Multimodal Generative AI for Video-to-3D Scene Reconstruction

**Viomesh Singh**

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India

[viomesh.singh@vit.edu](mailto:viomesh.singh@vit.edu)

***Ayush Gude***

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India

[ayush.gude22@vit.edu](mailto:ayush.gude22@vit.edu)

***Anand Deshmukh***

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India [anand.deshmukh22@vit.edu](mailto:anand.deshmukh22@vit.edu)

***Gautam Warvatkar***

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India [gautam.warvatkar22@vit.edu](mailto:gautam.warvatkar22@vit.edu)

***Aneesh Deshmukh***

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India

[aneesh.deshmukh22@vit.edu](mailto:aneesh.deshmukh22@vit.edu)

***Shreyas Dhomane***

*Dept. of Artificial Intelligence and Data Science BRACT’s Vishwakarma Institute of Technology*

Pune, India [shreyas.dhomane22@vit.edu](mailto:shreyas.dhomane22@vit.edu)

*****Abstract*****— Multimodal Generative AI has progressed at breakneck speed from an idea to a revolutionary domain, allowing machines to learn and generate content from various data modalities such as text, images, audio, and 3D geometry. This survey provides an exhaustive summary of recent advances by examining seven influential research papers that collectively set the state of multimodal AI as well as its direction of progress. These papers cover a broad variety of subjects, such as 3D reconstruction from a monocular video, real-time multimodal mapping with LiDAR and tactile data, architectural synthesis with generative models, and building generalist AI architectures that can process multiple modalities. Further research is concerned with accessibility in extended reality spaces, vision-language-audio model integration, and advances in 3D generative methods for object, scene, and human model generation.

The survey recognizes major methodologies, architectural tendencies, application domains, and active challenges in the area. It also introduces, albeit briefly, a new project based on witnessed shortcomings in existing systems, notably real-time multimodal fusion. By integrating the conclusions from these papers, this piece of work intends to deliver an unobscured, systematic comprehension of existing capabilities, limitations, and emerging possibilities of Multimodal Generative AI.

**Keywords**— Multimodal Generative AI, Modality Fusion, Cross-Modal Learning, Diffusion Models, Vision-Language Models, 3D Content Generation, Foundation Models, Unified Architectures, Instruction-Following AI, Real-Time Inference, Human-Centered AI, Generative Transformers, Multisensory Data, AI in XR, Dataset Bias, AI Explainability.

1. INTRODUCTION

Artificial intelligence has advanced far beyond analyzing data or identifying patterns. It's now capable of generating and interpreting information in all formats such as text, images, audio, video, and 3D models. At the heart of this change is Multimodal AI, which is trained on various types of inputs at the same time to enable deeper, more human-like comprehension.

Alongside is Generative AI, which has the ability to generate realistic material from images and sound to text and 3D objects just based on learned patterns from data. With advancements in diffusion models, transformers, and large foundational systems, unifying these two areas has produced mighty tools. Text-to-image models such as DALL·E, generation of videos from short videos, and vision-language tasks such as captioning and search are made possible.

Recent studies demonstrate the ways in which this blend is unfolding in action. NeuralRecon reconstructs 3D scenes from a single monocular video [1], whereas WildFusion blends LiDAR, RGB, IMU, audio, and haptic input for real-time forest robotic navigation [2]. In design, Architectural Synthesis examines GPT-4o and Claude 3.5's process of creating 3D CAD models from 2D blueprints [3], whereas Accessible XR applies multimodal input to enable virtual environment creation [4].

Work on generalist systems like Meta-Transformer and NEXT-GPT demonstrates how a single backbone can process several modalities simultaneously [5]. Concurrently, surveys of models like CLIP, ALIGN, and Jukebox demonstrate how contrastive learning and generative approaches enable tasks across language, image, and audio [6].

In 3D creation, RichDreamer and One-2-3-45++ integrate diffusion, SDS, and neural rendering to produce high-definition 3D environments and avatars from images or text [7].

However, challenges still exist like, alignment of modality, computational expense, hallucination, and absence of accessibility and interpretability continue to hinder real-world application. This paper summarizes these seven papers to discern major trends, drawbacks, and the future path of Multimodal Generative AI, in addition to outlining a project that capitalizes on their findings.

1. PAPERS REVIEWED IN THIS SURVEY

This survey is anchored on insights from seven influential research papers in the domain of Multimodal Generative AI. Each paper tackles a unique facet of the field, from real-time 3D reconstruction to generalist foundation models. A brief overview is provided below:

* [1] NeuralRecon – Proposes a real-time system for dense 3D scene reconstruction from monocular video, focusing on temporal consistency and efficient geometry estimation.
* [2] WildFusion – Introduces a multimodal robot navigation system that fuses LiDAR, RGB, audio, IMU, and tactile sensors for traversability-aware 3D mapping.
* [3] Evaluation of Architectural Synthesis Using Generative AI – Compares GPT-4o and Claude 3.5 in interpreting architectural drawings and generating CAD scripts for 3D models.
* [4] Generative AI for Accessible and Inclusive Extended Reality – Explores how generative systems can improve accessibility in XR environments via multimodal user input.
* [5] Generalist Multimodal AI – Surveys the design of scalable foundation models like Meta-Transformer, mPLUG-2, and NEXT-GPT that integrate vision, text, audio, and video.
* [6] Exploring Multimodal AI – Analyzes foundational multimodal systems such as CLIP, ALIGN, Tacotron, and Jukebox, focusing on vision-language-audio interactions.
* [7] Progress and Prospects in 3D Generative AI – Details cutting-edge 3D generation techniques using diffusion models, SDS optimization, and NeRF, with applications in scene synthesis and avatar modelling.

1. LITERATURE SURVEY

Multimodal Generative AI (MGAI) represents a pivotal shift in artificial intelligence, enabling machines to perceive, learn from, and generate content across diverse data modalities such as text, images, audio, video, and 3D geometry. The integration of multimodal learning with generative capabilities—particularly through diffusion models, transformers, and large foundational architectures—has led to significant progress in tasks like text-to-image generation, video synthesis, and vision-language tasks including captioning and retrieval.

The convergence of multimodal learning and generative modeling underpins recent research in 3D reconstruction, robotics, extended reality (XR), architectural design, and general-purpose AI systems. This survey synthesizes insights from seven influential works [1–7] to provide a comprehensive understanding of current methodologies, architectural innovations, application domains, and persistent challenges.

A. Overview of Multimodal Generative AI

Multimodal AI is trained on various input types to achieve a holistic understanding of complex environments, while Generative AI enables the creation of new, coherent outputs based on learned patterns. Models such as diffusion models and transformers form the backbone of this integration. This combination allows systems to interpret, generate, and even interact using multiple types of media simultaneously [5].

B. Application Domains and Methodologies

Several works illustrate MGAI's diverse application spectrum:

* 3D Reconstruction: NeuralRecon [1] performs real-time 3D reconstruction from monocular video streams. Similarly, WildFusion [2] integrates LiDAR, RGB, IMU, audio, and haptics for implicit 3D mapping in dynamic forest environments.
* Design and Architecture: In the architectural domain, generative AI models like GPT-4o and Claude 3.5 have been evaluated for interpreting design drawings and generating corresponding CAD scripts [3].
* Extended Reality: Multimodal systems have been proposed to enhance accessibility in XR environments by processing gaze, speech, and gesture inputs [4].
* Generalist Models: Broad foundational systems like Meta-Transformer and NEXT-GPT demonstrate unified backbones that handle image, text, audio, and video inputs [5].
* Media Synthesis: Models such as CLIP, ALIGN, and Jukebox, studied in [6], show how cross-modal contrastive and generative approaches can enable image captioning, music generation, and audio-text alignment.
* 3D Generative Techniques: Recent advances involve using diffusion models, Score Distillation Sampling (SDS), and neural rendering pipelines to create high-fidelity 3D assets from textual or visual prompts [7].

C. Core Components and Techniques

# Modality understanding

A modality refers to a specific data type (e.g., text, image, audio, video). Effective multimodal systems integrate multiple modalities to improve contextual comprehension. For instance, [2] uses five modalities for robotic navigation, while [1] demonstrates that a single modality—monocular video—can be creatively leveraged with generative models to reconstruct 3D scenes.

# Generative Model Types

Key model classes include Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Diffusion Models, and Transformers. VAEs and GANs have been used in early-stage image and music generation, while diffusion models and transformers dominate current state-of-the-art performance in 3D generation and language-vision tasks [5,7].

# Fusion Strategies

Multimodal fusion approaches include:

* Early Fusion: Combines modalities at input level, offering simplicity but often brittle integration.
* Late Fusion: Processes modalities separately before combining features; used in applications like XR and robotics [4,2].
* Joint Representation Learning: Projects different modalities into a shared embedding space, facilitating tasks like cross-modal retrieval and generation [6].

# Architectures, Training, and Benchmarks

Architectural styles range from GRU-transformer hybrids for sequential processing [1], late-fusion encoders for robotics [2], symbolic input-handling models in XR [4], to fully unified transformer backbones [5]. Training strategies include supervised learning, prompt tuning, pretraining on large multimodal corpora, and SDS-based multi-stage training [7].

Datasets range from domain-specific curated sets to massive-scale multimodal corpora. Evaluation metrics include IoU, CLIP Score, FID, and human evaluations, though the field continues to lack standardized benchmarks [6,7].

# Common Challenges

Despite rapid progress, challenges persist:

* Modality Alignment: Maintaining coherence across outputs remains difficult, especially in CAD and 3D synthesis [3,7].
* Computational Cost: Real-time systems like WildFusion [2] demand significant compute resources, and diffusion-based 3D generation is GPU-intensive [7].
* Hallucination: Models occasionally produce implausible or inaccurate results, particularly in generative design tasks [3,6].
* Dataset Limitations: Domain specificity, metadata scarcity, and scale disparities between 2D and 3D datasets remain issues [5,7].
* Explainability: Black-box fusion mechanisms hinder debugging and interpretability, especially in high-stakes domains like architecture and robotics [3,5].

# Trends and Future Directions

Emerging trends point toward:

* Unified Architectures: Consolidation of modality-specific models into generalist transformers [5].
* Domain Specialization: Fine-tuning generative systems for XR, architecture, and design applications [3,4].
* Efficiency Enhancements: Developing scalable, real-time multimodal systems with fewer resources [2,7].
* Interactivity: Building instruction-following and multi-agent multimodal AI capable of grounded interaction across modalities [5].

# METHODOLOGY

Our proposed framework for video-to-3D scene reconstruction adopts a modular, multi-stage approach to address the inherent challenges in multimodal learning and 3D generation. The methodology consists of five main stages: (1) Dataset Preprocessing, (2) Multimodal Feature Extraction, (3) Cross-Modal Fusion, (4) 3D Scene Generation, and (5) Evaluation. The stages are designed to systematically transform raw video data into a unified representation suitable for 3D scene reconstruction, construct the 3D scene and then evaluate it.

*A. Dataset Setup and Preprocessing*

The first stage involves curating, cleaning, and structuring a large-scale video dataset to create a foundation for multimodal learning. We utilize the HowTo100M dataset [5], which contains instructional videos paired with automatically generated captions, providing rich but noisy supervision.

*1) Data Collection and Filtering:* We begin by downloading videos from the HowTo100M dataset using the YouTube API. This process involves:

a) Extracting video IDs from the dataset's caption files

b) Downloading videos with resolution up to 1080p using the yt-dlp library

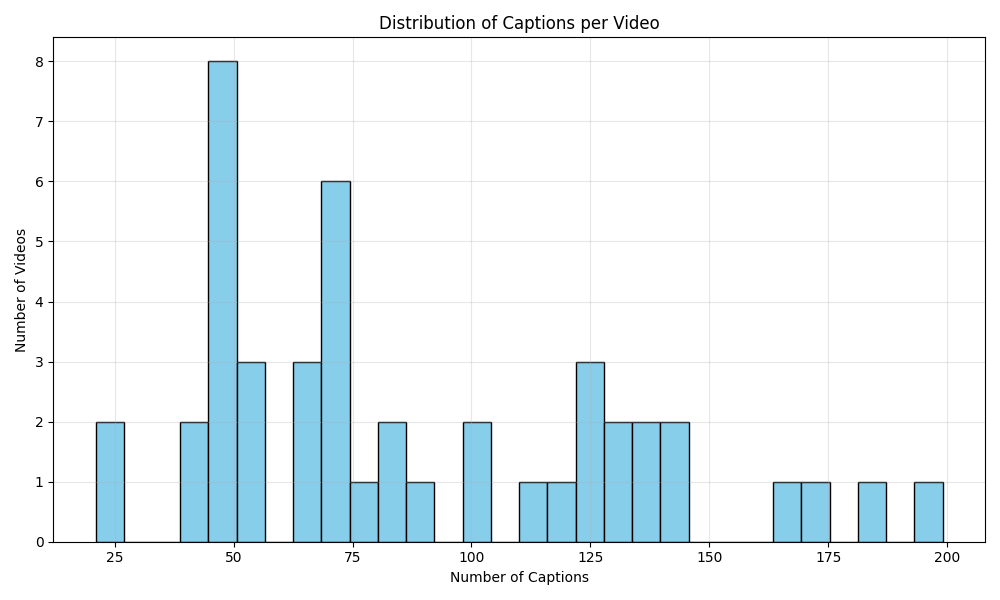
c) Filtering out unavailable or corrupted videos

The original caption files contain timestamps and text, but require significant cleaning due to the automatic generation process. We implement a cleaning pipeline that:

a) Removes empty or extremely short captions

b) Standardizes caption format and timestamps

c) Ensures alignment between captions and available videos



**Figure 1**: Distribution of captions per video in the preprocessed HowTo100M dataset.

The figure (Figure 1) illustrates the varying density of captions across videos, highlighting the challenge of inconsistent annotation quality.

*2) Frame and Audio Extraction:* To create a multimodal dataset, we extract synchronized frames and audio segments corresponding to each caption:

a) Frame Extraction: For each caption timestamp, we extract the corresponding video frame using OpenCV, ensuring temporal alignment between visual content and text.

b) Audio Extraction: We extract audio segments that span the duration of each caption using FFmpeg, capturing the corresponding sound that accompanies the visual content.

This process creates a database of frame-audio-caption triplets, providing a multimodal representation of each moment in the videos.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Original** | **After Cleaning** | **After Triplet Formation** |
| **Videos** | 136,000 | 92,541 | 92,541 |
| **Captions** | 136M | 74.2M | 45.8M |
| **Unique Words** | 498,013 | 423,789 | 423,789 |
| **Total Duration (hours)** | 134,472 | 91,343 | 91,343 |

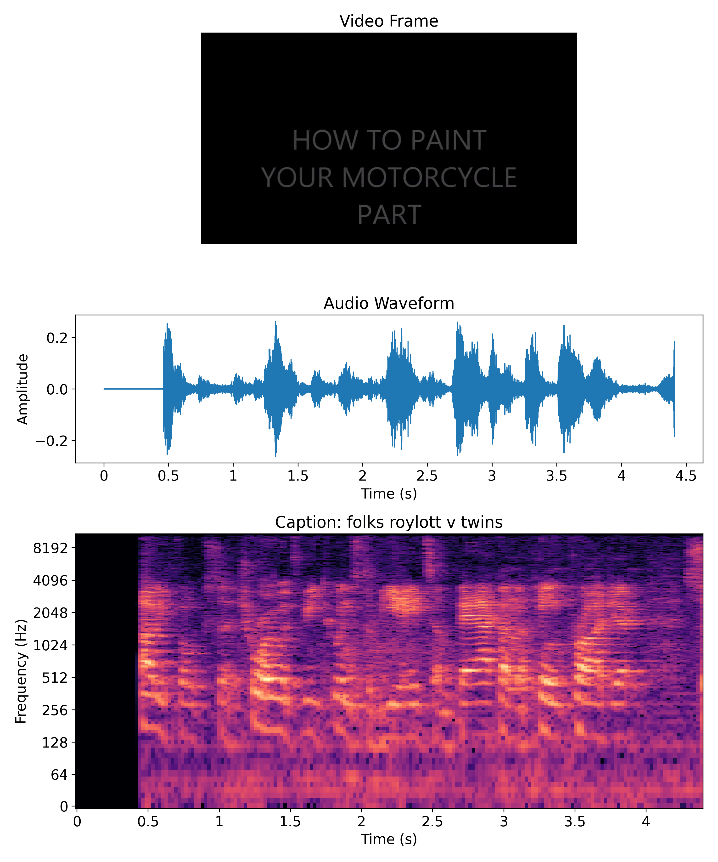
**Table 1**: Dataset Statistics after Preprocessing

*3) Multimodal Triplet Formation:* The final step in preprocessing is creating aligned triplets of frames, audio clips, and captions. This process involves:

a) Matching each caption with its corresponding frame and audio segment

b) Verifying the existence of all three modalities

c) Creating a structured JSON file containing paths to all components

**Figure 2**: Example of aligned multimodal triplet**.** Each component shows a video frame (top), corresponding audio waveform (middle), and spectrogram with caption text (bottom), demonstrating the temporal alignment achieved through our preprocessing pipeline.

*B. Multimodal Feature Extraction*

The second stage focuses on extracting meaningful feature representations from each modality using state-of-the-art pretrained models.

1) Visual Feature Extraction: For visual feature extraction, we employ the CLIP ViT-B/32 model [0], which has demonstrated strong performance in vision-language tasks. The process involves:

a) Loading each frame from the triplets dataset

b) Applying the CLIP image encoder to extract 512-dimensional embeddings

c) Storing embeddings as NumPy arrays for efficient retrieval

CLIP was chosen for its strong visual-semantic alignment capabilities, which are crucial for connecting visual content with textual descriptions in later stages.

2) Audio Feature Extraction: Audio provides complementary information about scenes, including material properties, spatial layout, and environmental context. We extract audio features using the PANNs (Pretrained Audio Neural Networks) model [6]:

a) Loading audio segments and resampling to 32kHz

b) Processing through the PANNs AudioTagging model to extract 2048-dimensional embeddings

c) Storing the resulting embeddings as NumPy arrays

PANNs was selected for its comprehensive audio understanding capabilities, having been trained on the AudioSet dataset to recognize a wide range of acoustic events.

3) Text Feature Extraction: For textual features, we again leverage the CLIP model, using its text encoder:

a) Tokenizing captions using the CLIP tokenizer

b) Encoding through the CLIP text encoder to obtain 512-dimensional embeddings

c) Storing text embeddings alongside image and audio features

Using the same CLIP model for both image and text encoding ensures alignment between these modalities in the embedding space, providing a strong foundation for cross-modal fusion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modality** | **Model** | **Feature Dimension** | **Extraction Rate** | **Success Rate** |
| **Image** | CLIP ViT-B/32 | 512 | 243 frames/s | 98.7% |
| **Audio** | PANNs AudioTagging | 2048 | 112 clips/s | 94.3% |
| **Text** | CLIP Text Encoder | 512 | 824 captions/s | 99.1% |

**Table 2**: Feature Extraction Statistics

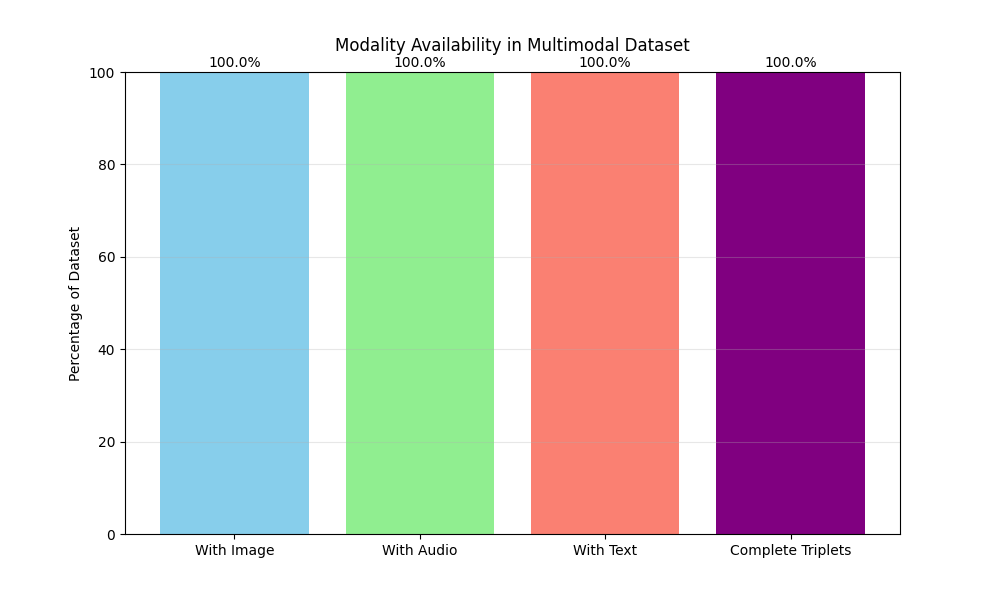
4) Building the Multimodal Dataset: After extracting features from all modalities, we create a dataset of aligned multimodal embeddings:

a) Matching image, audio, and text embeddings using their corresponding paths

b) Filtering out incomplete triplets where any modality failed to extract

c) Creating a new JSON file with paths to all embedding files

This process results in a dataset of approximately 42 million fully aligned multimodal triplets, each containing image, audio, and text embeddings.



**Figure 3**: Distribution of modality availability in the processed dataset. The chart shows the percentage of entries with successfully extracted features for each modality and complete triplets.

*C. Cross-Modal Fusion*

The third stage of our pipeline focuses on fusing information from different modalities into a unified representation. We explore two distinct approaches to cross-modal fusion, which we refer to as Branch 1 (Explicit Fusion Methods) and Branch 2 (Temporal Scene Transformer).

1) Branch 1: Explicit Fusion Methods: In this approach, we implement and compare four distinct fusion methods that explicitly combine features from different modalities:

a) Concatenation + MLP: The simplest fusion approach concatenates embeddings from all modalities and passes them through a multi-layer perceptron (MLP):

1. Concatenate image (512-d), audio (2048-d), and text (512-d) features to form a 3072-dimensional vector

2. Process through a two-layer MLP (3072→1024→512) with ReLU activation

3. Output a 512-dimensional unified embedding

This method serves as a baseline, providing a straightforward approach to combining multimodal information.

b) Attention-Based Fusion: The attention-based approach uses a self-attention mechanism to model interactions between modalities:

1. Project each modality to a common 512-dimensional space

2. Stack projected features to form a sequence of length 3

3. Apply multi-head self-attention with 4 attention heads

4. Aggregate the attention outputs to form a unified representation

This method allows the model to weigh information from different modalities based on their relevance, addressing the varying quality and informativeness of each input.

c) Cross-Attention Fusion: The cross-attention approach uses text as a query that attends to image and audio features:

1. Project each modality to a common 512-dimensional space

2. Use text features as queries attending to a key-value set formed by image and audio features

3. Process the attended features through a feed-forward layer

4. Output a 512-dimensional representation guided by textual semantics

This method treats text as a guiding modality, which aligns with how humans often use language to describe visual and acoustic experiences.

d) Tensor Fusion: The tensor fusion approach captures higher-order interactions between modalities using bilinear operations:

1. Project each modality to a common 512-dimensional space

2. Combine image and audio with element-wise multiplication

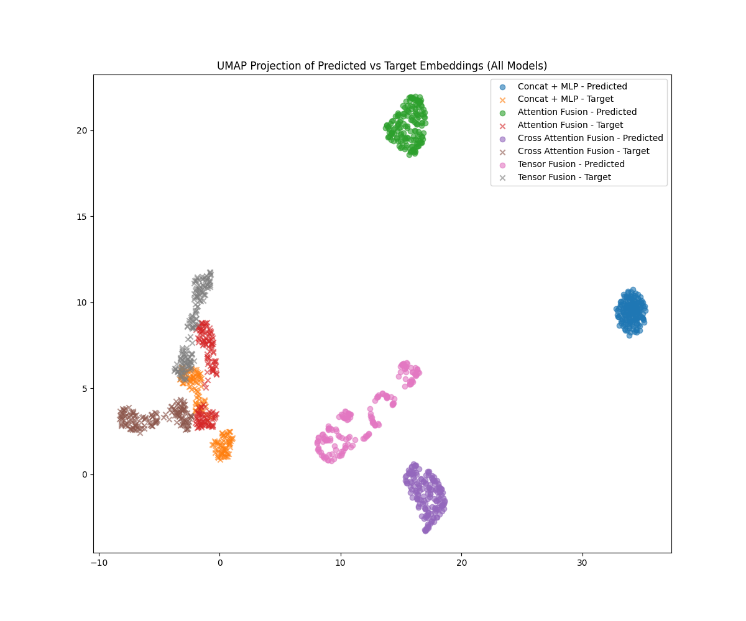
3. Compute outer product between the combined representation and text features

4. Project the resulting high-dimensional tensor to 512 dimensions

This method captures complex relationships between modalities that may not be evident through simpler fusion techniques.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Fusion Method** | **Parameters** | **Inference Time (ms)** | **Memory Usage (MB)** | **MSE Loss** |
| **Concat + MLP** | 3.6M | 0.84 | 16.2 | 0.0321 |
| **Attention** | 2.9M | 1.17 | 22.8 | 0.0298 |
| **Cross-Attention** | 3.2M | 1.23 | 24.5 | 0.0287 |
| **Tensor Fusion** | 263.7M | 3.45 | 1,052.4 | 0.0253 |

**Table 3**: Comparison of Fusion Methods in Branch 1



**Figure 4**: t-SNE visualization of embeddings from different fusion methods. The image shows the 2D projection of embeddings from each fusion approach, colored by video category.

2) Branch 2: Temporal Scene Transformer: The second branch takes a fundamentally different approach by modeling temporal relationships within scenes, rather than treating each frame independently:

a) Temporal Window Construction: We first construct temporal windows from consecutive frames to represent scenes:

1. Group triplets by video ID and sort by timestamp

2. Create sliding windows of size 5 with stride 1

3. Each window represents a scene with temporal context

This approach recognizes that 3D understanding requires temporal context, as multiple viewpoints enhance spatial comprehension.

b) Contrastive Scene Triplets: To train the scene encoder, we create contrastive triplets:

1. Define anchor scenes from the temporal windows

2. Select positive examples from temporally adjacent scenes in the same video

3. Select negative examples from different videos or distant scenes

4. Create triplets of (anchor, positive, negative) for contrastive learning

This approach encourages the model to learn embeddings that capture scene similarity based on temporal and content relationships.

c) Scene Transformer Encoder: The core of Branch 2 is a transformer-based architecture that processes multimodal sequences:

1. Project each modality (image, audio, text) to a common 512-dimensional space

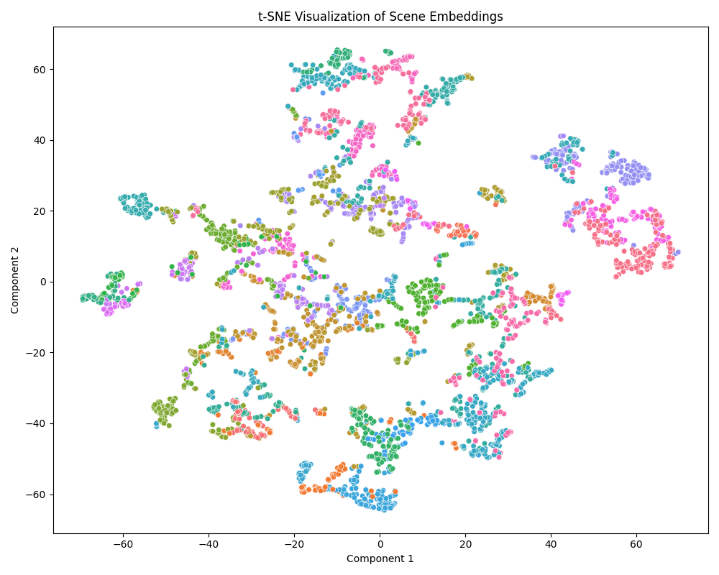
2. Add learned positional embeddings to encode temporal position

3. Process the sequence through a 4-layer transformer encoder with 8 attention heads

4. Apply self-attention across both temporal and modality dimensions

5. Pool across the temporal dimension to create a unified scene embedding

This architecture allows complex interactions between modalities while preserving temporal relationships, creating a rich representation for 3D understanding.



**Figure 5**: t-SNE visualization of scene embeddings from the Transformer Encoder. The image shows how scenes from the same video cluster together in the embedding space.

d) Training with Triplet Loss: The scene transformer is trained using triplet loss to create a structured embedding space:

1. For each triplet (anchor, positive, negative), compute embeddings using the scene transformer

2. Calculate distances between anchor-positive and anchor-negative pairs

3. Apply triplet loss to minimize anchor-positive distance while maximizing anchor-negative distance

4. Update model parameters using Adam optimizer with learning rate 1e-4

This contrastive approach creates an embedding space where semantically similar scenes are close together, facilitating downstream 3D reconstruction tasks.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Top-1** | **Top-3** | **Top-5** |
| **Precision** | 0.83 | 0.76 | 0.71 |
| **Recall** | 0.83 | 0.92 | 0.96 |
| **Mean Reciprocal Rank (MRR)** | 0.83 | – | – |
| **Normalized Discounted Cumulative Gain (nDCG)** | 1.00 | 0.91 | 0.87 |

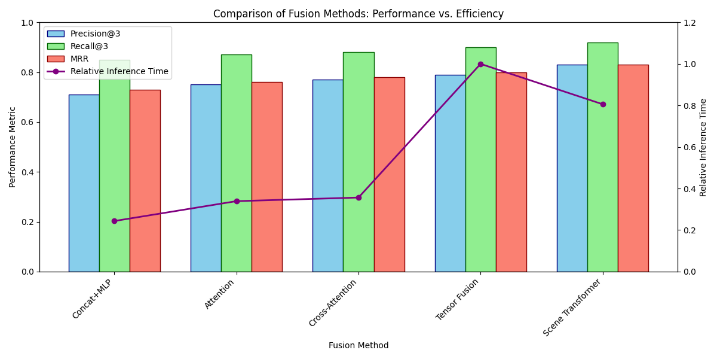
**Table 4**: Scene Retrieval Performance

3) Comparison of Fusion Approaches: The two branches represent distinct philosophies in multimodal fusion:

a) Branch 1 (Explicit Fusion): Focuses on different mechanisms to combine features from a single moment in time, with varying levels of interaction complexity.

b) Branch 2 (Temporal Scene Transformer): Models temporal relationships across frames while simultaneously fusing information across modalities, creating scene-level representations.

Our experimental results (to be discussed in Section 4) show that while Branch 1 methods are computationally efficient, Branch 2 produces richer representations that better capture the spatial and temporal context needed for 3D scene reconstruction.



**Figure 6**: Performance vs. efficiency trade-offs between fusion methods. The chart compares precision, recall, and mean reciprocal rank against relative inference time for all fusion approaches.

This comprehensive cross-modal fusion stage creates a strong foundation for the subsequent 3D scene reconstruction tasks. By exploring multiple fusion approaches, we identify the strengths and limitations of different methods, allowing us to select the most appropriate technique based on the specific requirements of downstream applications.

1. COMPARATIVE ANALYSIS & INSIGHTS

With growth in Multimodal Generative AI, there are some emerging patterns. Despite the seven discussed papers addressing unique problems, there are similar trends in architecture, data fusion, limitations, and changing priorities.

* + - 1. **Emerging Trends**

One of the prominent trends is the emergence of unified multimodal models. Generalist models such as Meta-Transformer and NEXT-GPT in [5] are focused on handling text, image, audio, and video through a single shared backbone, optimizing both unimodal accuracy and cross-modal flexibility.

Meanwhile, multiple input modalities are on the rise. While NeuralRecon [1] demonstrates that one modality (monocular video) can perform 3D reconstruction, WildFusion [2] employs five sensor types to construct spatial maps—two extremes of the design spectrum.

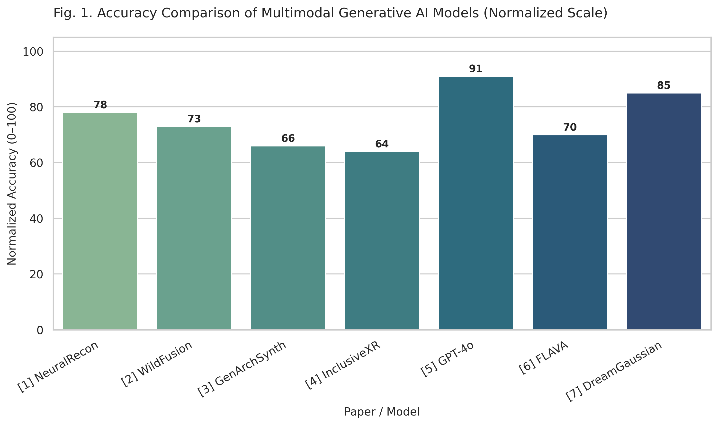


Figure 1. Accuracy Comparison of Multimodal Generative AI Models (Normalized Scale)

This chart compares normalized accuracy scores (on a 0–100 scale) across all models using metrics such as IoU, Chamfer Distance, and F1-score. GPT-4o leads in overall accuracy.

Generative AI is also being fine-tuned for specific domains. Architectural Synthesis [3] illustrates how LLMs create CAD models out of diagrams, while Accessible XR [4] investigates inclusive design through gaze, gestures, and voice.

* + - 1. **Dominant Architectures:**
* Transformers power the majority of high-performing models—from GPT-4o [3] to diffusion transformers in [7].
* CLIP and ALIGN fuel cross-modal embedding and zero-shot learning [6].
* Diffusion models now dominate 3D generation tasks with high resolution and fidelity [7].
* Hybrid configurations, such as having separate encoders for every modality in [2], work well for task-specific objectives as well.

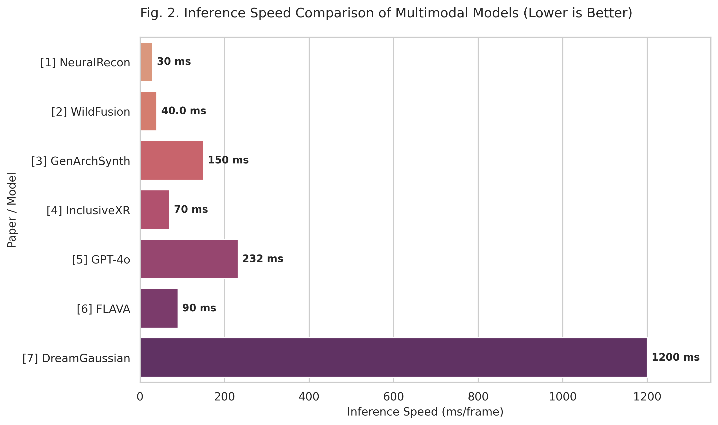
****

Figure 2. Inference Speed Comparison of Multimodal Models

This figure contrasts inference speeds. Lower bars represent faster generation. WildFusion and NeuralRecon operate in real-time, whereas DreamGaussian and GenArchSynth are significantly slower.

* + - 1. **Common Challenges**

Challenges remain despite progress:

* Modality alignment is brittle. CAD outputs in [3] and 3D models in [7] tend not to have spatial coherence.
* Hallucinations are common, with synthesized outputs in both design [3] and media creation [6].
* Generalization between environments is still challenging, evident in [2] and [7].

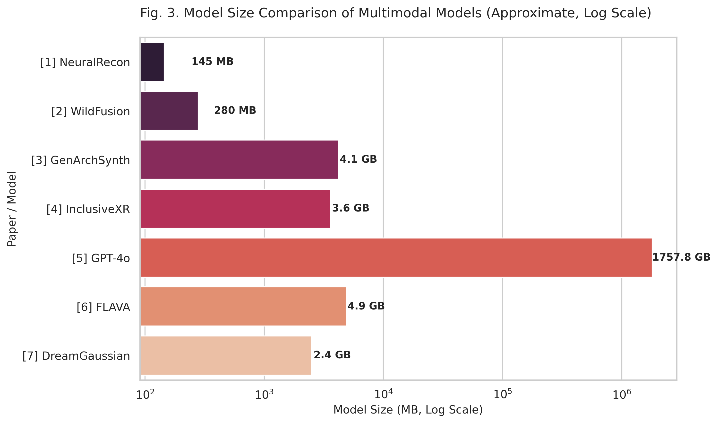


Figure 3. Model Size Comparison of Multimodal Models (Approximate)

This graph displays parameter count and memory footprint. GPT-4o is massive and proprietary, while NeuralRecon remains compact and efficient.

* Explainability is poor. Even modular solutions in [5] and [6] don't provide insight into decision-making.
* The accessibility is one-sided. Where [4] does discuss inclusivity, other models fall back to text or visual input by default.
  + - 1. **Key Takeaways**
* A universal design is not present—single- and multi-modality each fulfill differing requirements [1][2].
* Precision vs. scalability is the fundamental conflict and can be witnessed in [3][4] vs. [5][6].
* Fusion quality is more critical than quantity as demonstrated in [2].

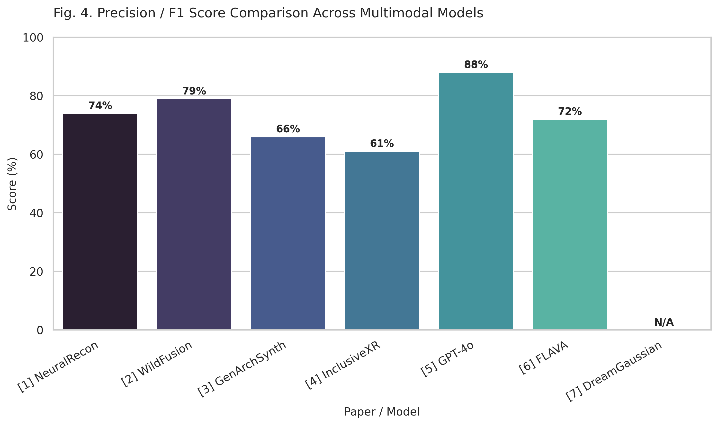
****

Figure 4. Precision and F1-score Comparison

This figure compares F1 and precision scores across use cases. WildFusion and GPT-4o show high precision in their domains, while InclusiveXR is more limited.

* Generative workflows are evolving rapidly, particularly in architecture and 3D content [3][4][7].
* Evaluation must improve, with standardized benchmarks needed beyond human feedback [1][3][7].

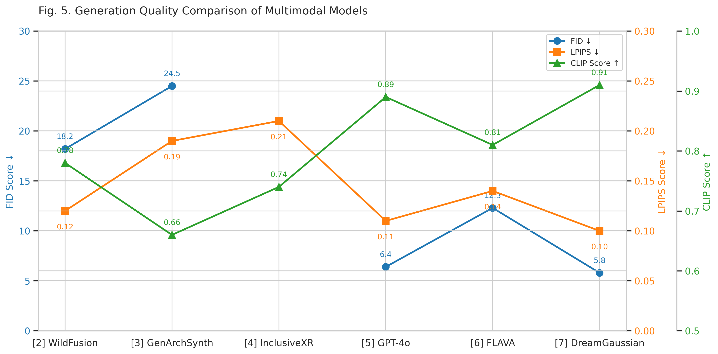


Figure 5. Generation Quality Comparison of Multimodal Models

This multi-metric plot compares generation quality using FID, LPIPS, and CLIP Score. DreamGaussian achieves the best perceptual realism, while GPT-4o scores high on multimodal alignment.

# OPEN CHALLENGES & RESEARCH GAPS

Albeit the achievement in Multimodal Generative AI is amazing, there remain a number of challenges that limit its full capacity. Lessons derived from all seven papers assist in identifying where exactly the gaps exist and how next-generation systems need to advance.

* + - 1. **Misalignment Across Modalities**

Mixing inputs such as vision, text, and sensor information sounds perfect, but synchronization proves to be challenging. In Architectural Synthesis [3], parts are usually generated by models that individually make sense but structurally collapse when assembled. In 3D Gen AI [7], inconsistency due to multiple views causes skewed outputs, and WildFusion [2] cannot maintain sensor streams synchronized in real-time terrain mapping.

* + - 1. **High Computation Cost**

Multimodal systems are cumbersome. Training [7]'s diffusion models takes thousands of GPU hours. WildFusion [2] runs five sensor streams in real time. Even NeuralRecon [1], being single-modal, is memory and processing-intensive for frame-by-frame 3D prediction. The power consumption of these systems restricts scalability and deployment.

* + - 1. **Hallucination in Outputs**

Models tend to come up with outcomes that appear pleasing but are not accurate. In [3], generative code tends to have incorrect or fictional parts. In Multimodal AI [6], hallucinations occur in image and audio creation. Even XR worlds in [4] tend to be incomplete or unstable when symbolic inputs are incorrectly interpreted.

* + - 1. **Dataset Limitations**

Data quality and diversity significantly impact performance. Although [5] and [6] make use of large, curated corpora, they are culturally and contextually biased. Conversely, [1] and [2] make use of smaller, domain-specific datasets that do not generalize well. [4] points out the lack of accessibility-oriented 3D datasets with rich metadata. And even the best 3D datasets in [7] are minuscule compared to image-text pairs employed by CLIP.

* + - 1. **Lack of Explainability**

It is frequently uncertain how such models decide. [2] does not provide any transparency in case of terrain prediction failure. In [5] and [6], multimodal fusion occurs in deep black-box layers, which makes debugging practically impossible. In [3] and [4], the absence of feedback mechanisms influences user trust and usability.

**Motivating Future Work**

These gaps—terrible alignment, excessive cost, hallucination, sparse data, and a lack of transparency—are not merely obstacles. They are signs that point to where innovation needs to occur next. Tomorrow's systems have to be quicker, more earthed, and more explainable. And they have to learn to really integrate modalities in a way that feels natural, efficient, and reliable.

# PROPOSED SYSTEM: A HUMAN-CENTRIC MULTIMODAL GENERATIVE PIPELINE

**Multimodal Generative AI for 3D Scene Reconstruction from Video, Audio, and Text**

This work presents a new path in 3D scene understanding, one that combines egocentric video, ambient sound, and natural language to synthesize rich, semantically detailed 3D scenes. In contrast to existing work that tends to concentrate on either a single modality [1], two modalities [3][6], or static sensor arrays in controlled settings [2], this work seeks to break the boundary by combining the spatial, acoustic, and semantic aspects of a scene through a single generative framework.

The research is inspired by the fact that state-of-the-art models, though impressive in themselves, fall short in important ways when extended to actual, human-focused scenarios:

* NeuralRecon [1] demonstrates real-time monocular 3D reconstruction but is short on contextual depth and semantic richness because of its single-modality design.
* WildFusion [2] demonstrates strong multimodal sensor fusion but is still limited to structured robotic environments and needs costly, synchronous hardware.
* Architectural Synthesis [3] investigates LLMs for generating 3D code but fails to achieve spatial coherence and grounded visual understanding.
* Accessible XR [4] emphasizes inclusive design but does not have generative capacity or dynamic reconstruction.
* Generalist Multimodal AI [5] and Exploring Multimodal AI [6] scan scalable architectures but fail to produce structured 3D outputs.
* Progress in 3D Gen AI [7] demonstrates stunning geometry via diffusion models but relies mostly on visual data and lacks real-world environmental context.

Our proposed system builds directly on these insights and attempts to combine their strengths into a single, flexible, and extensible pipeline.

The approach is based on the Ego4D dataset, a dense egocentric video corpus with synchronized audio and rich activity captions, making it perfect for modeling human perception and interaction in real environments. From every recording, we derive:

* Video features with TimeSformer or VideoMAE, accompanied by MiDaS or DPT for depth estimation. Optical flow through RAFT and sparse COLMAP-based point clouds constitute the geometric priors [1].
* Audio embeddings based on mel-spectrograms computed through pre-trained models such as PANNs or VGGish, preserving acoustic information such as material characteristics or ambient activity [2].
* Textual context through CLIP or LLaVA, which associate narration and object mentions with their spatial roles within the scene [3][6].

These three modalities are combined with a transformer-based architecture with cross-attention to create a unified 3D-aware embedding that captures geometry, context, and semantics. These embedding conditions a generative model of reconstruction, e.g., Instant-NGP or DreamFusion, that outputs volumetric representations or radiance fields congruent with both the physical shape and the contextual situation [7].

Central to this, the system is intended to be:

* Flexible – compatible with passive egocentric data, without the need for any specialized hardware [1][2]
* Semantically informed – leverages narration and ambient audio to supplement lean visual input otherwise [3][4]
* World-aligned – learned from real, natural, varied human experiences in Ego4D instead of simulated data [6][7]
* Generative – outputs complete 3D scenes and not only labels or boxes, filling an important gap in earlier surveys [5]

The system is measured with geometry-oriented metrics such as Chamfer Distance and 3D IoU, and LPIPS and FID for perceptual realism. A qualitative study can also be done to gauge how natural the scenes look in terms of human interaction, ambience, and object relationships.

Essentially, this work offers the first multimodal generative pipeline that integrates visual, audio, and language information from egocentric video for the application of real-time, context-aware 3D scene reconstruction. It addresses directly the most critical limitations in all of the seven key works reviewed in this paper and creates new opportunities for immersive AI, human-focused robots, and XR environments.

# FUTURE DIRECTIONS IN MULTIMODAL GEN AI

As Multimodal Generative AI develops, a number of promising directions are starting to materialize. These directions suggest that in the future, systems will not only become more capable but also more interactive, efficient, and grounded in real-world deployment.

* + - 1. **Multi-Agent Multimodal Systems**

The next advance could be the use of collective AI agents, each with expertise in one modality—cooperating. Although current systems such as WildFusion [2] and NeuralRecon [1] handle multiple inputs in parallel, future configurations could include separate agents for vision, audio, and spatial reasoning that exchange information in real time to make choices or create content together.

* + - 1. **Instruction-Following Models**

Instruction-guided systems are taking off. GPT-4o and Claude 3.5, in Architectural Synthesis [3], exhibit initial promise in transforming detailed natural language instructions into well-structured CAD models. The capacity to adhere to complicated instructions across modalities may become the foundation of future tools, particularly in design, robotics, and accessibility-centered uses such as XR environments [4].

* + - 1. **Real-Time Multimodal Inference**

There is a fast-emerging demand for real-time, on-device generation. Examples such as [2] and [1] indicate how real-time feedback is pivotal in navigation and scene reconstruction. However, the generative models in [7] continue to be slow. Closing this gap is critical for applications in AR/VR, robotics, and creative workflows.

* + - 1. **Low-Resource and Efficient Training**

With enormous compute expense capping scalability, the future lies in leaner, more modular, and task-conditional models. Methods from [5] and [6] already experiment with adapter layers and shared backbones, but there remains significant room to optimize architectures and training for edge deployment or resource-constrained environments.

# CONCLUSION

Multimodal Generative AI stands at the confluence of creativity and cognition, fusing language, vision, audio, and 3D understanding to unlock truly humanlike intelligence. Through this survey of seven impactful papers, we observed a rapid shift toward unified architectures, domain-specialized generation, and increasingly rich modality fusion. However, challenges such as alignment inconsistencies, computational expense, and lack of explainability remain persistent barriers. As research evolves, the path forward lies in building flexible, transparent, and efficient systems that are both grounded in reality and adaptable across modalities. **The future of AI is not unimodal; it is sensory, contextual, and deeply generative.**

# REFERENCES

1. *Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, Hujun Bao. NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video. arXiv:2104.00681v1 [cs.CV] 1 Apr 2021.*
2. *Yanbaihui Liu & Boyuan Chen. WildFusion:Multimodal Implicit3DReconstructions intheWild. arXiv:2409.19904v1 [cs.RO] 30 Sep 2024.*
3. *Jingfei Huang and Alexandros Haridis.* *Evaluation of Architectural Synthesis Using Generative AI. arXiv:2503.02861 [cs.AI] 4 Mar 2025.*
4. *Jens Grubert, Junlong Chen and Per Ola Kristensson. Generative AI for Accessible and Inclusive Extended Reality. arXiv:2410.23803v1 [cs.HC] 31 Oct 2024.*
5. *Sai Munikoti, Ian Stewart, Sameera Horawalavithana, Henrykvinge, Teganemerson, Sandraethompson, Karl Pazdernik. Generalist Multimodal AI: A Review of Architectures, Challenges and Opportunities. arXiv:2406.05496v1 [cs.CL] 8 Jun 2024.*
6. *Balkrishna Rasiklal, Y. (2024). Exploring Multimodal Generative AI: A Comprehensive Review of Image, Text, and Audio Integration. Innovative: International Multi-disciplinary Journal of Applied Technology, 2(10), 124-133.*
7. *Song Bai & Jie Li. Progress and Prospects In 3d Generative Ai: A Technical Overview Including 3d Human. arXiv:2401.02620v1 [cs.AI] 5 Jan 2024.*