applied sciences

MDPI

Review

AI and Generative Models in 360-Degree Video Creation: Building the Future of Virtual Realities

Nicolay Anderson Christian [1], Jason Turuwhenua [2] and Mohammad Norouzifard [1,2,*]

[1] School of Technology, Yoobee College of Creative Innovation, Auckland 1010, New Zealand
[2] Auckland Bioengineering Institute, University of Auckland, Auckland 1010, New Zealand
* Correspondence: mohammad.norouzifard@yoobeecolleges.com or m.norouzifard@auckland.ac.nz

Abstract

The generation of 360° video is gaining prominence in immersive media, virtual reality (VR), gaming projects, and the emerging metaverse. Traditional methods for panoramic content creation often rely on specialized hardware and dense video capture, which limits scalability and accessibility. Recent advances in generative artificial intelligence, particularly diffusion models and neural radiance fields (NeRFs), are examined in this research for their potential to generate immersive panoramic video content from minimal input, such as a sparse set of narrow-field-of-view (NFoV) images. To investigate this, a structured literature review of over 70 recent papers in panoramic image and video generation was conducted. We analyze key contributions from models such as 360DVD, Imagine360, and PanoDiff, focusing on their approaches to motion continuity, spatial realism, and conditional control. Our analysis highlights that achieving seamless motion continuity remains the primary challenge, as most current models struggle with temporal consistency when generating long sequences. Based on these findings, a research direction has been proposed that aims to generate 360° video from as few as 8–10 static NFoV inputs, drawing on techniques from image stitching, scene completion, and view bridging. This review also underscores the potential for creating scalable, data-efficient, and near-real-time panoramic video synthesis, while emphasizing the critical need to address temporal consistency for practical deployment.

Keywords: 360° video generation; generative AI; sparse-input synthesis; panoramic video; diffusion models; neural radiance fields

Academic Editor: Tien Chi Huang

Received: 24 June 2025
Revised: 21 August 2025
Accepted: 22 August 2025
Published: 24 August 2025

Citation: Christian, N.A.; Turuwhenua, J.; Norouzifard, M. AI and Generative Models in 360-Degree Video Creation: Building the Future of Virtual Realities. *Appl. Sci.* **2025**, *15*, 9292. https://doi.org/10.3390/app15179292

1. Introduction

In recent years, the rise of immersive technologies, such as virtual reality (VR), augmented reality (AR), and gaming environments, has driven increasing demand for high-quality 360-degree content. Unlike conventional 2D media, 360° panoramas offer a full spherical field of view, allowing users to look in any direction and experience scenes from multiple perspectives. This capability is central to immersive storytelling, interactive education, simulation training, and virtual tourism. In fact, 360° video is now widely adopted across multiple industries. As shown in Figure 1, its applications include professional sports, travel content, live entertainment, and even news broadcasting and film production [1,2]. Figure 1 indicates that adoption is strongest in visually dynamic domains such as sports and travel, where immersive perspectives enhance audience engagement, while sectors like news and documentaries adopt it more selectively. This distribution underscores the need for generation methods that balance visual quality with adaptability across varied content genres.

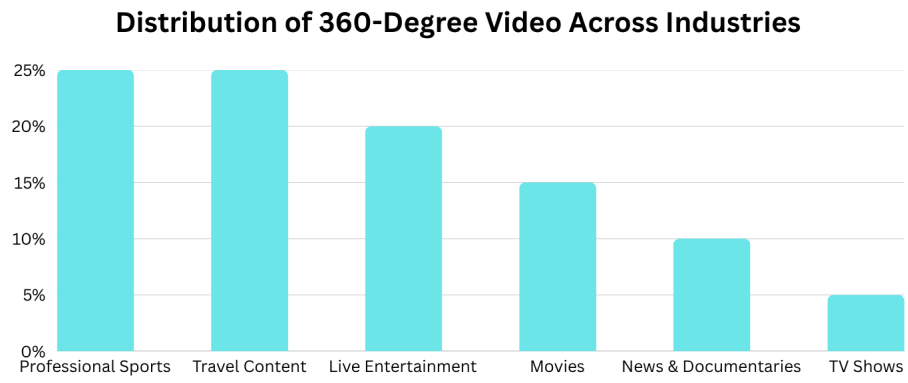

Figure 1. Distribution of 360° video applications across industries. The highest usage is observed in professional sports and travel content, highlighting the medium's strong concentration within the entertainment and film sector. In contrast, the lowest adoption of 360° video is reported in news, documentaries, and TV shows.

Traditionally, 360° content has been captured using specialized multi-camera rigs or panoramic recording setups. While effective, these approaches are often costly, hardware-dependent, and difficult to scale. As a result, there is growing interest in data-driven alternatives that can synthesize panoramic video content using artificial intelligence (AI). Recent advances in neural rendering have enabled promising progress in this area. For example, neural radiance fields (NeRFs) [3] have demonstrated impressive results in reconstructing static 3D scenes from multi-view images (frames). However, standard NeRF frameworks depend on consistent lighting and static geometry, which limits their effectiveness in dynamic or uncontrolled environments. NeRF in the Wild (NeRF-W) is an extension of the classic NeRF model, designed to address the limitations of NeRF when dealing with "in-the-wild" photo collections. The classic NeRF model required a static scene and controlled lighting, making it difficult to use with real-world images from the internet [4]. The NeRF-W model introduces two main innovations to overcome these challenges:

- **Static and Transient Components:** To address the static scene limitation in the classic NeRF model, the NeRF-W model decomposes the scene into static and transient components. The static component represents the permanent parts of the scene (e.g., a building), while the transient component models temporary elements like people or cars moving through the scene. This allows the model to reconstruct a clean, static representation of the scene, even with transient occluders present in the input images [4].
- **Appearance Embeddings:** The model learns a low-dimensional latent space to represent variations in appearance, such as different lighting conditions, weather, or post-processing effects (e.g., filters). By assigning an appearance embedding to each input image, NeRF-W can disentangle these variations from the underlying 3D geometry of the scene [4].

To support another challenge in classic NeRF models, which is the high complexity of outdoor and unbounded environments, the mipmap neural radiance fields 360 (Mip-NeRF 360) model [5] introduces hierarchical scene parameterization and online distillation, improving rendering quality and reducing aliasing in wide-baseline, unbounded scenes. Unlike classic NeRF, which employs uniform ray sampling and fixed positional encoding, Mip-NeRF 360 utilizes conical frustum-based integrated positional encoding and a coarse-to-fine proposal multi-layer perceptron to enhance fidelity. While these improvements significantly boost rendering quality, they come at the cost of increased computational

resources and training time. Nonetheless, the model remains limited to static imagery and does not support dynamic video generation [6].

To address temporal synthesis, Imagine360 proposes a dual-branch architecture that transforms standard perspective videos into 360° panoramic sequences. By incorporating antipodal masking and elevation-aware attention, the model generates plausible motion across panoramic frames [7]. Similarly, models like LAMP (Learn a Motion Pattern) aim to generalize motion dynamics from few-shot inputs, showing that motion priors can be learned and reused efficiently [8]. Recently, Wen [9] introduced a human–AI co-creation framework that enables intuitive generation of 360° panoramic videos via sketch and text prompts, highlighting the growing trend toward collaborative and creatively controllable generation pipelines.

In parallel, diffusion-based models have emerged as a powerful alternative to GANs and NeRFs for generative image and video tasks. SphereDiff mitigates the distortions introduced by equirectangular projection [10] through a spherical latent-space Mip-NeRF and distortion-aware fusion [11]. Meanwhile, Diffusion360 enhances panoramic continuity by applying circular blending techniques during generation, producing seamless and immersive results [12].

Conceptually, sparsity [13] describes a property where a vector or matrix contains a high percentage of zero or near-zero values. However, in the context of generating 360-degree videos from narrow-field images, sparsity primarily refers to a different, yet equally critical, challenge: the limited number and distribution of input images. This sparsity manifests as minimal overlap and insufficient visual information in images across the entire 360-degree field.

There are three sparsity bands used to describe behavior and output quality in 360° video generation, which are described as follows:

- Low sparsity band: The number of NFoV frames is high, angular baselines are small, and view overlap is substantial. Generative models typically produce the highest visual fidelity. Fine textures, consistent color gradients, and stable temporal coherence are preserved due to abundant spatial and angular information. Geometry-aware modules can more accurately reconstruct depth, and parallax artifacts are minimal. However, computational demands are higher, and redundant viewpoints can lead to inefficiencies in training and inference [14–16].
- Medium sparsity band: Moderate angular baselines and reduced overlap begin to challenge models' ability to maintain detail across large scene variations. While global scene structure remains plausible, fine detail and texture consistency may degrade, especially in peripheral regions. Temporal smoothness may also be affected, with occasional flickering in dynamic areas. This regime can serve as a trade-off point, balancing efficiency with acceptable quality for many real-world applications [17,18].
- High sparsity band: Viewpoints are widely spaced and overlap is minimal. Models struggle to interpolate unseen regions accurately. This often results in visible artifacts such as ghosting, stretched textures, and structural distortions, particularly in occluded or complex geometry regions. Temporal coherence suffers significantly in motion-rich scenes, and geometry-aware approaches may produce unstable depth estimates. While computational cost is low, the resulting videos typically require strong priors, advanced inpainting, or multimodal constraints to achieve acceptable quality [16,18,19].

Table 1 defines the three sparsity bands for 360° video generation based on spatial and temporal sampling. Low sparsity (dense) uses many views ($\geq$36 NFoV frames) with small angular baselines ($\leq$10°), high overlap ($\geq$70%), fast capture rates ($\geq$60 views/s), and highly uniform coverage ($\geq$80% entropy). Medium sparsity has moderate views (12–36)

with 10°–30° baselines, 30–70% overlap, 30–60 views/s capture, and moderately uniform coverage (40–80% entropy). High sparsity (sparse) involves few views (≤12) with wide baselines (>30°), low overlap (≤30%), slow capture (≤30 views/s), and uneven coverage (≤40% entropy).

Table 1. Sparsity bands for 360° video generation.

Sparsity Level	Number of NFoV Frames	Angular Baseline Between Views	Coverage (Overlap %)	Per-Second Sampling	View Coverage Entropy
Low	≥36	≤10° between views	≥70% overlap	≥60 views/s	High (≥80% of max H)
Medium	12–36	10°–30° between views	30–70% overlap	30–60 views/s	Moderate (40–80% of max H)
High	≤12	>30° between views	≤30% overlap	≤30 views/s	Low (≤40% of max H)

These developments collectively establish a strong foundation for this research, which explores how generative models can synthesize immersive 360° video from sparse visual inputs, widely spaced perspective images, typically fewer than ten, that provide incomplete coverage of the panoramic viewing sphere. As illustrated in Figure 2, the proposed pipeline situates NFoV inputs within a broader generative framework, highlighting how diffusion models and radiance-field methods can transform limited perspective data into coherent panoramic sequences. This schematic underscores the role of AI-driven synthesis in replacing hardware-intensive capture workflows, enabling both scalability and creative flexibility. By unifying the strengths of radiance-field modeling, motion-aware learning, and diffusion-based generation, this work aims to address the limitations of traditional hardware-dependent pipelines and pave the way toward scalable, cost-effective, and creatively flexible panoramic video generation.

The remainder of this paper is organized as follows. Section 2 outlines the methodology and literature review strategy. Section 3 presents related work in panoramic video generation. Section 4 describes the identified gaps. Section 5 compares computational and processing times. Section 6 discusses the limitations of existing approaches. Section 7 highlights challenges and opportunities for innovation. Section 8 concludes this paper with future research directions.

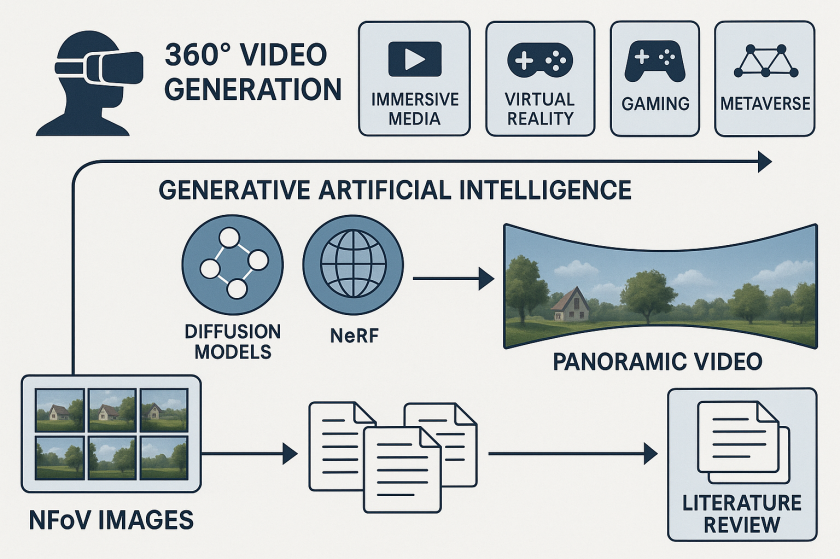

Figure 2. Schematic view of 360° video generation using generative algorithms. Generative models (e.g., diffusion models, NeRF models) enable the transformation of NFoV images into panoramic videos for immersive applications.

2. Methodology

This research is based on a structured literature review of peer-reviewed and preprint publications from 2021 to 2025. The review focused on 360° video generation, panoramic vision, neural radiance fields (NeRFs), diffusion models [20], and generative AI frameworks. Special attention was given to studies addressing sparse-input generation, such as synthesizing panoramic content from a limited number of static images or perspective views.

Relevant papers were sourced through IEEE Xplore, arXiv, and Google Scholar using targeted keyword combinations, including *"360 video diffusion model," "omnidirectional vision," "panoramic video generation," "text-to-360 synthesis," "NeRF 360 video," "image stitching,"* and *"bridging the gap."* These search terms were selected to identify works focusing on spatial consistency, stitching logic, and few-shot synthesis in the context of immersive media.

Studies were selected based on technical relevance, originality, full-text accessibility, and their contribution to the evolving field of AI-based panoramic generation. No experimental data, human subjects, or proprietary datasets were used in this research. The reviewed materials include both open-access and publisher-restricted papers, retrieved through academic databases and institutional access where applicable.

2.1. Research Questions

To guide the scope of this review and structure the analysis of current approaches, the following research questions were formulated:

Research Question 1 (RQ1): What generative approaches are currently used for 360° image and video synthesis?
This question investigates the main categories of generative models, such as NeRFs, GANs, and diffusion models, and how they have evolved to handle immersive panoramic content.

Research Question 2 (RQ2): What unique technical contributions do current models offer for panoramic image and video generation?
This question focuses on how different approaches introduce novel architectures, generation strategies, or domain-specific techniques to advance the field of 360° content synthesis.

Research Question 3 (RQ3): What technical challenges and research gaps remain in current 360° video generation systems?

This question emphasizes the need to address issues such as stitching artifacts, generalization across scenes, and efficient generation from limited viewpoints.

2.2. Literature Selection and Review Strategy

The literature selection process for this review followed a structured approach using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework. A total of 43 records were identified through database searches in IEEE Xplore, arXiv, and Google Scholar, with an additional 13 records identified through other sources such as academic recommendations and manual browsing. After removing duplicates, 56 records remained for screening. During the screening stage, four records were excluded based on a review of the titles and abstracts. The remaining 52 full-text articles were assessed for eligibility, from which 6 were excluded due to insufficient methodological detail or lack of direct relevance to 360° panoramic generation. The final review included 46 papers that contribute to the analysis of AI-driven approaches for 360° video and image synthesis. The PRISMA flow diagram in Figure 3 visualizes this process.

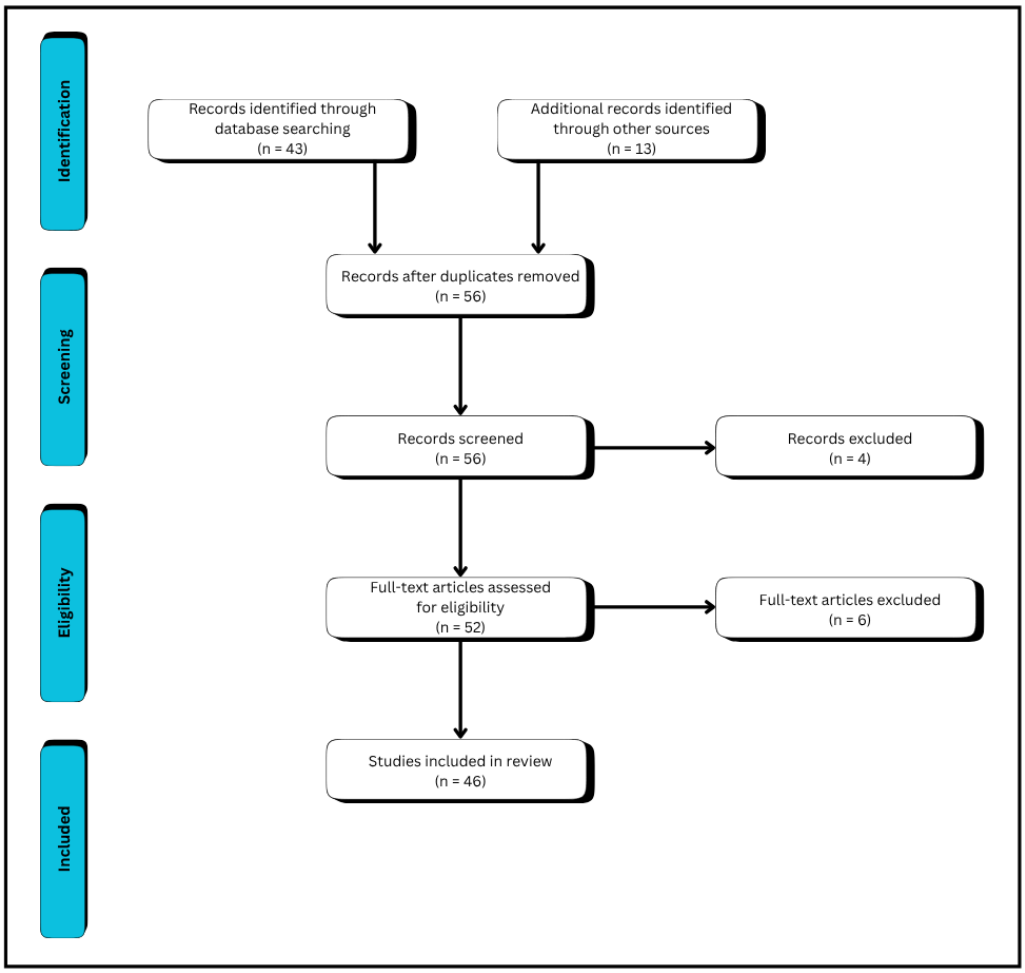

Figure 3. PRISMA chart of selected studies. The chart illustrates the systematic process of identifying, screening, and including studies in the review, ensuring methodological transparency. Although it reflects a structured approach, the selection remains limited by the scope of the databases and sources searched.

The literature search aimed to identify recent and high-impact studies related to 360° image and video generation using artificial intelligence. Searches were conducted across IEEE Xplore, arXiv, Google Scholar, and the proceedings of top-tier conferences such as CVPR. A range of search strings was used, both individually and in combination, to ensure comprehensive coverage of the domain. These are summarized in Table 2.

After the initial keyword search, all retrieved papers were screened using predefined inclusion and exclusion criteria to ensure technical relevance, methodological depth, and alignment with the research objectives. Tables 3 and 4 present these criteria in detail.

Following the selection process, 50 papers were reviewed, from which 12 core works were identified for in-depth analysis due to their significant technical contributions across categories such as diffusion-based generation, sparse-view stitching, neural radiance fields, and latent motion modeling. These core studies were selected based on their methodological depth, relevance to panoramic synthesis under sparse-input conditions, and alignment with the objectives of this thesis. The technical details were extracted from each paper to support the narrative summaries and structured comparisons in the following sections.

Table 2. Search keywords used in the literature search.

No.	Search Keywords
1	360 video diffusion model
2	360° video generation
3	Omnidirectional vision
4	Text-to-360 panorama synthesis
5	Panoramic image generation
6	NeRF 360 video
7	Bridging the gap (panoramic completion)
8	Image stitching for 360 video
9	Deep learning methods
10	Generative models

Table 3. Inclusion criteria.

No.	Criterion
1	Focuses on 360° image or video generation using AI or ML techniques.
2	Produces omnidirectional or panoramic outputs, either in image or volumetric format.
3	Introduces novel methods involving diffusion models, NeRFs, or text-to-360° pipelines.
4	Full-text available with sufficient technical and methodological detail.

Table 4. Exclusion criteria.

No.	Criterion
1	Studies focusing solely on standard 2D video generation without panoramic or 360° objectives.
2	Works addressing 3D reconstruction without spherical or omnidirectional rendering goals.
3	Hardware-specific solutions (e.g., camera rigs) without significant AI-based generation components.
4	Papers that lacked sufficient methodological detail, technical relevance, or a direct focus on 360° panoramic or generative synthesis approaches.

2.3. BERTopic for Topic Modeling

To organize and analyze the research papers, **BERTopic**, a transformer-based topic modeling method, was applied. This technique automatically grouped abstracts into semantically similar clusters, facilitating the identification of recurring themes and research directions within the field.

As shown in Figure 4, BERTopic works [21,22] by first embedding the text using a pre-trained language model. These embeddings are then reduced in dimensionality and clustered to discover patterns in the data. Each cluster is summarized with key terms using a class-based TF-IDF approach, allowing us to label and interpret the core ideas represented in each topic.

After applying this process to the collection of over 40 research papers, we identified **10 distinct topics**. These ranged from 360° video generation and scene reconstruction to image outpainting and spherical vision techniques. The top-ranked terms for each topic

are presented in Figure 5, providing a clear snapshot of the focus areas that emerged from the modeling [22]. This visualization highlights the semantic structure of current research, revealing that while some topics cluster around foundational techniques (e.g., NeRF-based depth modeling, diffusion-driven generation), others focus on application-oriented challenges such as VR/AR integration and panoramic video blending. These thematic clusters help contextualize the breadth of methods and objectives in the panoramic video generation literature.

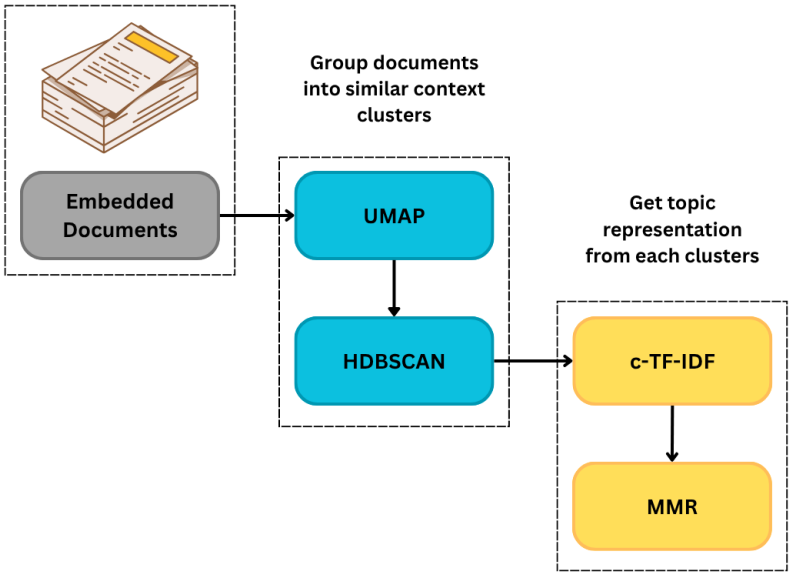

Figure 4. BERTopic modeling process, including embedding, clustering, and topic extraction. The model applies document embeddings and UMAP with HDBSCAN to cluster semantically similar documents, from which topics are extracted and refined using c-TF-IDF and Maximal Marginal Relevance (MMR). This approach offers a robust method for topic discovery but requires careful parameter tuning to yield meaningful results.

Each topic was manually interpreted to go beyond automated keyword summaries. This involved reviewing representative documents from each cluster to understand the context behind the extracted terms. Based on these insights, we assigned human-readable labels to the topics (e.g., "Temporal Scene Synthesis," "Diffusion-Based Rendering," and "Camera Calibration and Stitching Challenges") that more accurately reflected their underlying themes. This step helped ensure that the model outputs were not only statistically coherent but also semantically meaningful and aligned with the literature's actual research focus.

Using BERTopic not only saved time compared to manual categorization but also allowed for a more objective and data-driven structure for the literature review. It helped us understand the landscape of the field more clearly and informed the direction of our research.

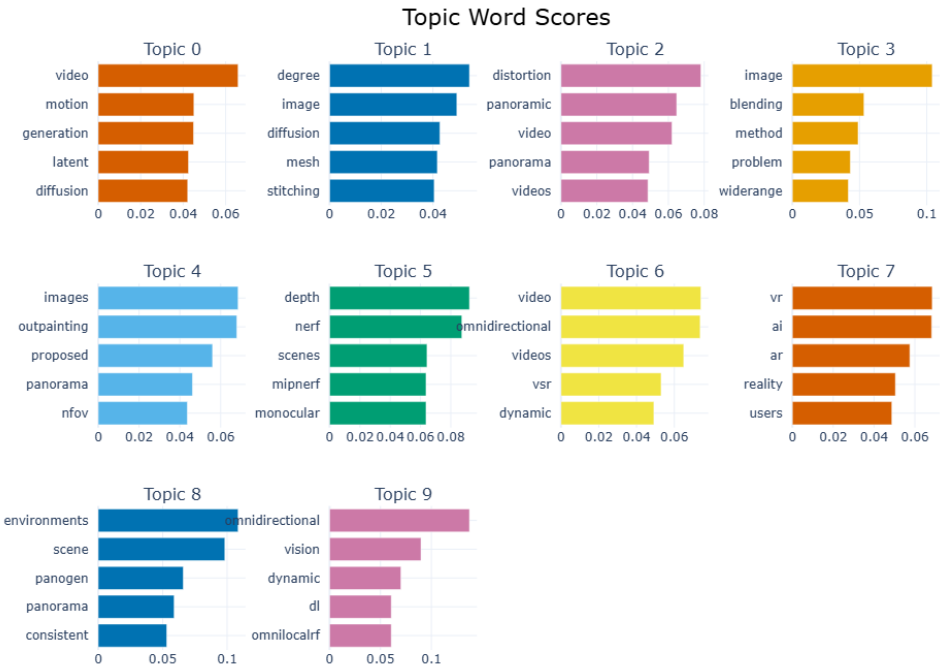

Figure 5. The top 10 topics discovered through BERTopic modeling, visualized with their top five representative keywords and corresponding c-TF-IDF scores. Each horizontal bar chart displays the relative importance of keywords within a topic, enabling a clear and quantitative understanding of the topic's content.

3. Related Works

Before diving into the technical progression of panoramic synthesis, it is essential to recognize the broader context in which this research area has evolved. The generation of 360° video content has gained substantial traction due to its applications in virtual reality (VR), immersive storytelling, remote training, and simulation environments. As the demand for richer, more engaging content has grown, so too has the necessity for automated and intelligent methods capable of synthesizing panoramic scenes with minimal distortion and maximum continuity. Early research in this space primarily focused on hardware-based solutions, such as omnidirectional camera arrays and fisheye lenses, to capture the full 360° field of view. However, these solutions introduced their own challenges, including high cost, calibration complexities, and issues in stitching overlapping views. Consequently, a growing body of work began exploring software-based approaches, laying the groundwork for algorithmic and machine learning models aimed at refining panoramic video generation, particularly in terms of spatial coherence, real-time processing, and perceptual realism.

Table 5 highlights foundational methods and recurring challenges in panoramic synthesis, providing a baseline against which more advanced techniques can be measured. It emphasizes how early reliance on manual stitching and limited computational resources shaped the trajectory of subsequent research. This context contrasts sharply with later developments that integrated deep learning to overcome these constraints. As research advanced, attention shifted toward reconstructing full scenes from sparse inputs, aiming to reduce data capture requirements while improving output quality. Table 6 summarizes the methods designed for view interpolation and scene completion, complementing the earlier table by showing how algorithmic sophistication evolved to address the initial limitations. The direct comparison of sparse-input techniques, as shown in Table 7, reveals performance trade-offs across methods, particularly in balancing fidelity and computational efficiency. This table serves as a bridge between scene reconstruction research

and generative approaches, illustrating which methods laid the strongest foundation for learning-based synthesis. Table 8 summarizes how generative 360° video models perform under different input sparsity regimes (low, medium, and high). Temporal coherence, geometric consistency, texture fidelity, and artifact rate are considered evaluation aspects. Performance trends are indicated qualitatively, showing that low sparsity generally yields smooth motion, accurate geometry, and minimal artifacts, while high sparsity leads to severe breakdowns in depth estimation, motion stability, and texture quality, with artifacts becoming prominent.

The introduction of generative models marked a significant leap in panoramic content creation. Table 9 catalogs the architectures and training paradigms, highlighting how adversarial and diffusion-based approaches unlocked more photorealistic and coherent results. This complements earlier tables by showing the shift from reconstruction-focused to creation-driven methods.

A further innovation emerged with models capable of generating motion in 360° spaces, moving beyond static panoramas. Table 10 outlines these methods, emphasizing how temporal consistency challenges were addressed alongside spatial coherence. This naturally extends the discussion from still-image synthesis to full video generation.

Table 11 shows that the performance of the selected 360° video generation models on the WEB360 benchmark dataset varies notably across methods. VideoPanda and PanoWan exhibit comparatively lower FVD values; VideoPanda also has the highest CLIPScore. In contrast, DynamicScaler exhibits the weakest FVD value, although its CLIPScore is not reported; 360DVD exhibits intermediate performance with both metrics reported.

Table 5. Key contributions of early methods for panoramic image synthesis.

Author	Year	Title	Unique Contribution
Sumantri et al. [23]	2020	360 Panorama Synthesis from a Sparse Set of Images with Unknown Field of View	Demonstrated layout estimation and panoramic reconstruction using GANs from sparse, unregistered images without known camera parameters.
Akimoto et al. [24]	2022	Diverse Plausible 360° Image Outpainting for Efficient 3DCG Background Creation	Enabled visually diverse 360° image outpainting for artistic and background creation use cases, emphasizing multimodal completions.
Liao et al. [25]	2023	Cylin-Painting: Seamless 360 Panoramic Image Outpainting and Beyond	Enhanced continuity in equirectangular projections using circular inference and perceptual optimization strategies.
Zheng et al. [26]	2025	Panorama Generation From NFoV Image Done Right	Introduced a geometry-aware approach for high-resolution panorama generation from narrow-FoV images, improving fidelity and spatial coherence.

Cross-view and text-guided panorama generation introduced new ways to control and customize the output. Table 12 contrasts these approaches with prior work by highlighting the integration of semantic conditioning and multimodal inputs, broadening the scope of possible applications.

Finally, the push toward immersive 3D panoramic environments is captured in Table 13, which synthesizes methods incorporating depth, geometry, and scene realism. This table illustrates how the field is converging toward highly interactive and lifelike

experiences, connecting the technological progression back to the immersive goals outlined at the start of this section.

Table 6. Key contributions of learning-based stitching and reconstruction methods.

Author	Year	Title	Unique Contribution
Nie et al. [27]	2022	Deep Rectangling for Image Stitching: A Learning Baseline	Proposed a supervised baseline to geometrically rectify stitched panoramas into visually natural rectangular formats.
Nie et al. [28]	2023	Parallax-Tolerant Unsupervised Deep Image Stitching	Introduced a parallax-resilient stitching method using semantic and geometric cues, trained without supervision.
Zhou et al. [29]	2024	RecDiffusion: Rectangling for Image Stitching with Diffusion Models	Applied diffusion models to refine stitched image geometry, achieving higher fidelity in panoramic rectangling.
Zheng et al. [26]	2025	Panorama Generation From NFoV Image Done Right	Developed a geometry-aware network for generating high-quality 360° panoramas from a few NFoV images.

3.1. Early Approaches and Classic Challenges in Panoramic Synthesis

The generation of 360° panoramic content from sparse or NFoV inputs has historically posed major challenges in computer vision. Classical methods typically depended on multi-camera rigs and dense, overlapping images, followed by stitching and geometric alignment. These techniques performed well in controlled environments but struggled with unstructured or sparse data, often resulting in visible seams, ghosting, or geometric distortion in the final panorama.

Table 5 outlines the early learning-based models that pioneered the use of neural architectures to generate panoramic scenes from fewer and less structured inputs. One of the earliest such efforts was by Sumantri et al. [23], who proposed a CNN–GAN pipeline for 360° panorama synthesis from a sparse set of unregistered images with unknown FoVs. Their model estimated scene layout and semantic consistency without requiring camera calibration, marking a departure from geometry-dependent approaches.

Building on this, Akimoto et al. [24] introduced a framework for diverse 360° outpainting tailored to 3DCG and creative background generation. Their model emphasized the generation of multiple plausible panoramic completions from a single viewpoint, enabling greater artistic flexibility but at the cost of geometric precision and realism.

Liao et al. [25] shifted the focus toward enhancing image continuity in the equirectangular projection (ERP) space. Their model, Cylin-Painting, introduced circular inference and perceptual loss strategies to mitigate boundary artifacts, achieving smoother transitions across stitched segments and improving perceptual quality in immersive viewing.

A more technically grounded solution was later proposed by Zheng et al. [26], who addressed the shortcomings of prior approaches through a geometry-aware network. Their method demonstrated high-resolution panoramic reconstruction from narrow-FoV images by correcting perspective inconsistencies and maintaining spatial coherence, making it one of the most robust early solutions for sparse-input panoramic generation.

As detailed in Table 5, these foundational works collectively illustrate the transition from traditional geometry-dependent stitching to early learning-based panoramic generation. The table highlights each method's unique contribution, ranging from layout estimation and diverse outpainting to ERP continuity and geometry-aware reconstruction, and

clearly shows how each method addressed a specific aspect of sparse-input panoramic synthesis. Presenting these methods side by side emphasizes their complementary strengths and limitations, providing a historical context that informs the evolution toward modern, fully generative approaches discussed in the next subsection.

3.2. From Sparse Views to Full Scenes: Learning to Stitch and Reconstruct

Following the limitations of classical stitching heuristics and GAN-based outpainting, researchers began to develop learning-based approaches that directly addressed challenges such as parallax misalignment, image rectangling, and incomplete scene geometry.

These methods, which are described in Table 6, aimed to produce visually and geometrically coherent 360° panoramas from sparse or unregistered NFoV [30] inputs.

One early step in this direction was Deep Rectangling by Nie et al. [27], who introduced a supervised learning baseline to correct the warped geometry often encountered in stitched panoramas. By transforming irregular panoramas into natural rectangular layouts, their model laid a foundation for more structured image composition in panoramic synthesis.

Extending this effort, Nie et al. [28] proposed an unsupervised stitching framework that tolerates large parallax shifts. The model leverages semantic cues and geometric reasoning to align disjointed images without explicit correspondence maps or calibration data, offering robustness in unstructured or dynamic scenes.

Building upon these developments, Zhou et al. [29] introduced RecDiffusion, a novel rectangling-aware diffusion model. Their approach refines stitched panoramas through generative sampling, preserving detailed textures and structural boundaries. By integrating diffusion processes into stitching, RecDiffusion achieves improved visual fidelity compared to earlier rectification techniques.

More recently, Zheng et al. [26] reframed the task as a geometry-aware reconstruction problem from sparse NFoV images. Their architecture accounts for perspective distortions and spatial continuity, enabling high-resolution panoramas even from very limited input. The model excels in generating seamless outputs without requiring dense image coverage.

As summarized in Table 6, these learning-based methods collectively demonstrate the progression from heuristic stitching to intelligent, geometry-aware reconstruction pipelines. The table highlights how each contribution addresses a distinct challenge, ranging from geometric rectangling and parallax tolerance to generative refinement and sparse-input reconstruction, illustrating the incremental steps needed to overcome the limitations of classical stitching. Presenting these methods side by side clarifies their complementary roles in advancing panoramic composition and sets the stage for the transition to fully generative synthesis approaches, which are explored in the next subsection on GAN- and diffusion-based models.

3.3. Sparse-Input Comparison

Table 7 provides a comparative overview of recent sparse-input 360° reconstruction and novel-view synthesis models. It summarizes key aspects, including input modalities and sparsity levels, supervision and prior strategies, temporal handling, geometry utilization, reported quantitative performance, and common failure modes. The table highlights how different approaches balance sparse-view input constraints with 3D reconstruction fidelity, temporal consistency, and rendering quality, illustrating the trade-offs between feed-forward, diffusion-based, and hybrid methods across both object-centric and large-scale unbounded scenes.

All compared methods address high-sparsity conditions (i.e., few input views with large angular baselines), making 360° generation particularly challenging. While some approaches incorporate temporal handling for video (e.g., MVSplat360 with strong SVD coherence), others focus solely on static reconstructions.

Table 7. Comparison of sparse-input models.

Model	Input Modality and Sparsity Band	Priors/Supervision	Temporal Handling	Geometry Use	Reported Metrics (Dataset, Res.)	Noted Failure Modes
MVSplat360 [31]	High Sparsity: Sparse posed images (as few as 5 views), wide-baseline observations	Pre-trained Stable Video Diffusion (SVD); leverages large-scale latent diffusion prior (LDM)	Temporally consistent video generation; strong SVD temporal coherence	Geometry-aware 3D reconstruction with coarse 3DGS; backprop from SVD improves geometry backbone	Superior visual quality on DL3DV-10K and RealEstate10K; photorealistic 360° NVS	Coarse 3DGS may show artifacts/holes; prior feed-forward methods fail under sparse observations
Free360 [32]	High Sparsity: *Unposed* views (e.g., 3–4) in unbounded 360° scenes	Dense stereo for coarse geometry; iterative fusion of reconstruction and generation; uncertainty-aware training	Not primarily video; targets 3D reconstruction from sparse views	Layered Gaussian representation; dense stereo for coarse geometry/poses; layer-specific bootstrap optimization to denoise and fill occlusions	Outperforms SoTA in rendering quality and surface reconstruction accuracy from 3–4 views	Dense stereo noisy in unbounded scenes → artifacts in novel views; unreliable depth; visibility issues
ReconX [33]	High sparsity: Sparse-view images (as few as two)	Large pre-trained *video* diffusion prior; 3D structure condition from global point cloud; confidence-aware 3DGS optimization	Reframes ambiguous reconstruction as temporal generation; detail-preserved frames with high 3D consistency	Builds global point cloud → 3D structure condition; recovers 3D from generated video using confidence-aware 3DGS	SoTA on common datasets (especially view interpolation); promising 360° extrapolation; robust on MipNeRF360 and Tanks&Temples; surpasses feed-forward baselines	Hard to maintain strict 3D view-consistency from pre-trained video models; limited full-360 evals; slower inference vs. feed-forward
UpFusion [34]	High Sparsity: Sparse reference images without poses (1, 3, or 6 unposed object views)	Pre-trained 2D diffusion (e.g., Stable Diffusion) initialization; scene-level transformer with query-view alignment; direct attention to input tokens; score distillation for 3D consistency	Novel-view synthesis and 3D inference for *objects* (not video-temporal)	Infers 3D object representations; extracts 3D-consistent models via score-based distillation	Evaluated on Co3Dv2 and Google Scanned Objects; strong gains vs. pose-dependent baselines; outperforms single-view 3D methods	Regression-style training can be blurry (mean-seeking under uncertainty)
NeO-360 [35]	High Sparsity: Single or a few posed RGB images in unbounded 360° outdoor scenes	Image-conditional tri-planar representation; trained on many unbounded 3D scenes	Not a video-temporal method; focuses on novel-view synthesis from single/few images	Tri-planar representation models 3D surroundings; decomposition via samples inside 3D bounding boxes	Outperforms generalizable baselines on NeRDS360 (75 unbounded scenes)	Struggles to predict global scene representation in complex street scenes with forward motion
Splatter-360 [36]	High Sparsity: Wide-baseline panoramic images	Spherical cost volume via spherical sweep; 3D-aware bi-projection encoder (monocular depth, ERP, cube-map); cross-view attention; randomized camera baselines	Real-time rendering; not a video-temporal method	Improves depth/geometry via spherical cost volume and bi-projection encoder	Outperforms SoTA NeRF/3DGS (PanoGRF, MVSplat, DepthSplat, HiSplat) on HM3D/Replica; PSNR, SSIM, LPIPS; depth metrics (Abs Diff/Rel, RMSE, $\delta < 1.25$); e.g., +2.662 dB PSNR vs. PanoGRF on HM3D	Multi-view matching struggles with occlusions and textureless areas

Most leverage geometry-aware strategies, ranging from coarse 3D Gaussian splats to triplanar or spherical volume representations.

3.4. Effects Across Sparsity Bands

Table 8 summarizes how the behavior of the models discussed in Table 7 varies across the three sparsity bands defined in Table 1. A comparison across four evaluation aspects, including temporal coherence, geometric consistency, texture fidelity, and artifact rate, is presented in a compact format.

We use qualitative indicators (↑ increase, → neutral/stable, ↓ decrease) to compactly convey trends. Quantitative call-outs from the literature are provided where available.

Table 8. Model behavior across sparsity bands.

Evaluation Aspect	Low Sparsity	Medium Sparsity	High Sparsity
Temporal Coherence	↑ Smooth sequences; stable motion continuity	→ Mostly stable, minor flicker in dynamic regions	↓ Severe breakdown; ghosting and jitter in motion-rich scenes
Geometric Consistency	↑ Accurate depth/parallax recovery	→ Coarse but plausible structure; errors at periphery	↓ Structural distortions, unstable depth estimates
Texture Fidelity	↑ Fine-grained detail, sharp textures	→ Moderate texture degradation, especially peripheral	↓ Blurring, stretched textures, loss of fine detail
Artifact Rate	↓ Minimal artifacts, efficient reconstruction	→ Noticeable artifacts in complex occlusions	↑ Frequent artifacts (ghosting, seams, inpainting errors)

Across models, low-sparsity regimes consistently yield the highest visual fidelity and temporal stability. Medium sparsity introduces trade-offs between efficiency and detail. In high-sparsity conditions, all methods degrade sharply, showing ghosting, stretched textures, or unstable depth. These patterns underscore the critical role of input density in determining generative quality and highlight the need for advanced priors or hybrid constraints to address the high-sparsity regime.

Notably, transformer models tend to preserve temporal coherence, even under moderate sparsity, while diffusion models show comparatively stronger texture recovery. Overall, the synthesis of trends reveals that no single model dominates across all aspects, pointing toward the potential of hybrid or ensemble methods to balance competing objectives.

3.5. Generative Models for Panoramic Image and Video Synthesis

As panoramic generation matured beyond stitching and completion, a new wave of research emerged, leveraging deep generative models to synthesize 360° content directly. Generative adversarial networks (GANs) initially laid the groundwork, but diffusion models rapidly took center stage due to their improved mode coverage, output stability, and scalability. These approaches emphasize semantic richness, geometric continuity, and immersive detail, transforming how panoramic content is created.

Table 9 describes one of the early contributions in this space, which came from Dastjerdi et al. [37], who introduced a guided co-modulated GAN architecture for full-field-of-view extrapolation. Their model supported user-guided panoramic outpainting and scene completion by combining domain-specific conditioning with co-modulated style generation. While it produced convincing wide-angle imagery, it relied heavily on fine-tuned inputs and exhibited challenges in preserving long-range structural coherence.

Table 9. Key contributions of generative models for panoramic synthesis.

Author	Year	Title	Unique Contribution
Dastjerdi et al. [37]	2022	Guided Co-Modulated GAN for 360° Field of View Extrapolation	Introduced user-guided extrapolation using co-modulated GANs to enhance realism in panoramic outpainting tasks.
Feng et al. [12]	2023	Diffusion360: Seamless 360° Panoramic Image Generation Based on Diffusion Models	Proposed circular blending during denoising to eliminate left–right discontinuities in panoramic images.
Zhang et al. [38]	2024	Taming Stable Diffusion for Text-to-360° Panorama Image Generation (PanFusion)	Adapted Stable Diffusion with ERP-aware attention for unpaired, training-free 360° text-to-image generation.
Wang et al. [39]	2024	Customizing 360° Panoramas Through Text-to-Image Diffusion Models	Enabled text-driven panoramic scene customization through latent diffusion with prompt-based control.
Park et al. [11]	2025	SphereDiff: Tuning-Free Omnidirectional Panoramic Image and Video Generation	Developed a spherical latent representation for high-fidelity panoramic generation without task-specific tuning.

A notable shift occurred with the advent of diffusion models. Feng et al. [12] presented Diffusion360, which applied denoising diffusion probabilistic models (DDPMs) to generate seamless panoramic images. By implementing a circular blending mechanism during denoising, the model successfully eliminated left–right seams in equirectangular projections, an issue often encountered in panoramic formats.

Following this, Zhang et al. [38] proposed PanFusion, an adaptation of Stable Diffusion tailored for text-to-360° image generation. The model introduced ERP-aware attention mechanisms, specifically Equirectangular-Perspective Projection Attention (EPPA), to mitigate projection distortion and enhance layout realism. Crucially, PanFusion supports unpaired training, making it accessible for broader applications without annotated datasets.

To advance customization, Wang et al. [39] developed a diffusion-based pipeline for text-controlled panoramic generation. Their approach emphasizes layout manipulation and stylistic flexibility through prompt engineering, allowing users to tailor scene composition interactively. This highlights an emerging focus on controllability in generative workflows.

Pushing diffusion synthesis into the spherical domain, Park et al. [11] introduced SphereDiff. This model utilizes a spherical latent space and distortion-aware fusion strategies to produce omnidirectional imagery with high fidelity. Notably, it operates in a tuning-free manner, demonstrating strong generalization across datasets and use cases without task-specific retraining.

As summarized in Table 9, these generative approaches collectively illustrate the transition from completion-based pipelines to fully AI-driven synthesis of panoramic content. The table highlights each model's unique contribution, from co-modulated GAN conditioning and user-guided outpainting to ERP-aware diffusion and spherical latent modeling, showing how successive innovations have addressed key challenges such as left–right seam removal, projection distortion, and controllable content generation. Presenting these models side by side clarifies how generative methods not only achieve higher visual fidelity and semantic richness than traditional stitching but also establish a foundation for scalable panoramic video synthesis, which is explored in the following subsection.

3.6. From Still to Motion: Diffusion and Latent Models for 360° Video

The success of diffusion-based methods in static panoramic image synthesis has catalyzed a new wave of research focused on video generation. These efforts aim to extend spatial coherence to the temporal domain, enabling immersive 360° video synthesis from

minimal inputs, such as a single image, a short prompt, or a sparse sequence of frames. Key challenges include maintaining motion continuity, controlling generation dynamics, and preserving the structural integrity of panoramic views over time.

Table 10 outlines early foundational work, including that by Zhou et al. [40], who introduced MagicVideo, a latent diffusion model designed for efficient and high-quality video synthesis. By leveraging compact representations and temporal modeling, the framework enables realistic motion dynamics while significantly reducing computational overhead.

Table 10. Key contributions of generative models for panoramic and image-to-video synthesis.

Author	Year	Title	Unique Contribution
Zhou et al. [40]	2022	MagicVideo: Efficient Video Generation with Latent Diffusion Models	Achieved high-quality video synthesis using latent diffusion and efficient training for natural video dynamics.
An et al. [41]	2023	Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation	Added a parameter-free temporal-shift module to reuse image diffusion models for temporally coherent video generation.
Chen et al. [42]	2023	Motion-Conditioned Diffusion Model for Controllable Video Synthesis (MCDiff)	Enabled user-guided video synthesis using sparse motion strokes and flow-based motion conditioning.
Zhang et al. [43]	2023	I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models	Cascaded two-stage diffusion model to progressively refine visual fidelity and motion quality in video generation.
Ni et al. [44]	2023	Conditional Image-to-Video Generation with Latent Flow Diffusion Models	Combined latent flow fields with diffusion-based generation for optical flow-aware video synthesis.
Wang et al. [45]	2024	360DVD: Controllable Panorama Video Generation with 360° Video Diffusion Model	Introduced a 360-Adapter for AnimateDiff and released the WEB360 dataset for 360° panoramic video-text synthesis and benchmarking.
Tan et al. [7]	2024	Imagine360: Immersive 360 Video Generation from Perspective Anchor	Proposed a dual-branch framework with antipodal masking to lift perspective videos into full panoramic space.
Hu et al. [46]	2025	LaMD: Latent Motion Diffusion for Image-Conditional Video Generation	Decomposed motion and content into latent representations, enabling motion-conditioned video synthesis from static images.

Building on this, An et al. [41] proposed Latent-Shift, which includes a parameter-free temporal-shift module to adapt existing image diffusion models for coherent video generation. This modification allows the reuse of pre-trained networks while introducing temporal awareness without retraining from scratch.

Chen et al. [42] contributed MCDiff, a motion-conditioned diffusion model that allows users to guide video synthesis using sparse motion strokes. By incorporating flow-based conditioning, the model offers precise motion control, enhancing flexibility in generation tasks.

In the realm of progressive refinement, Zhang et al. [43] presented I2VGen-XL, a cascaded two-stage diffusion framework for image-to-video synthesis. The model first establishes coarse temporal dynamics and then enhances fidelity through a secondary refinement stage, producing smooth and high-resolution outputs.

Ni et al. [44] tackled optical flow integration by introducing a latent flow-guided diffusion model. Their system combines flow estimation with diffusion to better control inter-frame consistency and directional motion, advancing the quality of conditional image-to-video generation.

A significant leap toward 360° applications came from Wang et al. [45], who introduced 360DVD, a plug-and-play extension of AnimateDiff adapted for panoramic video

generation. The framework includes a 360-Adapter module and a latitude-aware design to preserve spatial realism across spherical views. Additionally, the authors released WEB360, a benchmark dataset tailored to panoramic video-text synthesis.

Tan et al. [7] further advanced 360° video generation with Imagine360, a dual-branch architecture that transforms perspective videos into immersive panoramic sequences. By combining antipodal masking with elevation-aware attention, the model effectively bridges conventional and panoramic domains even under limited input conditions.

Finally, Hu et al. [46] proposed LaMD, a latent motion diffusion model that decomposes motion and content into separate latent variables. This two-stage pipeline supports image-to-video generation with high control over motion representation, pushing the boundaries of one-shot synthesis.

As summarized in Table 10, these models collectively map the evolution from static panoramic image generation to dynamic, temporally coherent 360° video synthesis. The table highlights each model's unique contribution, ranging from latent diffusion and temporal-shift strategies to motion-conditioned pipelines and spherical adaptations, clarifying how incremental innovations have addressed key challenges such as motion continuity, panoramic distortion, and controllable video generation. Presenting these methods side by side illustrates not only the diversity of approaches but also their complementary roles in advancing toward fully immersive, generative video systems, providing a clear bridge to the cross-modal and multi-input frameworks discussed in the next subsection.

3.7. Performance of 360° Video Generation Models on the WEB360 Benchmark Dataset

Recent works in text-to-360-degree video generation have been evaluated using the WEB360 benchmark dataset. Table 11 summarizes the text-to-360° video generation performance of the main models that report comparable metrics on WEB360 (a public dataset of ∼2k captioned panoramic video clips [45]; see https://github.com/Akaneqwq/360DVD/blob/main/download_bashscripts/4-WEB360.sh, accessed on 1 August 2025). The Fréchet Video Distance (FVD) [47,48] metric (lower is better) is reported for the overall video quality, and a CLIP-based score [49] (higher is better) for text–video alignment. As shown in Table 11, 360DVD [45] established the first baseline on WEB360, but newer diffusion-based models have achieved lower FVD values and higher CLIP alignment. In particular, VideoPanda [50] and PanoWan [51] improved the FVD by ∼30% over 360DVD. VideoPanda's multi-view diffusion yielded an FVD ≈ 1258 and a CLIPScore ≈ 29.8, versus 360DVD's 1750 and 28.4 on the same prompts. The PanoWan model achieved a similar FVD ≈ 1281 while addressing panoramic distortions.

The DynamicScaler model [52], a tuning-free method that stitches local generations, lagged behind with the highest FVD (>2100), indicating poorer visual quality. When comparing these results, it is important to note the inconsistencies in the metrics and protocols across studies. For example, 360DVD introduced bespoke metrics (i.e., a CLIP-based score for evaluating text-to-360° alignment), whereas later works (VideoPanda and PanoWan models) used standard CLIP similarity scores or VideoCLIP models. This means that the alignment scores are not directly comparable between papers [50–52]. Likewise, some report FVD values on *equirectangular frames* versus on projected normal-view frames, which can halve the numeric value. We have chosen a single setting per model for consistency, but such differences in metric computation limit direct comparisons. Another difference lies in the evaluation data. VideoPanda and 360DVD were evaluated on held-out WEB360 prompts, while PanoWan introduced a new, larger dataset (PanoVid) and reported metrics on that benchmark. This makes absolute scores not strictly comparable to those on WEB360; relative improvements are more meaningful [45,50,51].

Table 11. Performance of selected 360° video generation models on the WEB360 benchmark dataset. *Note*: "–" indicates that the metric was not reported.

Model	FVD Metric	CLIPScore
360DVD	∼1750	∼28.4
DynamicScaler	>2146	–
VideoPanda	∼1258	∼29.8
PanoWan	∼1281	–

As shown in Table 11, the traditional frame-reconstruction metrics, PSNR and SSIM, are seldom reported for text-to-video generation on WEB360, since there is no ground truth for open-ended prompts. They appear mainly in *conditional generation* tasks, such as image-to-360° video outpainting in VideoPanda, where a PSNR ≈ 17.6 dB and an SSIM ≈ 0.636 outperformed a panorama outpainting baseline (13.4 dB, 0.485). In summary, few models follow the same evaluation protocol on WEB360; only two or three can be directly compared. We outline these limitations in Table 11 and emphasize that current WEB360 results are in an early stage but consistently show the progress of newer diffusion-based approaches over the initial 360DVD baseline.

Figure 6 provides an illustrative summary of the performance of four distinct models, including 360DVD, DynamicScaler, VideoPanda, and PanoWan, based on the FVD metric. Figure 6 is derived from the values reported in Table 11, which provides the underlying numerical results for both the FVD and CLIPScore. Together, Table 11 and Figure 6 highlight not only the relative strengths of the models but also the gaps in the reported metrics, such as the absence of CLIPScore values for DynamicScaler and PanoWan.

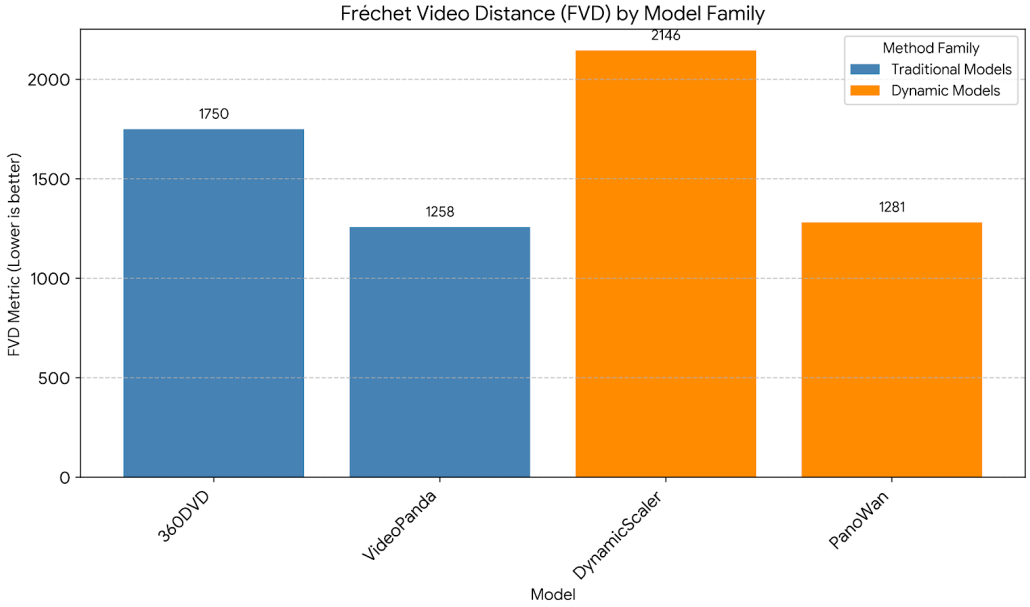

Figure 6. Grouped bar chart illustrating the FVD metric for four video generation models. These models are categorized into two families: dynamic models (DynamicScaler, PanoWan) and diffusion-based models (360DVD, VideoPanda). *Note*: The values are based on the selected subset of four models from the WEB360 dataset and are illustrative. The evaluation protocols may differ across models, and missing values are not imputed.

3.8. Cross-View and Text-Guided Panorama Generation

As generative models have evolved, a compelling trajectory has emerged: panoramic content generation from abstract, non-visual modalities such as aerial imagery, language prompts, and contextual navigation cues. These approaches push the boundaries of traditional input–output mapping by learning to synthesize immersive 360° scenes from highly compressed or semantically distant representations. Bridging these modality gaps requires not only spatial coherence but also semantic consistency, especially in applications like navigation, simulation, or storytelling.

Table 12 shows that one of the earliest efforts in this area was by Wu et al. [53], who introduced PanoGAN, a GAN-based model that translates top-down aerial imagery into ground-view panoramic scenes. By leveraging adversarial feedback and dual-branch discriminators, the model refines scene realism and structural alignment, showcasing the potential of cross-view synthesis.

Soon after, Li et al. [54] expanded the scope of conditioning inputs by integrating textual instructions with spatial reasoning in Panogen. Designed for vision-and-language navigation (VLN), the system generates panoramic scenes guided by both task context and descriptive prompts. This multimodal grounding enables panoramic generation aligned with movement goals and spatial semantics.

Table 12. Key contributions of cross-view and text-guided panorama generation models.

Author	Year	Title	Unique Contribution
Wu et al. [53]	2022	Cross-View Panorama Image Synthesis	Developed PanoGAN, a GAN with adversarial feedback to generate ground-view panoramas from top-down aerial imagery.
Li et al. [54]	2023	Panogen: Text-Conditioned Panoramic Environment Generation for Vision-and-Language Navigation	Combined text prompts with VLN context to generate spatially aligned panoramic scenes for navigation tasks.
Yang et al. [55]	2024	CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer	Introduced an expert transformer for generating high-quality panoramic videos from textual prompts.
Choi et al. [56]	2024	OmniLocalRF: Omnidirectional Local Radiance Fields from Dynamic Videos	Reconstructed static panoramic views by filtering out dynamic content from 360° videos using local radiance fields.
Zhang et al. [57]	2025	PanoDiT: Panoramic Video Generation with Diffusion Transformer	Proposed a panoramic video diffusion model with multi-view attention for text-guided scene generation.
Xiong et al. [58]	2025	PanoDreamer: Consistent Text to 360° Scene Generation	Introduced trajectory-guided semantic warping for consistent text-to-360 panorama synthesis aligned with camera motion paths.

Moving toward video and temporal modeling, Yang et al. [55] proposed CogVideoX, a text-to-video diffusion model equipped with an expert transformer architecture. While applicable to general video generation, the system's compositional understanding and prompt conditioning extend well to panoramic scenarios, particularly in guided or interactive applications.

In a related but orthogonal direction, Choi et al. [56] developed OmniLocalRF, a model that reconstructs static panoramic views from dynamic 360° video. Although not language-guided, the method leverages bidirectional optimization and local radiance fields to filter out transient content, supporting clarity and structure in panorama reconstruction from noisy or dynamic inputs.

More recently, Zhang et al. [57] introduced PanoDiT, a diffusion transformer tailored for panoramic video synthesis from textual prompts. By using multi-view attention, the model maintains coherence across spherical views and time, producing structurally aligned and visually smooth sequences.

Finally, Xiong et al. [58] proposed PanoDreamer, which takes text-to-360 generation a step further through trajectory-guided semantic warping. This framework aligns output scenes with virtual camera paths and natural language prompts, achieving high consistency and realism even in complex layouts with occlusion-aware expansion.

As summarized in Table 12, these works collectively chart the evolution from conventional image-conditioned panorama generation to models that can synthesize 360° content from abstract, multimodal, and often highly compressed inputs. The table highlights how each contribution tackles a different challenge, ranging from cross-view translation and text-to-scene synthesis to trajectory-aware video generation, illustrating how semantic alignment and structural coherence are progressively addressed across modalities. Presenting these methods side by side clarifies how cross-view and language-driven approaches expand the creative and practical potential of panoramic generation, establishing a clear bridge to the multimodal and 3D spatial environments explored in the next subsection.

3.9. Toward Immersive and Hyper-Realistic 3D Panoramic Environments

The evolution of panoramic generation is steadily moving toward immersive and explorable 3D environments that go beyond static visuals. This shift is driven by increasing demands for depth-aware rendering, spatial interactivity, and real-time deployment in virtual reality and augmented reality (VR and AR) settings. A growing body of work has begun addressing these challenges, introducing innovations that bridge visual quality, semantic context, and deployment efficiency.

Table 13 shows that a foundational step in this direction was introduced by Huo and Kuang [59], who developed TS360, a reinforcement learning-based framework designed to optimize the streaming of 360° video content. Although not a generative model, TS360 highlights the importance of bandwidth-aware delivery systems that support real-time applications in immersive environments, an essential complement to content generation efforts.

To improve visual fidelity, especially in scenarios constrained by projection artifacts, Yoon et al. [60] proposed SphereSR, a projection-independent super-resolution framework. By using a continuous spherical representation, the model enhances panoramic image quality across arbitrary projections, offering a more faithful and seamless viewing experience.

Building on this, Baniya et al. [61] presented a deep learning system tailored to enhance both spatial and temporal resolution in 360° video. Their work directly addresses the high visual standards required for immersive media, particularly in VR applications, by ensuring fluid playback and detailed visuals in dynamic panoramic scenes.

As research advanced from resolution enhancement to immersive scene creation, Yang et al. [62] introduced LayerPano3D, a model that constructs 3D panoramic environments using layered representations generated from text prompts. This approach enables multi-view navigation and real-time virtual exploration while maintaining structural realism, thus bridging generative models and 3D spatial computing.

Further extending the capabilities of panoramic environments, Kou et al. [63] developed OmniPlane, a recolorable scene representation that disentangles geometry and appearance in dynamic panoramic videos. This facilitates real-time editing, relighting, and context adaptation, making the system especially valuable for interactive or adaptive media applications.

In a complementary direction, Behravan et al. [64] explored the use of vision-and-language models for context-aware 3D object creation in augmented reality. While their work does not focus solely on panoramic generation, it introduces techniques for semantically grounded object synthesis, which could be extended to enrich panoramic environments with intelligent, interactive elements.

Table 13. Key contributions of immersive and 3D-oriented panoramic generation models.

Author	Year	Title	Unique Contribution
Huo et al. [59]	2022	TS360: A Two-Stage Deep Reinforcement Learning System for 360° Video Streaming	Applied reinforcement learning to optimize bandwidth-efficient, high-quality streaming of 360° video content.
Yoon et al. [60]	2022	SphereSR: 360° Image Super-Resolution with Arbitrary Projection via Continuous Spherical Representation	Proposed a projection-independent spherical SR method for enhancing 360° image quality without introducing spatial artifacts.
Baniya et al. [61]	2023	Omnidirectional Video Super-Resolution Using Deep Learning	Designed a deep learning system to enhance both spatial and temporal resolutions in 360° video for immersive applications.
Yang et al. [62]	2024	LayerPano3D: Layered 3D Panorama for Hyper-Immersive Scene Generation	Generated 3D panoramic environments from text prompts using layered representations, enabling real-time virtual exploration.
Kou et al. [63]	2025	OmniPlane: A Recolorable Representation for Dynamic Scenes in Omnidirectional Videos	Developed a dynamic panoramic scene representation separating geometry and appearance to support relighting and editing.
Behravan et al. [64]	2025	Generative AI for Context-Aware 3D Object Creation Using Vision–Language Models in AR	Demonstrated vision–language-guided generation of semantically adaptive 3D objects for AR, suggesting potential for integrating interactive object synthesis into panoramic environments.

As summarized in Table 13, these models collectively illustrate the field's progression from enhanced 360° video delivery and super-resolution toward fully generative, interactive 3D panoramic environments. The table highlights how each work contributes a distinct capability, ranging from bandwidth-efficient streaming and projection-independent super-resolution to layered 3D scene construction, dynamic relighting, and context-aware object synthesis, clarifying how these innovations address the dual challenge of visual fidelity and spatial interactivity. Presenting these methods side by side underscores the convergence of panoramic generation and spatial computing, showing how technical advances in delivery, resolution, and semantic integration are shaping the path to truly immersive and explorable 360° environments.

4. Identified Gaps

Although significant advances have been made in 360° panoramic image and video generation, the existing body of literature reveals several persistent gaps that limit the scalability, generalizability, and deployability of current approaches.

A major limitation lies in the reliance on dense, overlapping, or structured input data. Many models, whether based on stitching [23,27], GANs [37], or diffusion [11,12], assume input completeness or paired view correspondences to function effectively. Even state-of-the-art solutions such as Imagine360 [7] and PanoDreamer [58] require perspective video streams or semantic layout priors, making them less applicable in scenarios involving sparse or non-overlapping NFoV images. While methods like SphereDiff and PanFusion mitigate projection issues, they remain tuned for image-level outputs and often lack generalization under low-data regimes. This limitation reduces accessibility in real-world cases

where full scene coverage is unavailable, such as mobile capture, quick scans, or low-cost drone footage.

Temporal modeling in generative video synthesis is also constrained by input assumptions and lacks flexibility under few-shot conditions. Works like 360DVD [45], LaMD [46], and MagicVideo [40] demonstrate strong results in controlled settings but still rely on continuous motion prompts or full video sequences. Models like Latent-Shift [41] and MCDiff [42] propose efficient mechanisms for motion conditioning, yet few studies have examined how they can be adapted for sparse temporal inputs or motion hallucination from static frames. This gap limits their practical use in applications where continuous video is not available or motion must be inferred from minimal data, such as surveillance, fast prototyping, or archival conversion.

Another critical gap is the underutilization of geometric information such as scene depth, symmetry, and layout. While methods like 360MonoDepth [65] and Hara et al. [66] highlight the role of depth and symmetry in improving spatial realism, these concepts are not yet embedded in most panoramic generation pipelines. Diffusion-based models like SphereDiff [11] operate in latent space without incorporating explicit geometry, and many stitching or blending approaches [28,67] still treat geometry as a post-processing correction rather than an integral modeling component. As a result, generated content often lacks spatial coherence or realism when deployed in 3D environments like VR, AR, or game engines.

Controllability and cross-modal conditioning remain promising but underdeveloped directions. Text-to-360-degree generation models such as Panogen [54], PanoDiT [57], and Customizing360 [39] have made early strides in prompt-based generation but still rely heavily on training-time conditioning or paired datasets. Generalization to unseen prompts, free-form editing, or hybrid inputs remains limited, particularly in the context of dynamic video generation or immersive interaction. Recent advancements in AR content creation [64] demonstrate how vision–language models can generate scene-aware 3D assets based on environmental context and textual intent, highlighting a path forward for panoramic systems to become more adaptive and context-sensitive. This restricts personalization and user interaction in consumer-facing applications like virtual tourism or narrative-based VR experiences.

Lastly, real-time deployment, streaming readiness, and lightweight inference are seldom addressed. While some works explore architectural efficiency [40,42] or adaptive streaming strategies [59], these aspects are rarely integrated with generative quality. This gap limits the real-world applicability of current systems in domains such as VR/AR, mobile media, or simulation training, where performance, latency, and interactivity are as critical as fidelity. This affects scalability in environments with bandwidth, latency, or device constraints, particularly in education, live events, or mobile deployment scenarios.

In summary, while the field has progressed rapidly across several fronts, it continues to face challenges in sparse-input generalization, temporal realism, geometry integration, controllability, and deployment efficiency. These limitations directly motivate this thesis, which proposes a unified framework that synthesizes 360° panoramic video from sparse NFoV inputs while maintaining temporal coherence and visual fidelity. By integrating geometric priors, motion-aware learning, and efficient generative pipelines, this work aims to bridge key gaps in both technical capability and practical usability.

5. Discussion

The findings from this review illustrate a rapidly evolving field of research in panoramic image and video generation, driven by the convergence of generative AI, radiance-field modeling, and immersive media demands. Across the spectrum, from clas-

sical image stitching and outpainting to diffusion-based 360° synthesis and multimodal guidance, researchers have made significant strides in generating semantically rich, geometrically coherent panoramic content. However, several critical limitations and emerging trends shape the trajectory of future development.

Across this evolution, three primary families of generative approaches (GAN-based methods, radiance-field models such as NeRF, and diffusion models) define the field's progression in both capability and design philosophy. GAN-based pipelines formed the first wave of generative panoramic methods, offering rapid image-to-panorama outpainting and computational efficiency. However, they frequently suffered from mode collapse, texture artifacts, and limited geometric reasoning, which constrained their ability to produce consistent results in sparse or unstructured scenarios. Radiance-field approaches marked a significant shift by introducing explicit 3D awareness and depth-consistent view synthesis, achieving strong geometric fidelity but at the cost of high data requirements and slow training or inference. Diffusion models now represent the current frontier, unifying high visual fidelity, semantic controllability, and multimodal integration through iterative denoising in high-dimensional latent spaces. Taken together, these model families illustrate a clear trajectory: from texture-focused generative methods to geometry-aware radiance fields, and ultimately to controllable, scalable, and multimodal pipelines capable of supporting truly immersive 360° media production.

One of the most prominent trends is the shift from deterministic stitching and inpainting to generative architectures that leverage GANs and diffusion models. These models have enabled richer semantic understanding and improved visual fidelity, allowing for more coherent scene completion and text-guided synthesis. Yet a considerable number of these methods rely on dense input data or structured conditioning signals, limiting their utility in real-world scenarios where only sparse or unregistered NFoV images are available. The assumption of data richness represents a barrier to scalability and accessibility, especially in low-resource or mobile capture environments. This concern is echoed in mobile delivery surveys, which highlight the challenge of scaling 360° content under bandwidth constraints and device limitations [68].

Temporal modeling has also emerged as a major research focus, particularly in transforming still inputs into temporally coherent video sequences. While models like 360DVD and Imagine360 introduce mechanisms for panoramic video synthesis, they often depend on motion priors, dense video captions, or perspective-based inputs. These conditions make them less applicable to use cases requiring few-shot or one-shot generation from minimal visual cues. Furthermore, generative video models frequently lack robust control over temporal consistency, leading to flickering, drift, or incoherent motion propagation across frames.

Scene geometry and depth awareness remain underutilized in most panoramic generation pipelines. Despite the introduction of 360°-specific solutions like SphereDiff and SphereSR, the majority of models treat the spherical canvas as a projection rather than a spatial structure. This overlooks the importance of geometric cues such as symmetry, parallax, and depth, which are essential for producing immersive, explorable environments. Methods such as 360MonoDepth and spherical symmetry-based generation, while addressing geometric aspects, often remain isolated from the broader generative pipeline and have yet to be integrated into end-to-end frameworks.

Recent advances in radiance-field models illustrate how geometry can be leveraged more effectively. The Mip-NeRF 360 [5] model introduces a mipmapping-based, anti-aliased representation that supports level-of-detail rendering and high-fidelity reconstruction of unbounded 360° environments, mitigating aliasing and edge distortions that classic NeRF models cannot handle. Similarly, the NeRF-W model [4] incorporates appearance

embeddings to manage unstructured input with varying lighting and occlusions, achieving robust synthesis from sparse and inconsistent viewpoints. These innovations highlight how radiance-field methods are evolving toward practical panoramic use, bridging the gap between classical geometry reasoning and modern generative pipelines.

The relevance of geometric realism is further reinforced by recent applications in architecture, engineering, and construction (AEC), where 360° media has been adopted for training, simulation, and virtual site monitoring [69].

Another trend is the growing emphasis on interactivity and control. Recent models have begun supporting text-to-360-degree generation, user-guided outpainting, and conditional scene customization. While these capabilities offer greater flexibility, they are still constrained by limitations in layout reasoning, input conditioning complexity, and prompt sensitivity. Cross-view synthesis techniques further demonstrate the feasibility of generating panoramas from abstract or top-down inputs, but these approaches frequently suffer from realism trade-offs and limited generalization across domains.

Finally, practical deployment considerations remain largely absent. Few models account for real-time processing, lightweight inference, or streaming efficiency, critical factors for applications in VR, AR, or edge-device rendering. Although frameworks like TS360 and MagicVideo address deployment and compression indirectly, there is a pressing need for panoramic generation systems that balance quality with performance and accessibility. Recent surveys on bandwidth optimization and edge delivery of 360° video [1,68] emphasize this point, highlighting how the infrastructure for delivering immersive content still faces limitations, even when generation is technically solved.

In summary, the field is trending toward more powerful and flexible generation tools, yet many models fall short of addressing the core challenges related to sparse-input handling, geometric reasoning, temporal coherence, and real-time usability. The insights from this review underscore the importance of unified, efficient, and generalizable architectures that can scale across use cases from creative applications to immersive media production. This forms the basis for the research direction proposed in this thesis, which aims to bridge these gaps through an integrated approach to sparse-input, generative panoramic video synthesis.

5.1. Computational and Deployment Considerations

While substantial progress has been made in the visual quality and flexibility of 360° image and video generation models, computational performance and deployment feasibility remain underreported in the literature. Among the models reviewed, SphereDiff [11], MVSplat360 [31], and 360DVD [45] provide some information regarding training time, GPU memory usage, inference latency, and deployment scalability (see Section 5.2).

Nevertheless, architectural characteristics suggest that many of these models may pose significant computational demands. For instance, diffusion-based models like SphereDiff are known for their iterative sampling process, which can result in slow inference speeds and high memory consumption. Similarly, 360DVD leverages view-dependent rendering techniques inspired by NeRFs, which are computationally intensive and difficult to optimize for real-time scenarios. Imagine360's dual-branch temporal design, incorporating attention-based motion synthesis and spatial transformations, likely incurs high GPU memory requirements during training and inference, especially for longer video sequences.

The lack of standardized benchmarks on runtime efficiency, hardware requirements, or deployment constraints still makes it difficult to assess the practical readiness of these models for use in latency-sensitive or resource-constrained environments. This gap is particularly concerning for applications in mobile VR/AR, simulation training, or remote collaboration, where responsiveness and computational efficiency are as critical as visual fidelity.

5.2. Comparison of Hardware and Processing Times

Table 14 presents a comparative overview of recent 360° video generation models, including both NeRF-based and diffusion-based approaches. The table summarizes key attributes such as model parameters, training compute requirements, inference latency or FPS, the resolution tested, and the hardware used for training or inference. NeRF, Mip-NeRF, and Mip-NeRF 360 are MLP-based methods that require per-scene training, with training times ranging from several hours on multiple TPU cores to over a day on a V100 GPU (NVIDIA Corporation, 2788 San Tomas Expressway Santa Clara, CA 95051, USA), while inference speed is reported only for NeRF. Diffusion-based methods, including 360DVD, VideoPanda, SphereDiff, MVSplat360, and SpotDiffusion, vary widely in computational cost and runtime: some require extensive multi-GPU training, whereas others, like SphereDiff and SpotDiffusion, operate without additional training but may have long inference times or high VRAM requirements; for instance, SphereDiff requires $\leq$40 GB of VRAM, while the VRAM usage of other models is not reported. This table provides a concise reference for evaluating the trade-offs between model complexity, computational resources, and runtime performance in 360° video generation.

Table 14. Performance comparison of recent 360° video generation models *.

Model	Parameters	Training Time	Inference Latency	Resolution Tested	Hardware/ Runtime
NeRF	≈0.6 M	1× NVIDIA V100 GPU, ≈24–48 h per scene	≈30s per frame (≈0.03 FPS)	800 × 800 (Blender), ≈1000 × 1000 (LLFF)	NVIDIA V100 GPU, TensorFlow
Mip-NeRF	612 K	32× TPUv2 cores, ≈2.8 h	–	800 × 800 (Blender, multiscale)	TPUv2 32 cores, JAX
Mip-NeRF 360	9.9 M	32× TPUv2 cores, ≈6.9 h	–	Varied (real 360° scenes)	TPUv2 32 cores, JAX
360DVD	–	100 K steps on WEB360 (GPUs not reported)	–	512 × 1024	–
VideoPanda	–	Trained on 32× NVIDIA A100 GPUs (GPU hours not reported)	–	–	32× A100 GPUs
SphereDiff	–	–	30 s per 512 × 512 image; 20 min per video	512 × 512 per view (90° FoV)	Inference on A100-40 GB or RTX 3090 24 GB
MVSplat360	–	100 K steps on 1–8× A100 GPUs	Inference very slow (multi-step diffusion)	DL3DV-10K frames (4 K original, downsampled)	Training: 1–8× A100; inference: similar GPU
SpotDiffusion	–	–	7 min per 512 × 2048 panorama	512 × 2048	–

* SphereDiff requires $\leq$40 GB VRAM, while the VRAM usage of other models was not reported. *Note*: "–" indicates that it was not reported.

6. Limitations

While this study offers a comprehensive review of AI-driven 360° image and video generation, it is not without limitations. First, the scope of the literature was restricted to publicly available academic and preprint publications from 2021 to 2025. While this range captures recent innovations, it may exclude relevant industrial systems or proprietary mod-

els that have not been documented in open-access venues. Consequently, the findings may not fully reflect the most cutting-edge or large-scale implementations used in commercial applications.

Additionally, the review emphasized papers indexed in IEEE, arXiv, CVPR, and similar repositories, which may introduce selection bias. Although an effort was made to include a diverse set of models and perspectives, from stitching-based methods to diffusion transformers, the selection process may have overlooked significant contributions published in less prominent forums or outside the core computer vision community. This could limit the generalizability of the review's conclusions across subfields like robotics, 3D graphics, or immersive interface design.

Another limitation stems from the inherently subjective nature of evaluating generative models. Many papers rely on qualitative visuals or user preference studies to assess output fidelity, which introduces inconsistency in model comparison. Objective metrics for 360° content, especially for video, remain underdeveloped, making it difficult to benchmark models uniformly. This hinders the ability to draw conclusive insights about model performance across tasks and datasets.

Furthermore, while this study aims to identify architectural and conceptual gaps, it does not experimentally validate new models. As such, its insights remain grounded in secondary analysis and author-reported findings. Future work could complement this review with empirical studies or implementation-based evaluations to measure practical performance and scalability.

Lastly, the scope of this thesis focuses on generating panoramic video content from sparse inputs such as 8–10 NFoV images. While this approach is promising for enhancing accessibility and efficiency, it also introduces constraints related to motion realism, occlusion handling, and depth consistency. These trade-offs may limit the fidelity of generated content in edge cases and complex scene structures. Addressing these limitations will require further refinement of geometry-aware, temporally conditioned generative pipelines.

7. Challenges

Although panoramic video generation has made significant strides, several persistent challenges continue to delay its deployment in real-world applications. Based on literature prevalence and practical impact, the key challenges are summarized below.

7.1. Sparse-Input Handling (Highest Priority)

The most significant challenge lies in handling sparse inputs, as seen in many state-of-the-art models such as 360DVD [45] and Diffusion360 [12], which assume access to dense, overlapping multi-view inputs. When only a few NFoV images are available, these models struggle to preserve spatial coherence and often hallucinate content [26]. Partial remedies include techniques like the Splatter360 model [36], which leverages point cloud splatting for sparse inputs, but its performance still degrades with very few views.

7.2. Temporal Coherence

Maintaining consistent motion across frames is challenging, especially with sparse or near-static inputs. Approaches such as Imagine360 [7] and LAMP [8] utilize motion priors or interpolation techniques; however, their effectiveness decreases significantly with reduced frame density [46]. Flicker and jitter reduce realism and immersion, which are critical for VR and simulation applications.

7.3. Geometry Awareness

Most generative pipelines overlook structural cues such as depth, symmetry, or parallax, focusing on 2D projections or flattened panoramas. Methods like SphereDiff [11]

and 360MonoDepth [65] introduce geometry-aware modules, but these are rarely integrated into full video synthesis workflows, resulting in visually plausible but structurally unrealistic outputs.

7.4. Controllability and Semantic Alignment

Aligning AI-generated content with user intent remains limited. Systems often require retraining or paired data to support text, sketch, or layout-driven control [70]. This constrains creative flexibility and interactive authoring in applications like education, storytelling, or co-creation environments.

7.5. Real-Time Performance and Deployment Constraints

High-fidelity models typically demand substantial computation, large memory footprints, and specialized hardware, hindering their deployment on mobile or bandwidth-constrained platforms. Techniques like adaptive streaming and lightweight architectures [59] exist, but they are rarely combined with full-quality panoramic synthesis.

By systematically addressing these prioritized challenges, from sparse-input handling to deployment constraints, the next generation of 360° panoramic video models can achieve realistic, controllable, and efficient immersive experiences, bridging current experimental prototypes with practical applications.

8. Conclusions and Future Directions

This research presents a comprehensive review of 360° panoramic image and video generation techniques, tracing the field's evolution from early stitching-based methods to modern generative frameworks powered by diffusion models, neural radiance fields, and text-based conditioning. By examining over 40 peer-reviewed and preprint studies, the review categorizes key contributions in areas such as sparse-view synthesis, temporal modeling, multimodal generation, geometric awareness, and immersive rendering.

The findings reveal a clear trajectory toward more flexible and high-fidelity systems. Models like 360DVD [45], Imagine360 [7], and SphereDiff [11] exemplify progress in motion continuity, semantic understanding, and spherical representation. However, limitations persist. Many systems depend on dense input data, structured prompts, or computationally intensive inference. Generalization to sparse perspectives, depth-consistent generation, and real-time processing remains an open research challenge.

In light of these observations, this review establishes a strong foundation for addressing the next wave of innovation in panoramic generation. As immersive applications grow across VR, simulation, training, and entertainment, there is an urgent need for scalable, controllable, and geometry-aware pipelines that can function under real-world constraints.

The direction proposed in this thesis contributes to this vision by focusing on panoramic video generation from as few as 8–10 NFoV images, bridging the gap between efficiency and expressiveness. Through this unified, data-efficient approach, this project aims to push the boundaries of what is possible in 360° content synthesis while offering practical benefits for both academic research and industry deployment.

To address the identified gaps, the proposed method integrates several key innovations. By using sparse NFoV inputs, it directly tackles the challenge of data efficiency and removes reliance on dense camera setups. Incorporating temporal attention and latent motion prediction enables the generation of temporally coherent video, even under limited input conditions. Additionally, the inclusion of geometric priors, such as depth cues or symmetry constraints, enhances spatial realism and alignment. Finally, by leveraging prompt-adaptive control mechanisms, the system supports flexible conditioning through sketch, text, or hybrid inputs, reducing dependency on paired datasets and enabling creative control.

These design choices position the framework as a practical and adaptable solution for immersive content creation in bandwidth-limited, mobile, or creator-driven environments.

Future Directions

As the demand for immersive and interactive media continues to grow, several promising avenues for future research in 360° panoramic generation emerge directly from the challenges in Section 7. Addressing these gaps requires solutions that tackle multiple constraints: sparse-input handling, temporal coherence, geometry awareness, controllability, and deployment efficiency in an integrated manner.

One-shot and few-shot panoramic video generation remains especially challenging because most current models rely on dense input sequences or strong motion priors. In real-world scenarios, however, creators and developers often have access to only a handful of perspective frames due to hardware, bandwidth, or cost constraints. Future research should focus on entropy-aware input conditioning, few-shot motion interpolation, and conditional diffusion bridges to produce temporally coherent 360° video from minimal inputs while preserving spatial realism. Leveraging latent motion prediction and temporal attention mechanisms will further stabilize motion across long sequences. Achieving this would directly reduce dependency on multi-camera rigs, making panoramic generation feasible for mobile VR content creation, low-cost simulation environments, and independent creator workflows.

Ensuring structural realism across sparse multi-view sequences will require embedding geometry- and scene-aware representations, leveraging depth, symmetry, and parallax cues, into generative pipelines. This integration would address current limitations where models overlook structural cues, resulting in outputs that are visually plausible but geometrically inconsistent.

Another critical direction is improving the controllability and expressiveness of generative models. While text-driven synthesis and layout conditioning are now common, they often require paired data or task-specific fine-tuning, limiting deployment flexibility. Future systems should emphasize training-free or prompt-adaptive frameworks that allow dynamic control over generation using hybrid inputs such as sketches, depth cues, or interactive prompts. This capability would empower creators to iterate rapidly without retraining models, enabling more accessible and scalable production pipelines. Wen [9] demonstrates the potential of such systems in human–AI co-creation for 360° video, where real-time bidirectional interaction allows user intent to directly guide generative outcomes.

Future directions also include embedding contextual and semantic awareness into panoramic generation. Behravan et al. [64] highlight the value of hybrid vision–language conditioning in AR, suggesting that panoramic synthesis could benefit from scene-aware cues such as spatial layout, object roles, and environmental affordances to produce content that is both coherent and meaningful in context. Additionally, Liu and Yu [71] illustrate the potential of emotionally responsive video generation, which can communicate mood, influence perception, or support goal-directed experiences. Integrating these capabilities into 360° video generation could enable emotionally resonant applications in education, virtual tourism, and therapeutic simulations, where engagement depends on both realism and affective impact.

Finally, addressing real-time performance and deployment constraints will require optimizing inference efficiency, employing adaptive streaming strategies, and developing lightweight architectures that balance quality and speed. These approaches will ensure deployability across a wide range of hardware platforms, from high-end VR systems to mobile devices. By aligning these innovations, future systems can deliver panoramic

video generation that is not only realistic and controllable but also efficient and accessible, bridging the gap between experimental prototypes and scalable real-world applications.

Author Contributions: Conceptualization, N.A.C., M.N. and J.T.; methodology, N.A.C.; formal analysis, N.A.C. and M.N.; writing—original draft preparation, N.A.C., M.N.; writing—review and editing, M.N. and J.T.; visualization, N.A.C., M.N.; supervision, M.N. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The studies supporting this review are derived from publicly available resources cited within the article. No new data were created or analyzed in this study.

Acknowledgments: The authors would like to thank all those who contributed indirectly to this work through inspiring discussions, critical insights, and support within the broader academic environment.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wong, E.S.; Wahab, N.H.A.; Saeed, F.; Alharbi, N. 360-degree video bandwidth reduction: Technique and approaches comprehensive review. *Appl. Sci.* **2022**, *12*, 7581. [CrossRef]
2. Shafi, R.; Shuai, W.; Younus, M.U. 360-degree video streaming: A survey of the state of the art. *Symmetry* **2020**, *12*, 1491. [CrossRef]
3. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
4. Martin-Brualla, R.; Radwan, N.; Sajjadi, M.S.; Barron, J.T.; Dosovitskiy, A.; Duckworth, D. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7210–7219.
5. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022. [CrossRef]
6. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 5835–5844. [CrossRef]
7. Tan, J.; Yang, S.; Wu, T.; He, J.; Guo, Y.; Liu, Z.; Lin, D. Imagine360: Immersive 360 Video Generation from Perspective Anchor. *arXiv* **2024**, arXiv:2412.03552. [CrossRef]
8. Wu, R.; Chen, L.; Yang, T.; Guo, C.; Li, C.; Zhang, X. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv* **2023**, arXiv:2310.10769.
9. Wen, Y. "See What I Imagine, Imagine What I See": Human-AI Co-Creation System for 360° Panoramic Video Generation in VR. *arXiv* **2025**, arXiv:2501.15456.
10. Ray, B.; Jung, J.; Larabi, M.C. A low-complexity video encoder for equirectangular projected 360 video content. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1723–1727.
11. Park, M.; Kang, T.; Yun, J.; Hwang, S.; Choo, J. SphereDiff: Tuning-free Omnidirectional Panoramic Image and Video Generation via Spherical Latent Representation. *arXiv* **2025**, arXiv:2504.14396.
12. Feng, M.; Liu, J.; Cui, M.; Xie, X. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. *arXiv* **2023**, arXiv:2311.13141. [CrossRef]
13. Wei, L.; Zhong, Z.; Lang, C.; Yi, Z. A survey on image and video stitching. *Virtual Real. Intell. Hardw.* **2019**, *1*, 55–83. [CrossRef]
14. Yao, X.; Hu, Q.; Zhou, F.; Liu, T.; Mo, Z.; Zhu, Z.; Zhuge, Z.; Cheng, J. SpiNeRF: Direct-trained spiking neural networks for efficient neural radiance field rendering. *Front. Neurosci.* **2025**, *19*, 1593580. [CrossRef]
15. Zhang, Q.; Huang, C.; Zhang, Q.; Li, N.; Feng, W. Learning geometry consistent neural radiance fields from sparse and unposed views. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 8508–8517.

16. Zhang, Y.; Zhang, G.; Li, K.; Zhu, Z.; Wang, P.; Wang, Z.; Fu, C.; Li, X.; Fan, Z.; Zhao, Y. DASNeRF: Depth consistency optimization, adaptive sampling, and hierarchical structural fusion for sparse view neural radiance fields. *PLoS ONE* **2025**, *20*, e0321878. [CrossRef] [PubMed]
17. Niemeyer, M.; Barron, J.T.; Mildenhall, B.; Sajjadi, M.S.; Geiger, A.; Radwan, N. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5480–5490.
18. Younis, T.; Cheng, Z. Sparse-View 3D Reconstruction: Recent Advances and Open Challenges. *arXiv* **2025**, arXiv:2507.16406. [CrossRef]
19. Truong, P.; Rakotosaona, M.J.; Manhardt, F.; Tombari, F. Sparf: Neural radiance fields from sparse and noisy poses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4190–4200.
20. Luo, C. Understanding diffusion models: A unified perspective. *arXiv* **2022**, arXiv:2208.11970. [CrossRef]
21. Gupta, P.; Ding, B.; Guan, C.; Ding, D. Generative AI: A systematic review using topic modelling techniques. *Data Inf. Manag.* **2024**, *8*, 100066. [CrossRef]
22. Cheddak, A.; Ait Baha, T.; Es-Saady, Y.; El Hajji, M.; Baslam, M. BERTopic for enhanced idea management and topic generation in Brainstorming Sessions. *Information* **2024**, *15*, 365. [CrossRef]
23. Sumantri, J.S.; Park, I.K. 360 panorama synthesis from a sparse set of images with unknown field of view. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 2386–2395.
24. Akimoto, N.; Matsuo, Y.; Aoki, Y. Diverse plausible 360-degree image outpainting for efficient 3dcg background creation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11441–11450.
25. Liao, K.; Xu, X.; Lin, C.; Ren, W.; Wei, Y.; Zhao, Y. Cylin-Painting: Seamless 360 panoramic image outpainting and beyond. *IEEE Trans. Image Process.* **2023**, *33*, 382–394. [CrossRef]
26. Zheng, D.; Zhang, C.; Wu, X.M.; Li, C.; Lv, C.; Hu, J.F.; Zheng, W.S. Panorama Generation From NFoV Image Done Right. *arXiv* **2025**, arXiv:2503.18420. [CrossRef]
27. Nie, L.; Lin, C.; Liao, K.; Liu, S.; Zhao, Y. Deep rectangling for image stitching: A learning baseline. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5740–5748.
28. Nie, L.; Lin, C.; Liao, K.; Liu, S.; Zhao, Y. Parallax-tolerant unsupervised deep image stitching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7399–7408.
29. Zhou, T.; Li, H.; Wang, Z.; Luo, A.; Zhang, C.L.; Li, J.; Zeng, B.; Liu, S. Recdiffusion: Rectangling for image stitching with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 2692–2701.
30. Trepkowski, C.; Eibich, D.; Maiero, J.; Marquardt, A.; Kruijff, E.; Feiner, S. The effect of narrow field of view and information density on visual search performance in augmented reality. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; pp. 575–584.
31. Chen, Y.; Zheng, C.; Xu, H.; Zhuang, B.; Vedaldi, A.; Cham, T.J.; Cai, J. MVSplat360: Feed-forward 360 scene synthesis from sparse views. In Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24, Vancouver, BC, Canada, 10–15 December 2024.
32. Bao, C.; Zhang, X.; Yu, Z.; Shi, J.; Zhang, G.; Peng, S.; Cui, Z. Free360: Layered Gaussian Splatting for Unbounded 360-Degree View Synthesis from Extremely Sparse and Unposed Views. *arXiv* **2025**, arXiv:2503.24382. [CrossRef]
33. Liu, F.; Sun, W.; Wang, H.; Wang, Y.; Sun, H.; Ye, J.; Zhang, J.; Duan, Y. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. *arXiv* **2025**, arXiv:2408.16767. [CrossRef]
34. Nagoor Kani, B.R.; Lee, H.Y.; Tulyakov, S.; Tulsiani, S. UpFusion: Novel View Diffusion from Unposed Sparse View Observations. In Proceedings of the Computer Vision—ECCV 2024: 18th European Conference, Milan, Italy, 29 September–4 October 2024; Proceedings, Part LXXVI; Springer: Berlin/Heidelberg, Germany, 2024; pp. 179–195. [CrossRef]
35. Irshad, M.Z.; Zakharov, S.; Liu, K.; Guizilini, V.; Kollar, T.; Gaidon, A.; Kira, Z.; Ambrus, R. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 9187–9198.
36. Chen, Z.; Wu, C.; Shen, Z.; Zhao, C.; Ye, W.; Feng, H.; Ding, E.; Zhang, S.H. Splatter-360: Generalizable 360° Gaussian Splatting for Wide-baseline Panoramic Images. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 21590–21599.
37. Dastjerdi, M.R.K.; Hold-Geoffroy, Y.; Eisenmann, J.; Khodadadeh, S.; Lalonde, J.F. Guided co-modulated gan for 360° field of view extrapolation. In Proceedings of the 2022 International Conference on 3D Vision (3DV), Prague, Czech Republic, 12–16 September 2022; pp. 475–485.

38. Zhang, C.; Wu, Q.; Gambardella, C.C.; Huang, X.; Phung, D.; Ouyang, W.; Cai, J. Taming Stable Diffusion for Text to 360° Panorama Image Generation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024.
39. Wang, H.; Xiang, X.; Fan, Y.; Xue, J.H. Customizing 360-degree panoramas through text-to-image diffusion models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 4933–4943.
40. Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; Feng, J. Magicvideo: Efficient video generation with latent diffusion models. *arXiv* **2022**, arXiv:2211.11018.
41. An, J.; Zhang, S.; Yang, H.; Gupta, S.; Huang, J.B.; Luo, J.; Yin, X. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv* **2023**, arXiv:2304.08477.
42. Chen, T.S.; Lin, C.H.; Tseng, H.Y.; Lin, T.Y.; Yang, M.H. Motion-conditioned diffusion model for controllable video synthesis. *arXiv* **2023**, arXiv:2304.14404. [CrossRef]
43. Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; Zhou, J. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv* **2023**, arXiv:2311.04145.
44. Ni, H.; Shi, C.; Li, K.; Huang, S.X.; Min, M.R. Conditional image-to-video generation with latent flow diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18444–18455.
45. Wang, Q.; Li, W.; Mou, C.; Cheng, X.; Zhang, J. 360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 6913–6923.
46. Hu, Y.; Chen, Z.; Luo, C. LaMD: Latent Motion Diffusion for Image-Conditional Video Generation. *Int. J. Comput. Vis.* **2025**, *133*, 4384–4400. [CrossRef]
47. Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. FVD: A New Metric for Video Generation. 2019. Available online: https://openreview.net/pdf?id=rylgEULtdN (accessed on 1 August 2025).
48. Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv* **2018**, arXiv:1812.01717.
49. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. **2022**, arXiv:2104.08718. [CrossRef]
50. Xie, K.; Sabour, A.; Huang, J.; Paschalidou, D.; Klar, G.; Iqbal, U.; Fidler, S.; Zeng, X. VideoPanda: Video Panoramic Diffusion with Multi-view Attention. *arXiv* **2025**, arXiv:2504.11389.
51. Xia, Y.; Weng, S.; Yang, S.; Liu, J.; Zhu, C.; Teng, M.; Jia, Z.; Jiang, H.; Shi, B. PanoWan: Lifting Diffusion Video Generation Models to 360° with Latitude/Longitude-aware Mechanisms. *arXiv* **2025**, arXiv:2505.22016.
52. Liu, J.; Lin, S.; Li, Y.; Yang, M.H. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 6144–6153.
53. Wu, S.; Tang, H.; Jing, X.Y.; Zhao, H.; Qian, J.; Sebe, N.; Yan, Y. Cross-view panorama image synthesis. *IEEE Trans. Multimed.* **2022**, *25*, 3546–3559. [CrossRef]
54. Li, J.; Bansal, M. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 21878–21894.
55. Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv* **2024**, arXiv:2408.06072.
56. Choi, D.; Jang, H.; Kim, M.H. OmniLocalRF: Omnidirectional Local Radiance Fields from Dynamic Videos. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024. [CrossRef]
57. Zhang, M.; Chen, Y.; Xu, R.; Wang, C.; Yang, J.; Meng, W.; Guo, J.; Zhao, H.; Zhang, X. PanoDiT: Panoramic Videos Generation with Diffusion Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Phfodelphia, PA, USA, 20 February–4 March 2025; Volume 39, pp. 10040–10048.
58. Xiong, Z.; Chen, Z.; Li, Z.; Xu, Y.; Jacobs, N. PanoDreamer: Consistent Text to 360-Degree Scene Generation. *arXiv* **2025**, arXiv:2504.05152. [CrossRef]
59. Huo, Y.; Kuang, H. Ts360: A two-stage deep reinforcement learning system for 360-degree video streaming. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
60. Yoon, Y.; Chung, I.; Wang, L.; Yoon, K.J. Spheresr: 360° image super-resolution with arbitrary projection via continuous spherical image representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5677–5686.
61. Baniya, A.A.; Lee, T.K.; Eklund, P.W.; Aryal, S. Omnidirectional video super-resolution using deep learning. *IEEE Trans. Multimed.* **2023**, *26*, 540–554. [CrossRef]
62. Yang, S.; Tan, J.; Zhang, M.; Wu, T.; Li, Y.; Wetzstein, G.; Liu, Z.; Lin, D. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *arXiv* **2024**, arXiv:2408.13252.

63. Kou, S.; Zhang, F.L.; Nazarenus, J.; Koch, R.; Dodgson, N.A. OmniPlane: A Recolorable Representation for Dynamic Scenes in Omnidirectional Videos. *IEEE Trans. Vis. Comput. Graph.* **2025**, *31*, 4095–4109. [CrossRef] [PubMed]
64. Behravan, M.; Matković, K.; Gračanin, D. Generative AI for context-aware 3D object creation using vision-language models in augmented reality. In Proceedings of the 2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR), Lisbon, Portugal, 27–29 January 2025; pp. 73–81.
65. Rey-Area, M.; Yuan, M.; Richardt, C. 360monodepth: High-resolution 360deg monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3762–3772.
66. Hara, T.; Mukuta, Y.; Harada, T. Spherical image generation from a few normal-field-of-view images by considering scene symmetry. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6339–6353. [CrossRef]
67. Lu, C.N.; Chang, Y.C.; Chiu, W.C. Bridging the visual gap: Wide-range image blending. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 843–851.
68. Mahmoud, M.; Rizou, S.; Panayides, A.S.; Kantartzis, N.V.; Karagiannidis, G.K.; Lazaridis, P.I.; Zaharis, Z.D. A survey on optimizing mobile delivery of 360 videos: Edge caching and multicasting. *IEEE Access* **2023**, *11*, 68925–68942. [CrossRef]
69. Shinde, Y.; Lee, K.; Kiper, B.; Simpson, M.; Hasanzadeh, S. A Systematic Literature Review on 360° Panoramic Applications in Architecture, Engineering, and Construction (AEC) Industry. *J. Inf. Technol. Constr.* **2023**, *28*, 405–437. [CrossRef]
70. Partarakis, N.; Zabulis, X. A review of immersive technologies, knowledge representation, and AI for human-centered digital experiences. *Electronics* **2024**, *13*, 269. [CrossRef]
71. Liu, C.; Yu, H. Ai-empowered persuasive video generation: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–31. [CrossRef]