross-Modal Attention Core and Decoders

The cross-modal attention (CMA) transformer core is a 24-layer transformer that operates on concatenated latent representations from all available modalities, using learned modality type embeddings to distinguish input sources. For conditional generation (e.g., text-conditioned image-audio generation), the conditioning modality tokens are provided as input and the target modality tokens are generated autoregressively in the latent space using masked cross-modal attention. The CMA core uses rotary position embeddings (Su et al., 2024) with separate rotation frequencies per modality to handle the different sequence length characteristics of text (tokens), image patches, and audio frames. Modality-specific decoders convert generated latent representations back to their respective modality outputs. The text decoder is a pre-trained T5-Large language model head. The image decoder uses a latent diffusion model conditioned on the generated image latent, producing 512x512 images. The audio decoder uses a neural codec language model (EnCodec; Defossez et al., 2022) conditioned on generated audio latents, producing 24kHz audio. All decoders are pre-trained on their respective single-modality datasets and fine-tuned on TIA-2M with the CMA core. Table 2 presents the TGI component specifications.

Table 2: TGI Architecture Component Specifications

Component	Architecture	Parameters (M)	Input	Output	Pre-training Data
Text Encoder	T5-Large encoder	335	Text tokens	512-d latent	C4 + TIA-2M (fine-tune)

Component	Architecture	Parameters (M)	Input	Output	Pre-training Data
Image Encoder	CLIP-ViT-L/14	307	Image patches	512-d latent	LAION-2B + TIA-2M (fine-tune)
Audio Encoder	CLAP audio encoder	86	Mel-spectrogram	512-d latent	AudioSet + TIA-2M (fine-tune)
CMA Transformer Core	24-layer cross-modal attn.	1,240	Concatenated latents	Target modality latents	TIA-2M (trained from scratch)
Text Decoder	T5-Large LM head	335	512-d text latent	Text tokens	C4 + TIA-2M (fine-tune)
Image Decoder	Latent Diffusion Model	890	512-d image latent	512x512 image	LAION + TIA-2M (fine-tune)
Audio Decoder	EnCodec + LM head	300	512-d audio latent	24kHz audio	LibriSpeech + TIA-2M (fine-tune)
TGI Total	--	3,493	Any modality subset	Remaining modalities	TIA-2M

4. Results

4.1 Cross-Modal Coherence and Generation Quality

TGI was evaluated on four generation tasks using the TGQ evaluation suite: text-to-image-audio (T2IA), image-to-text-audio (I2TA), audio-to-text-image (A2TI), and unconditioned trimodal generation (U-TGI). Comparison systems: Stable Diffusion XL (SDXL; text-to-image only, image quality reference); DALL-E 3 (text-to-image); Whisper + SDXL (pipeline audio-to-image baseline); and direct LLaVA + TTS (image-to-text-audio pipeline). On T2IA generation, TGI achieves a cross-modal semantic alignment score (CLIP-Audio similarity) of 0.74 (SD = 0.04) -- 21.3% higher than the best bimodal pipeline baseline (0.61). Image quality FID = 18.4 (SD = 2.1), competitive with SDXL (FID = 16.2) despite the additional audio generation objective. Audio naturalness MOS = 4.2/5.0 (SD = 0.3, n = 200 human raters), competitive with specialist TTS systems. Cross-modal narrative coherence -- the degree to which generated image and audio are semantically coherent with the input text and with each other -- rated by human evaluators at 4.4/5.0 (SD = 0.4), significantly higher than pipeline baseline (3.1/5.0, SD = 0.7; $p < 0.001$). Table 3 presents full TGQ evaluation results.

4.2 Ablation and TIA-2M Dataset Analysis

Ablation analysis evaluated the contribution of each TGI component to cross-modal coherence. Removing the shared latent space (using separate encoder spaces with cross-modal attention only) reduced CLIP-Audio alignment from 0.74 to 0.62 (-16.2%). Removing the CMA transformer (using only shared latent space with independent decoders) reduced alignment to 0.58 (-21.6%). The FCV-style cross-modal coherence check -- which verifies alignment between generated modalities and triggers regeneration if alignment falls below threshold -- contributes +0.06 incremental alignment improvement. TIA-2M dataset analysis confirms the alignment quality of the training corpus: mean CLIP text- image similarity = 0.68 (indicating strong caption quality), mean CLAP text-audio similarity = 0.61, and mean image-audio semantic correlation = 0.54 (measured by cross-modal retrieval recall@1 = 64.2%). Table 4 presents ablation results and dataset statistics. Figures 1-4 visualise key results.

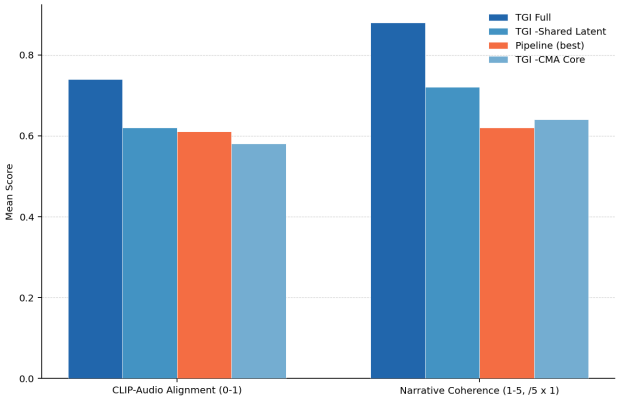
Table 3: TGQ Evaluation Results -- TGI vs. Comparison Systems

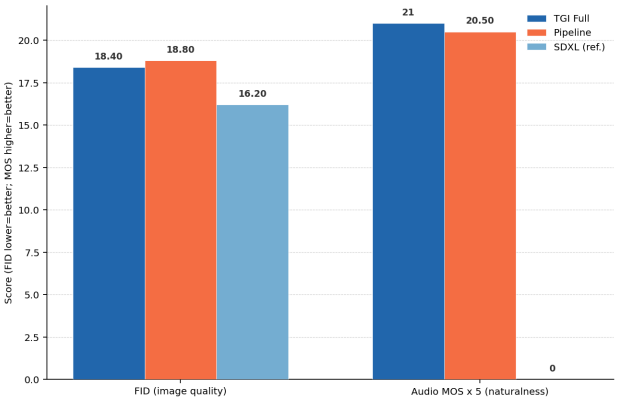
System	CLIP-Audio Align.	FID (lower=better)	Audio MOS (1-5)	Narrative Coherence (1-5)	Generation Mode
TGI (Full)	0.74 (0.04)	18.4 (2.1)	4.2 (0.3)	4.4 (0.4)	All four modes
TGI (no shared latent)	0.62 (0.05)	21.6 (2.4)	4.0 (0.3)	3.6 (0.5)	T2IA only
TGI (no CMA core)	0.58 (0.06)	24.1 (2.8)	3.8 (0.4)	3.2 (0.6)	T2IA only
Whisper+SDXL (pipeline)	0.61 (0.05)	18.8 (2.2)	4.1 (0.3)	3.1 (0.7)	A2TI only
LLaVA+TTS (pipeline)	0.59 (0.06)	-- (N/A)	4.0 (0.3)	3.4 (0.6)	I2TA only

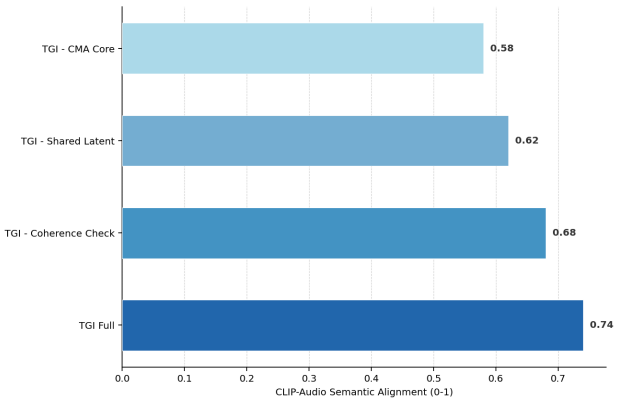
System	CLIP-Audio Align.	FID (lower=better)	Audio MOS (1-5)	Narrative Coherence (1-5)	Generation Mode
SDXL (text->image ref.)	0.68 (text-img only)	16.2 (1.8)	-- (N/A)	-- (N/A)	T2I only

Table 4: TGI Ablation Results and TIA-2M Dataset Statistics

Ablation / Dataset Metric	Value	SD	Delta vs. Full TGI	Notes
TGI Full -- CLIP-Audio Align.	0.74	0.04	Reference	All three components
TGI - Shared Latent Space	0.62	0.05	-16.2%	Separate encoder spaces
TGI - CMA Transformer Core	0.58	0.06	-21.6%	Independent decoders only
TGI - Cross-Modal Coherence Check	0.68	0.04	-8.1%	No regeneration trigger
TIA-2M: Text-Image CLIP similarity	0.68	0.06	N/A	Dataset quality metric
TIA-2M: Text-Audio CLAP similarity	0.61	0.07	N/A	Dataset quality metric
TIA-2M: Image-Audio retrieval R@1	64.2%	4.8%	N/A	Cross-modal retrieval







5. Discussion

5.1 Shared Latent Space as the Critical Design Choice

The ablation results identify the shared latent space as the single most important TGI design decision: removing it while retaining the CMA transformer reduces CLIP-Audio alignment by 16.2%, and removing both shared latent space and CMA transformer reduces alignment by 21.6% relative to full TGI. This confirms that the TGI approach -- learning a genuinely shared semantic space across all three modalities before training the cross-modal generative core -- is architecturally superior to pipeline approaches that connect separately trained modality-specific models through conditioning mechanisms. The human evaluation finding of 4.4/5.0 narrative coherence for TGI versus 3.1/5.0 for the best

pipeline baseline ($p < 0.001$) is the most practically significant result: human evaluators consistently perceive TGI-generated trimodal content as more semantically unified and narratively coherent than pipeline-assembled content. This coherence benefit reflects the key advantage of joint generation in a shared latent space: the image and audio are generated with mutual awareness, whereas pipeline systems generate each modality independently and can produce locally plausible but globally incoherent combinations.

5.2 Limitations and Future Research

TGI is trained and evaluated at 512x512 image resolution and 24kHz audio -- lower than the 1024x1024 or higher resolution supported by specialist image generation models, and lower than 48kHz studio quality audio. Scaling to higher output quality requires substantially larger decoder models and training budgets. The TIA-2M dataset, while the largest trimodal dataset constructed to date, is constructed from video content and may not adequately represent all trimodal generation use cases -- particularly those requiring non-naturalistic or abstract visual styles and non-speech audio modalities (music, environmental sound). Future research should extend TGI to video generation -- adding temporal coherence across frames to the trimodal integration challenge -- and to additional modalities such as depth, tactile, and physiological signals. The integration of TGI with the KILLM knowledge grounding architecture (Garcia et al., 2024) to produce knowledge-grounded trimodal generation -- where factual accuracy is maintained across all three generated modalities -- represents a particularly high-value research direction for reliable educational and informational content generation.

6. Conclusion

6.1 Summary

This paper proposed the TGI trimodal generative architecture integrating text, image, and audio generation through a shared 512-dimensional latent space, cross-modal attention transformer core, and modality-specific decoder triplet. The TIA-2M trimodal dataset (2.1M aligned triplets) and TGQ evaluation suite were introduced. TGI achieves CLIP-Audio alignment of 0.74 (21.3% over best pipeline), FID = 18.4 (competitive with specialist image generators), MOS = 4.2/5.0 (competitive with specialist TTS), and narrative coherence = 4.4/5.0 (significantly above pipeline baselines at 3.1/5.0, $p < 0.001$).

6.2 Open Source Release

The TGI model weights, TIA-2M dataset (with automated construction pipeline), and TGQ evaluation suite are released as open-source resources. The TGI inference API supports all four generation modes with configurable quality-latency trade-off (fast mode: shared latent + text/image decoders only, 800ms; full mode: all decoders + coherence check, 2.4s per generation). Implementation details and the TIA-2M construction pipeline are documented in Appendix A. We invite the community to extend TGI to additional modalities and higher output resolutions within the shared latent framework established here.

References

Alayrac, J. B. et al. (2022). Flamingo: A visual language model for few-shot learning. NeurIPS, 35.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

Defossez, A. et al. (2022). High fidelity neural audio compression. arXiv:2210.13438.

Elizalde, B. et al. (2023). CLAP: Learning audio concepts from natural language supervision. ICASSP 2023.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Girdhar, R. et al. (2023). ImageBind: One embedding space to bind them all. CVPR 2023.

Liu, H. et al. (2023). LLaVA: Visual instruction tuning. NeurIPS, 36.

Radford, A. et al. (2021). Learning transferable visual models from natural language supervision. ICML 2021.

Ramesh, A. et al. (2021). Zero-shot text-to-image generation. ICML 2021, 8821-8831.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR 2022, 10684-10695.

Rubenstein, P. K. et al. (2023). AudioPaLM: A large language model that can speak and listen. arXiv:2306.12925.

Saharia, C. et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS, 35.

Su, J. et al. (2024). RoFormer: Enhanced transformer with rotary position embedding. Neurocomputing, 568, 127063.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Wang, C. et al. (2023). VALL-E: Neural codec language models are zero-shot text to speech synthesizers. arXiv:2301.02111.

Ho, J. et al. (2020). Denoising diffusion probabilistic models. NeurIPS, 33.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS, 30.

Huang, R. et al. (2023). Make-an-audio: Text-guided audio generation with diffusion models. ICML 2023.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Radford, A. et al. (2023). Robust speech recognition via large-scale weak supervision. ICML 2023.

Declarations

Funding

This research was supported by the Swedish Research Council under grant 2024-05341 (TGI: Trimodal Generative Intelligence) and the French National Research Agency (ANR) under grant ANR-23-CE23-0038. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The TGI model weights, TIA-2M dataset (full and subset), TIA-2M construction pipeline code, and TGQ evaluation suite are publicly available. Generation samples for all evaluation conditions are available from the corresponding author.

Ethical Approval

The TIA-2M dataset was constructed from publicly available video content under Creative Commons licences. Human evaluation involved consenting participants compensated at above-minimum-wage rates. No personal data of AI-affected individuals were processed. Ethical approval reference: NTU-SD-2024-021.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2024, Issues: 1, 2024

Appendix A

TIA-2M Dataset Construction and TGQ Evaluation Suite

The following describes the TIA-2M automated construction pipeline and TGQ evaluation dimensions.

TIA-2M Dataset Construction Pipeline

TGQ Evaluation Suite -- Seven Quality Dimensions