Generative AI Across Modalities: Insights From Our Research on Domain-Aware Content Generation

Giovanna Castellano, Emanuele Colonna, Nicola Fanelli, Lucrezia Laraspata, Ivan Rinaldi, Albero G. Valerio and Gennaro Vessio*

Department of Computer Science, University of Bari Aldo Moro, Italy

Abstract

Generative AI is rapidly advancing, enabling the creation of diverse and high-quality content across multiple modalities, including text, images, audio, and video. However, much of this progress remains confined to general-purpose tasks, often neglecting real-world applications' unique complexities and domain-specific requirements. In this paper, we present a unified perspective on our research at the Computational Intelligence Laboratory (CILab), University of Bari Aldo Moro, aimed at advancing domain-aware generative AI across five core tasks: text-to-text, text-to-image, text-to-video, image-to-text, and image-to-audio. Our work introduces methodological innovations such as text-guided multi-mask inpainting with localized attention, explainable image-to-text pipelines for medical reporting, and generative sign language video synthesis using diffusion models. We further contribute new datasets and propose novel evaluation frameworks tailored to domain-specific constraints, including creativity, factual alignment, inclusivity, and human-centered assessment. Through applications in healthcare, cultural heritage, accessibility, and human capital management, we show how generative AI can be used for content creation and as a driver of creativity, inclusivity, and knowledge transfer. This paper argues for a shift toward more grounded, explainable, and domain-specific generative models. We aim to pave the way for impactful AI applications in complex, real-world scenarios by addressing key gaps in current practices.

Keywords

Generative AI, Multimodal Generation, Cross-Modal Learning, Large Language Models

1. Introduction

Generative AI has emerged as a transformative paradigm in machine learning, enabling systems to create novel content across various data modalities. From text and images to audio and video, generative models have demonstrated the ability to produce high-quality, diverse, and often surprising outputs, reshaping creative industries, communication tools, and decision-support systems. This progress has been largely driven by advances in large-scale pretraining, diffusion models, and multimodal architectures, which have unlocked new cross-modal generation and understanding capabilities.

However, despite these remarkable achievements, much of the generative AI landscape remains centered on general-purpose tasks and benchmark datasets, often neglecting the complexities and requirements of real-world, domain-specific applications. Domains such as healthcare, cultural heritage, accessibility, and human capital management present unique challenges, including the need for factual accuracy, interpretability, domain alignment, and creative adaptability beyond the biases of pretrained distributions.

Addressing these challenges requires moving beyond one-size-fits-all approaches and developing tailored methods to harness generative models' capabilities while embedding domain knowledge, multimodal reasoning, and user-centered evaluation strategies. It also demands a critical reflection on

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

giovanna.castellano@uniba.it (G. Castellano); emanuele.colonna@uniba.it (E. Colonna); nicola.fanelli@uniba.it (N. Fanelli); l.laraspata3@phd.uniba.it (L. Laraspata); i.rinaldi4@phd.uniba.it (I. Rinaldi); a.valerio31@phd.uniba.it (A. G. Valerio); gennaro.vessio@uniba.it (G. Vessio)

0000-0002-6489-8628 (G. Castellano); 0009-0009-0932-3424 (E. Colonna); 0009-0007-6602-7504 (N. Fanelli); 0009-0003-8136-9140 (L. Laraspata); 0009-0003-9597-064X (I. Rinaldi); 0009-0008-6101-965X (A. G. Valerio); 0000-0002-0883-2691 (G. Vessio)

the gaps left by current generation and evaluation practices, particularly in scenarios where creativity, semantics, and human-like interaction play a crucial role.

This paper presents a unified perspective on our research at the Computational Intelligence Laboratory (CILab), University of Bari Aldo Moro, where we explore the frontiers of generative AI across modalities. Our work spans five core generative tasks—text-to-text, text-to-image, text-to-video, image-to-text, and image-to-audio—each addressing a specific challenge in domain-aware content generation. We introduce methodological innovations, propose new datasets and evaluation frameworks, and demonstrate practical applications in diverse fields such as healthcare, cultural heritage, and accessibility.

Through these contributions, we aim to showcase the technical advances in multimodal generative AI and its potential to support creativity, inclusivity, and knowledge transfer in complex real-world scenarios.

2. Text-to-Text

Our research explored the capabilities of large language models (LLMs) for text-to-text generation, focusing on the automated creation of human resources (HR) questionnaires. Despite their widespread use as decision-support tools in organizations, survey questionnaire design remains a labor-intensive task, requiring meticulous attention to structure, logical flow, and relevance to ensure the collection of meaningful and unbiased data.

In [1], we addressed this gap by introducing a novel application of generative AI in the human capital management domain. Specifically, we contributed a new dataset of HR-related surveys and investigated the generative capabilities of two state-of-the-art GPT models. Our research also delved into critical aspects of LLM behavior, such as instruction alignment and generated content's semantic and stylistic fidelity.

Existing approaches to evaluating generated questionnaires predominantly rely on standard metrics borrowed from related tasks, including BLEU and ROUGE for syntactic similarity, and cosine similarity for assessing semantic overlap. However, these metrics fail to capture key attributes specific to questionnaires, such as their ability to engage respondents, the coherence of question flow, and the diversity of topics covered. To address these limitations, we designed a comprehensive evaluation framework that assesses generated questionnaires from multiple complementary perspectives: instruction alignment, syntactic and lexical diversity, semantic similarity to human-authored questions, and topic diversity (serendipity). In collaboration with Talentia Software, we further conducted human evaluation studies to assess the indistinguishability of AI-generated questionnaires from those created by experts. Our findings revealed that while both GPT models achieved high intra-questionnaire diversity and serendipity, human evaluators could still discern AI-generated content, particularly noticing stylistic inconsistencies and reduced variability in suggested answers.

These insights highlight the promise and current limitations of using LLMs for questionnaire generation. They underscore the need for further refinement to achieve outputs that are structurally sound and indistinguishable from human-created surveys. As part of our future work, we plan to explore reinforcement learning strategies to better align generated outputs with human preferences and enhance the explainability of LLMs in this sensitive application domain.

3. Text-to-Image

Our research on text-to-image generation focuses on the challenging task of image inpainting—specifically, completing missing or occluded regions of an image with semantically plausible and stylistically coherent content. Beyond conventional inpainting, our work addresses the more creative variant of object inpainting, where the goal is not merely restoration, but the imaginative addition of new elements that can transform the meaning or composition of an image.

In [2], we introduced a novel computer vision task: *text-guided multi-mask inpainting*, where multiple distinct regions within an image are simultaneously inpainted, each driven by its own textual prompt

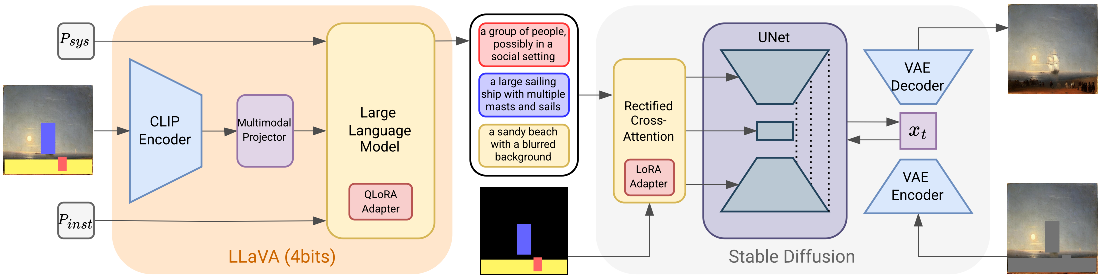

Figure 1: Overview of our text-guided multi-mask inpainting framework [2]. Our approach enables the simultaneous inpainting of multiple regions, each guided by an automatically generated textual prompt. An MLLM generates creative prompts directly from the corrupted image. At the same time, a rectified cross-attention mechanism ensures that each prompt exclusively influences its corresponding masked region, resulting in globally coherent and semantically enriched completions.

(Fig. 1). Although one might approach this problem by applying a standard inpainting model sequentially to each region, our experiments demonstrated that inpainting all regions concurrently leads to more globally coherent and harmonious outcomes.

A critical limitation of existing inpainting models, such as Stable Diffusion, is their tendency to produce generic or implausible completions when explicit textual guidance is absent or insufficiently rich. To overcome this, we leverage a multimodal large language model (MLLM) to automatically generate creative and context-aware prompts directly from the corrupted image. This process injects human-like imagination into the generation loop, helping to counteract biases and limitations inherited from pretraining on conventional datasets.

However, a key technical challenge arises in ensuring that each generated prompt influences only its intended region during inpainting, avoiding semantic leakage or interference between regions. To address this, we adapted the rectified cross-attention mechanism—originally proposed for freestyle layout-to-image synthesis—to the multi-mask inpainting setting. This mechanism enforces localized attention, ensuring the model conditions each region exclusively on its associated prompt, enabling precise and contextually appropriate generation.

We evaluated our approach on digitized paintings, which pose an especially demanding benchmark due to the need for creative reasoning and strict adherence to artistic styles. Our experiments showed that the proposed pipeline significantly improves over state-of-the-art baselines in terms of creativity, coherence, and stylistic consistency. Beyond creative editing, our method opens avenues for practical applications in art restoration and conservation, offering tools to assist historians and conservators in generating plausible reconstructions of damaged or missing regions, with the potential for fine-tuning on specific artists or art movements to ensure stylistic fidelity.

4. Text-to-Video

One of our research directions investigates the application of generative AI for enhancing communication accessibility, specifically through the automatic generation of sign language videos from written text. This area is significant for deaf and hard-of-hearing communities, for whom accessible communication technologies remain limited. Despite recent advances, bidirectional translation between spoken/written language and sign language continues to pose significant challenges due to sign language's unique linguistic, visual, and spatial complexity. Addressing these challenges requires integrating methods from computer vision, sequence modeling, and natural language processing.

In [3], we introduced a novel text-to-LIS (Italian Sign Language) generation model, moving beyond conventional avatar-based or rule-based approaches by leveraging generative AI techniques (Fig. 2). Our approach extends latent diffusion architectures, which have shown remarkable success in cross-modal generation tasks, adapting them to the specific challenges of sign language synthesis. The model

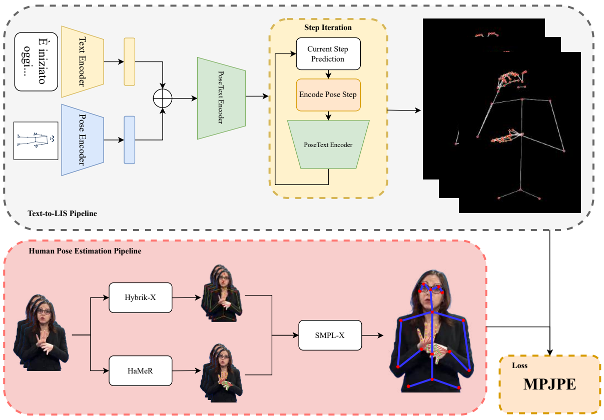

Figure 2: Overview of our text-to-LIS generation framework [3]. The system encodes written Italian into a latent space and generates LIS video sequences using a tailored diffusion process. Key components include temporal modeling to capture sign language's sequential nature and integrating non-manual features (e.g., facial expressions, body posture) to enhance grammatical correctness and naturalness.

encodes the semantic content of the input text into a latent space, from which continuous and expressive sign language videos are synthesized through a purpose-built diffusion process.

A key innovation of our framework lies in integrating temporal modeling components that capture the sequential dynamics intrinsic to sign language while maintaining spatial consistency critical to its grammar. Additionally, our model addresses essential linguistic phenomena such as co-articulation effects and incorporates non-manual features, including facial expressions and body posture, which are fundamental for conveying grammatical nuance and emotion in sign language communication.

We trained the model on a curated dataset comprising Italian text paired with LIS video annotations, enriched with temporal alignments and linguistic features to guide the generation process. Our text-to-LIS model represents an important step toward more inclusive and accessible communication technologies. We plan to extend this research toward bidirectional translation models and explore real-time applications that could support everyday communication scenarios.

5. Image-to-Text

While generative AI has made significant strides in medical imaging, the opacity of deep learning models remains a critical barrier to their widespread adoption in clinical practice. Clinicians often require transparent, interpretable outputs that align closely with medical standards and factual data. To address this challenge, we proposed an annotation-free, explainable image-to-text pipeline in [4], which integrates semantic segmentation, anatomical localization, and instruction-tuned language models to generate clinically meaningful reports directly from MRI scans (Fig. 3).

Our pipeline eliminates the need for annotated report datasets by introducing a structured, interpretable approach. First, the system performs 3D semantic segmentation to identify pathological regions, followed by anatomical localization using the Julich Brain Atlas to contextualize findings within precise brain structures. The extracted information—lesion location, volume, and segmentation confidence—is then encoded into a structured JSON format. This structured representation is a grounded prompt for an instruction-tuned LLM, ensuring the generated report is factual and aligned with the clinical data.

A key innovation of our method lies in this grounded prompting strategy, which significantly reduces hallucinations—a known limitation of LLMs in medical applications—and guarantees that the generated text reflects the actual imaging findings. Furthermore, our pipeline integrates an error mitigation strategy that prioritizes reporting on the five most affected anatomical regions, effectively reducing segmentation-related inaccuracies by over 96%.

We validated our method on two clinically relevant tasks: glioma segmentation and multiple sclerosis

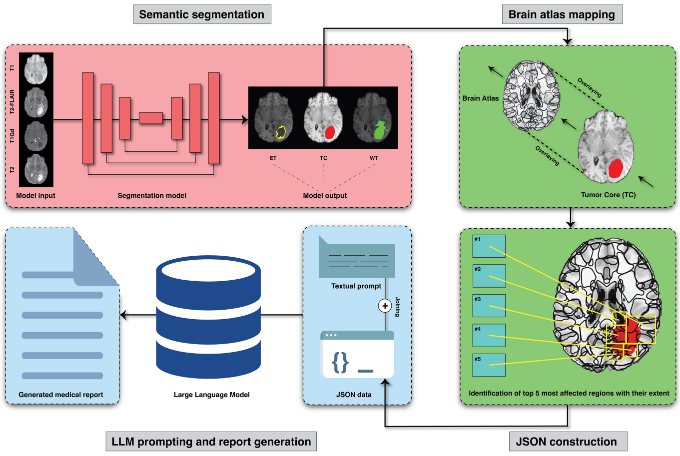

Figure 3: Overview of our explainable image-to-text pipeline for MRI report generation [4]. The system integrates 3D semantic segmentation, anatomical localization using the Julich Brain Atlas, and instruction-tuned LLMs. A structured JSON format encodes lesion characteristics, including spatial location, volume, and model confidence, serving as a factual grounding for report generation. This approach ensures clinically meaningful, human-readable outputs while minimizing hallucinations and improving transparency.

lesion detection. The results demonstrated high segmentation accuracy and consistent generation of clinically sound, human-readable reports. To rigorously assess the interpretability and factual alignment of the reports, we introduced novel evaluation metrics that measure coverage of critical findings, coherence of the narrative, and alignment with objective imaging data.

This work represents a step toward bridging the gap between black-box deep learning systems and the transparency requirements of clinical workflows, paving the way for safer and more trustworthy AI-assisted medical reporting.

6. Image-to-Audio

As part of our research into cross-modal generation, we explored the underinvestigated task of image-to-audio generation, focusing specifically on synthesizing music from digitized artworks. While generative models have recently achieved impressive results in text-to-music generation, leveraging complex visual stimuli such as artworks to condition music creation remains largely unexplored. This gap arises from the inherent challenges posed by artworks, which convey meaning, mood, and emotion through abstract elements like color, composition, and style, rather than through explicit, descriptive cues as found in text.

Recent advances in generative modeling, particularly latent diffusion models, have opened new possibilities for high-quality, efficient content generation across modalities. However, applying these models to the image-to-music domain introduces unique complexities, given visual art's subjective and deeply interpretive nature. Translating visual semantics into expressive and emotionally resonant music requires sophisticated multimodal understanding and bridging the cognitive gap between the visual and auditory domains.

To address these challenges, we introduced Art2Mus [5], a novel deep learning framework designed to generate music from digitized artworks, their textual descriptions, or a combination of both modalities. Alongside the model, we developed a new dataset of artwork-music pairs and proposed a domain-specific evaluation metric to assess the semantic and emotional alignment between the visual input and the generated music.

Art2Mus represents an initial step toward bridging the gap between visual art and music through generative AI. Potential avenues for extension include experimenting with alternative embedding models and pairing strategies to enhance cross-modal alignment. Furthermore, incorporating metadata—such as historical context, artist background, or the emotional tone associated with the artwork—could enrich

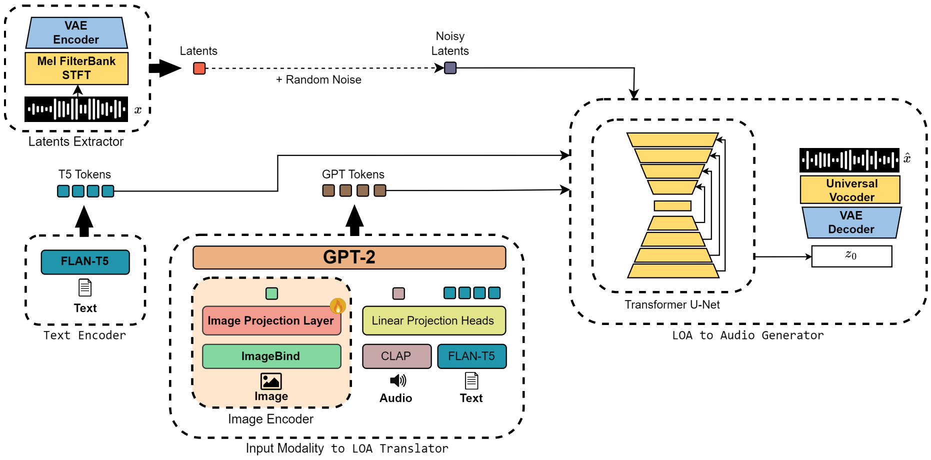

Figure 4: Overview of the Art2Mus framework [5]. The system leverages digitized artworks, optionally combined with textual descriptions, to condition a latent diffusion model for music generation.

the generation process, enabling more personalized and contextually aware musical outputs.

Acknowledgments

Giovanna Castellano acknowledges support from the NRRP project *FAIR – Future AI Research* (PE00000013), Spoke 6 – *Symbiotic AI* (CUP H97G22000210007), funded by NextGenerationEU.

Declaration on Generative AI

We used GPT-4o and Grammarly to assist with grammar refinement, spelling correction, and stylistic improvements during the preparation of this manuscript. All content was subsequently reviewed, revised, and approved by the authors, who take full responsibility for the final version of the work.

References

[1] L. Laraspata, F. Cardilli, G. Castellano, G. Vessio, Enhancing Human Capital Management through GPT-driven Questionnaire Generation, in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), CEUR-WS.org, 2024.

[2] N. Fanelli, G. Vessio, G. Castellano, I Dream My Painting: Connecting MLLMs and Diffusion Models via Prompt Generation for Text-Guided Multi-Mask Inpainting, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 6073–6082.

[3] E. Colonna, A. Arezzo, D. Roberto, D. Landi, F. Vitulano, G. Vessio, G. Castellano, Towards Italian Sign Language Generation for Digital Humans, in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), CEUR-WS.org, 2024.

[4] G. Castellano, S. de Benedictis, K. Trufanova, A. G. Valerio, G. Vessio, From Segmentation to Explanation: Generating Textual Reports from MRI with LLMs, Available at SSRN 4974224 (2024).

[5] I. Rinaldi, N. Fanelli, G. Castellano, G. Vessio, Art2Mus: Bridging Visual Arts and Music through Cross-Modal Generation, in: European Conference on Computer Vision, Springer, 2025, pp. 173–186.