A Systematic Review of Consistent Multi-Character Image Generation

Thathsarani Sandarekha[a]

[a]Robert Gordon University, School of Computing, Aberdeen, United Kingdom

ABSTRACT

The consistent generation of multiple characters in images poses significant challenges in preserving identity fidelity, spatial coherence, and semantic alignment across diverse scenarios. This review provides an in-depth analysis of state-of-the-art diffusion-based approaches addressing these challenges through innovative attention mechanisms, lightweight and training-free strategies, and modular frameworks. Key advancements are evaluated using robust metrics such as CLIP-based alignment, identity consistency, and layout diversity, showcasing improvements in fidelity and narrative coherence. The paper also explores transformative applications in storytelling, gaming, and animation, while identifying critical challenges like scalability, generalization, and real-time efficiency. By synthesizing these developments, this review highlights the progress made and outlines future directions to advance the field of multi-character image generation.

Keywords: Multi-character generation, Visual storytelling, Text-to-image, Identity preservation, Character consistency.

1. INTRODUCTION

Despite progress in this domain, several persistent challenges remain. Chief among them are issues related to identity blending, where visual features of characters begin to merge when placed in proximity, and spatial misalignment, which leads to unnatural compositions in crowded scenes. Scalability also remains a limitation, as current pipelines often falter when attempting to render high-density scenes containing more than three characters[1,2] Diffusion-based approaches, which form the backbone of modern text-to-image synthesis, offer superior image quality but often lack mechanisms to explicitly enforce identity consistency or resolve spatial conflicts when multiple subjects are present.[3] Likewise, attention-based control strategies such as cross-attention provide spatial anchoring but are prone to degradation when dealing with tightly packed, multi-subject compositions.[4]

This review synthesizes recent advancements in multi-character image generation, with a focus on systems that maintain identity fidelity, compositional coherence, and visual alignment. It evaluates both training-based and training-free pipelines, as well as inference-time strategies that use segmentation, graph guidance, layout control, and prompt engineering to achieve consistent results. The paper also assesses evaluation protocols and benchmarks that have emerged to support this growing subfield. Key strengths and limitations of state-of-the-art methods are critically discussed, offering insights into the broader applicability of current solutions and outlining areas for future research.

This paper is organized as follows: Section 2 presents the methodology for the literature review, including eligibility criteria, search strategies, and commentary on related work. Section 3 discusses existing methods in detail. Section 4 identifies current technical challenges. Section 5 describes evaluation metrics. Section 6 surveys practical applications, and Section 7 outlines future research directions. Section 8 concludes with a summary and roadmap.

Further author information: (Send correspondence to T.S.)
T.S.: E-mail: thathsarani.20211422@iit.ac.lk

2. METHODOLOGY

2.1 Eligibility Criteria

To ensure a focused and rigorous review, only studies that addressed multi-character generation in the context of image synthesis were selected. Works that proposed methods for character consistency, spatial layout control, semantic disentanglement, or appearance preservation across multiple subjects were included. Importantly, the scope was limited to approaches that targeted multi-character scenes, particularly those using text-to-image models or diffusion-based systems. Studies that dealt exclusively with single-character personalization (e.g., DreamBooth,[5] Textual Inversion in isolation) or face-only editing tasks without addressing compositional concerns were excluded.

Further exclusion criteria applied to works that focused solely on video generation or segmentation without any connection to multi-subject image synthesis. However, multi-character video generation systems such as Follow-Your-MultiPose were included if they offered novel solutions to identity preservation across temporally coherent image sequences and shared architecture with static image pipelines.

2.2 Search Strategy

A comprehensive search strategy was designed to identify relevant studies on multi-character image generation and consistency techniques in visual storytelling. The search was conducted across six academic databases: Google Scholar, IEEE Xplore, ACM Digital Library, arXiv, SpringerLink, and ScienceDirect. The search period was limited to the years 2014 to 2024, ensuring both the inclusion of foundational work and the most recent advances in the field.

To capture the full breadth of relevant literature, a broad keyword set and Boolean logic combinations were used. Initial keywords included: "multi-character image generation," "consistent character image generation," "multi-character representation in images," "visual storytelling with multiple characters," "multi-concept generation," "multi-ID synthesis," and "image synthesis with multiple characters."

The initial query resulted in 45 papers, from which 20 studies were ultimately selected for inclusion following a two-phase screening process. First, duplicate records were removed and titles and abstracts were screened for relevance to multi-character generation. Second, full-text evaluations were conducted to ensure the selected works addressed either identity preservation or contextual alignment in multi-character image synthesis. Exclusion criteria included works that focused exclusively on single-character personalization, general scene understanding without generation, or unrelated domains such as object detection and captioning.

The PRISMA flow diagram 1 summarizes the identification, screening, eligibility assessment, and inclusion process. This diagram provides a transparent overview of how the final corpus of 20 papers was selected from the initial pool, in alignment with systematic review best practices.

2.3 Related Reviews

To the best of our knowledge, no prior systematic or narrative review has been published that specifically focuses on consistent multi-character image generation. While some surveys have explored related domains—such as text-to-image synthesis, generative modeling, or visual storytelling—none have critically examined the subset of methods aimed at preserving character identity and compositional fidelity in multi-subject scenes. This review therefore addresses a clear gap in the literature by providing a targeted synthesis of algorithms, techniques, and evaluation strategies for consistent multi-character image generation.

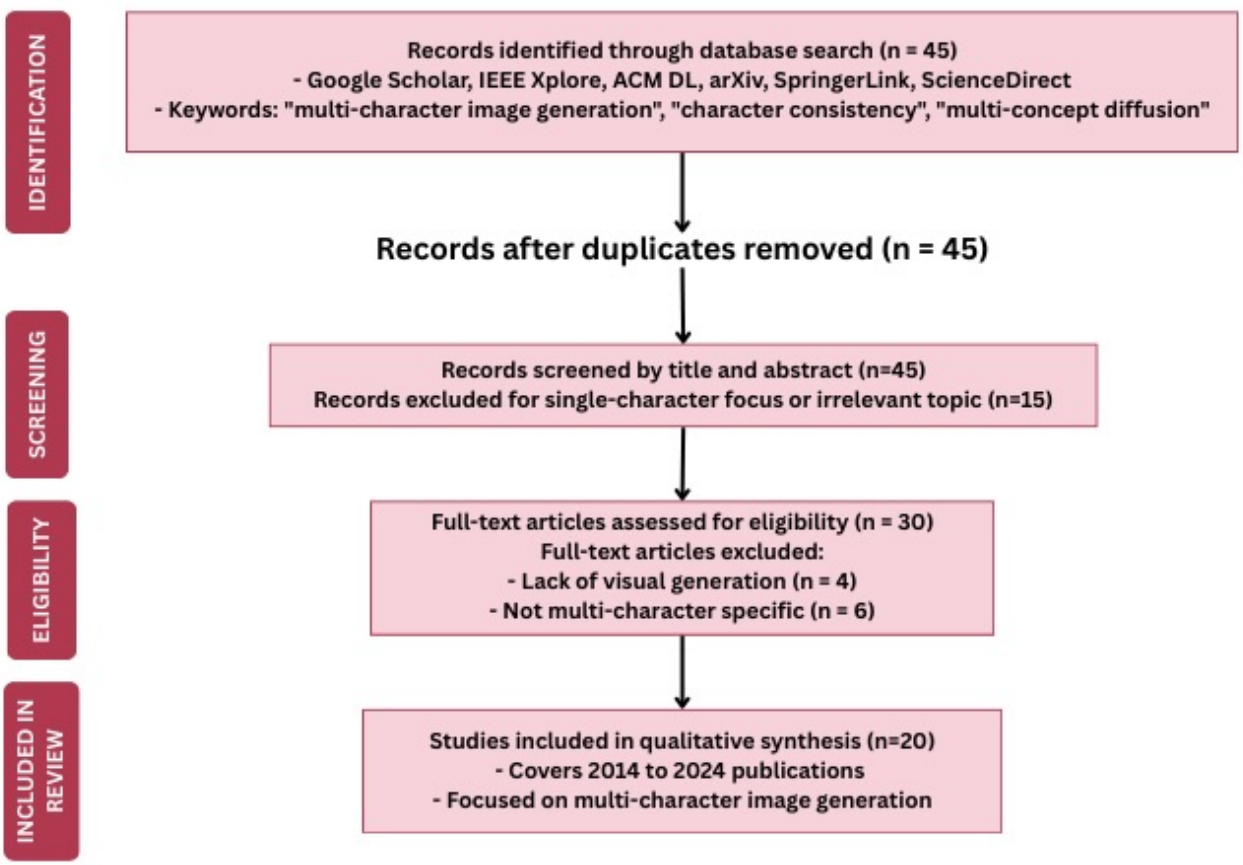

Figure 1. PRISMA flow diagram

3. EXISTING APPROACHES

3.1 Rule-Based and Probabilistic Pre-Diffusion Foundations

Before the emergence of diffusion-based generative models, early approaches to multi-character scene synthesis were grounded in probabilistic planning, rule-based modeling, and motion graph techniques. These methods prioritized scene-level coherence, realistic character motion, and plausible inter-character interactions, laying important conceptual groundwork for later systems. One of the most influential works in this space is Generating and Ranking Diverse Multi-Character.[6] Interactions system, Generating and Ranking Diverse Multi-Character Interactions, which introduced a generate-and-rank pipeline to choreograph multi-character scenes using Subject-Verb-Object (SVO) triplets. Drawing from a motion-captured dataset of professional stunt actors, the system used probabilistic synthesis and a PageRank-inspired scoring model to produce and rank thousands of interaction sequences based on criteria such as deformation energy, spatial mismatch, collision avoidance, and scene diversity.

Generating and Ranking Diverse Multi-Character[6] method stood out for its dual focus on interaction diversity and physical plausibility, with support for up to 20-character scenes and advanced features like interaction patch replacement, cycle synchronization, and user-editable action swapping. However, the system operated on motion data rather than generating new visuals and was limited to modeling only pairwise interactions per event. The lack of appearance customization and dependence on manually labeled motion clips also restricted scalability. Despite these limitations, its emphasis on layout coherence, diversity, and ranking mechanisms foreshadowed core objectives in modern diffusion-based and language-guided image generation systems, making it a foundational contribution to the field of consistent multi-character storytelling.

3.2 Modular Diffusion Pipelines for Scene and Layout Control

Modular diffusion pipelines represent a significant advancement in multi-character story generation by structuring the image synthesis process into distinct stages such as prompt planning, layout construction, and controllable diffusion. Systems like TaleCrafter[7] and TheaterGen[8] begin by using large language models to decompose story text into structured prompts and spatial layouts, enabling precise placement of characters through bounding boxes or layout sketches. These layouts act as blueprints for the generation stage, where techniques such as LoRA modules, Prompt Books, or attention masks are applied to maintain visual coherence. DreamStory[9] innovates by eliminating external layout inputs and instead relying on segmentation maps and masked attention mechanisms to separate subject features and maintain identity consistency across scenes without retraining.

While modular pipelines offer strong identity preservation and layout control, they face several limitations. Systems like CogCartoon[10] use lightweight, reusable character plugins for scalable identity management, but struggle with dynamic expressions or stylistic variation. Additionally, many modular frameworks treat layout and appearance consistency as separate components, leading to issues like character drift or spatial misalignment over time. Most also assume static camera perspectives and fixed scene configurations, making them less effective in dynamic or cinematic contexts. Despite these challenges, modular pipelines provide a powerful foundation for scalable and controllable story generation. Future systems must unify identity modeling with spatial reasoning and narrative flow to support more complex, interactive storytelling applications.

3.3 Multimodal and MLLM-Guided Joint Text-Image Story Generation

Recent advances in multimodal large language models (MLLMs) have enabled unified frameworks for joint text-image story generation, moving beyond layout-based diffusion pipelines. These systems aim to generate visually coherent narratives by reasoning over semantic, temporal, and stylistic consistency across multiple frames. SEED-Story is a leading example, leveraging a Multimodal Attention Sink to capture long-range dependencies across up to 30 frames, with a fine-tuned SDXL-based decoder generating stylized visuals that maintain character identity and pose. It introduces the large-scale StoryStream dataset to benchmark multi-turn generation, though its reliance on animated data and large-scale retraining limits generalization and real-world adaptability.

Other systems like DiffSensei[11] and Dialogue Director[12] extend this paradigm into specialized domains. DiffSensei targets manga generation with a two-stage pipeline that combines masked attention for panel layout with an MLLM-driven personalization module, offering strong region-level control and expression fidelity but at high computational cost. Dialogue Director, on the other hand, focuses on dialogue-to-visual translation using a three-agent pipeline—script parser, cinematographer, and storyboard planner—integrating cinematic conventions like camera angles and shot composition. While these models improve semantic grounding and narrative alignment, they face challenges in generalization, efficient scaling, and handling free-form or dynamic scenes. As MLLMs evolve, future research will need to address joint optimization of dialogue flow, character consistency, and visual realism to fully realize story-aware, multi-character generation.

3.4 Training-Free Personalized Character Generation

Training-free methods have emerged as a scalable solution for consistent character generation, offering fast inference and modular reuse without requiring model fine-tuning. These approaches inject subject-specific information into pre-trained diffusion models using techniques like embedding augmentation, attention manipulation, or patch-level feature injection. PortraitBooth[13] exemplifies this by combining facial embeddings with latent-level identity loss and localized cross-attention masking to enable high-fidelity portrait generation and expression variation. While effective, its scope is limited to single-subject portrait compositions. Character-Adapter[14] expands this capability to full-body generation using prompt-guided regional adapters, enabling multi-subject control but relying on accurate segmentation for effectiveness. ConsiStory further advances scalability through shared attention and feature injection across batches, allowing character consistency across scenes, though it inherits stylistic and boundary limitations from its SDXL backbone.

Despite their efficiency, training-free pipelines face challenges in handling occlusion, pose diversity, and long-range temporal consistency. They often depend on static reference embeddings or segmentation priors that limit adaptability in open-domain storytelling. Still, these methods form a vital foundation for modular, identity-aware generation frameworks. Their plug-and-play architecture makes them highly compatible with layout

control, stylization, and scene variation modules, providing a practical alternative to resource-intensive tuning-based systems. As future work explores hybrid pipelines and region-aware generation, training-free systems will remain central to building consistent multi-character narratives at scale.

3.5 Inference-Time Attention Modulation and Masked Cross-Attention

Inference-time attention modulation methods offer a training-free solution to the challenge of preserving character distinction in multi-subject scenes. Systems like InstantFamily introduce masked cross-attention mechanisms that direct subject-specific prompts and embeddings to predefined spatial regions, effectively preventing identity blending and feature leakage. Similarly, SPDiffusion[15] employs early-stage semantic protection by computing cross-attention maps and enforcing binary attention masks to ensure attribute disentanglement, particularly in multi-concept compositions. Both methods show strong performance in dense scenes, with InstantFamily supporting pose control and SPDiffusion excelling in attribute-level separation. However, these systems depend heavily on accurate spatial priors, such as face embeddings or segmentation quality, and often falter when masks are misaligned or under-specified.

Extending this concept to video, Follow-Your-MultiPose[16] applies pose-guided spatial masking and prompt filtering to maintain subject coherence across temporal frames, offering efficient multi-character video generation without retraining. Although these methods scale well in static or moderately dynamic settings, they are limited in their ability to handle complex interactions, dynamic camera motion, or depth-aware positioning. Overall, inference-time attention modulation presents a modular and scalable approach to character disentanglement, but future advancements must enhance spatial precision and scene adaptability to support more interactive and immersive storytelling environments.

3.6 Segmentation-Guided and Mask-Based Feature Decoupling

Segmentation-guided and mask-based feature decoupling methods offer a robust strategy for managing character identity in crowded scenes where spatial overlap or occlusion is common. Unlike attention modulation techniques that separate identities during the generation process, these methods explicitly define spatial regions in advance using masks or layout priors. MuDI[17] exemplifies this approach by generating subject-specific masks and injecting LoRA[18] modules through region-constrained attention selectors. This modularity enables each character to be synthesized in isolation and composed later, significantly improving multi-ID consistency in scenes with three or more subjects. However, MuDI's dependence on accurate segmentation limits its generalizability, and its sequential composition strategy introduces computational overhead at scale.

OMG[19] employs a two-stage framework, first generating a coarse layout and then injecting identity-specific noise into masked regions using concept blending. This allows for flexible integration of identity modules like InstantID or LoRA and supports occlusion-aware rendering without retraining. Both MuDI and OMG outperform traditional personalization baselines in dense scenarios, but their success hinges on high-quality segmentation and layout inference. When mask accuracy falters, these systems can suffer from visual artifacts and compositional degradation. Nonetheless, by encoding spatial priors explicitly, segmentation-guided methods provide strong identity separation and layout control, making them valuable for building scalable, multi-character storytelling systems.

3.7 Tuning-Based Identity and Style Preservation Frameworks

Tuning-based frameworks remain indispensable for achieving high-fidelity character consistency across complex visual narratives, especially when scenes involve varied poses, long-term character appearances, or stylistic adaptation. These systems fine-tune either lightweight adapters or projection layers per subject, offering superior identity robustness compared to training-free approaches. OneActor[20] exemplifies this by learning cluster-conditioned embedding offsets, allowing efficient and accurate identity preservation with minimal tuning time. It supports multi-subject generation through spatial masking and cluster guidance but faces limitations when identity clusters overlap or scenes grow too crowded. StoryMaker[21] takes a spatially constrained approach, fusing facial and full-body embeddings via a Positional-aware Perceiver Resampler (PPR), reinforced by segmentation-aware attention losses. Its ability to preserve full-body attributes and support ControlNet-based pose manipulation makes it particularly effective in multi-character story contexts.

Despite their accuracy, these methods come with trade-offs. Each new character requires a separate tuning phase, limiting scalability and increasing computational demands, especially when merging multiple tuned characters in the same scene. Moreover, they often lack joint optimization mechanisms for identity and spatial composition, leading to occasional misalignments. Nevertheless, for applications where visual fidelity is paramount—such as illustrated narratives, stylized character renderings, or cinematic storyboards—tuning-based pipelines offer unmatched control and consistency, provided the required reference data and compute resources are available.

3.8 Graph-Based and Knowledge-Enhanced Consistency Modeling

Graph-based approaches to multi-character generation enhance consistency by modeling characters not just through appearance but also through their semantic relationships and contextual roles within a story. StoryWeaver[22] exemplifies this direction by introducing a Character Concept Graph (C-CG) and a Knowledge-Enhanced Spatial Guidance (KE-SG) module, which embed object, attribute, and relational data into the generation pipeline. These structures guide attention to respect character positioning and interaction boundaries, improving identity preservation across multi-character scenes. Evaluated on the TBC-Bench dataset, StoryWeaver outperforms visual-only baselines like DreamBooth[5] and Mix-of-Show[23] by capturing narrative-aligned character dynamics such as friendship, rivalry, and familial ties.

However, the effectiveness of graph-based models like StoryWeaver is constrained by their reliance on structured captions, vision-language model accuracy, and stylistically homogeneous training data. The approach assumes even spatial layouts and struggles with abstract narratives, dense scenes, or evolving relationships that require real-time adaptation. Still, by anchoring character identity in narrative context rather than just pixels, these systems offer a principled pathway toward scalable and semantically rich multi-character generation—particularly useful for comic books, visual novels, and animated storytelling where relational dynamics are central to scene coherence.

3.9 Multi-Subject and Pose-Guided Video Generation

Extending multi-character generation into the video domain introduces new challenges such as maintaining character identity, pose alignment, and temporal coherence across frames. Follow-Your-MultiPose[16] addresses this by decomposing multi-character prompts into subject-specific identity tokens and spatial regions, aligning them with pose sequences using OpenPose. Each character is assigned its own ControlNet branch and spatial prompt mask, allowing the system to preserve identity and action consistency across video frames without retraining. This training-free pipeline outperforms prior methods like Text2Video-Zero[24] and Follow-Your-Pose[25] in frame-wise consistency and identity separation, demonstrating that inference-only mechanisms can be effectively extended to dynamic, multi-subject video generation.

Despite these advantages, Follow-Your-MultiPose struggles with scalability beyond three characters due to high computational overhead and lacks the ability to model inter-character interactions such as synchronized gestures or gaze coordination. The system assumes static backgrounds and fixed camera perspectives, limiting cinematic expressiveness. It also lacks temporal memory mechanisms, making it sensitive to pose jitter and identity drift during frame transitions. Still, video-based generation offers immense potential for animated storybooks, motion comics, and educational media. Future systems must integrate temporal attention and narrative modeling to support long-form, stylistically coherent, multi-character storytelling in dynamic environments.

3.10 Benchmark Datasets and Evaluation Trends

The evaluation of multi-character generation systems has advanced through the introduction of specialized benchmarks that assess identity preservation, spatial consistency, and semantic alignment. Datasets like StoryStream provide long-form narratives paired with over 250,000 annotated frames, enabling evaluation of stylistic coherence and temporal consistency in MLLM-based models like SEED-Story.[26] Meanwhile, DS-500 focuses on character-level fidelity across short sequences using tools like CLIP-T and D&C-DS to isolate and compare embeddings frame by frame. TBC-Bench adds relational grounding to the mix by measuring the correct rendering of relationships such as siblings or adversaries using frame accuracy and class-wise F1 scores. In domain-specific contexts,

MangaZero evaluates panel-level pose variation and dialogue integration in manga-style compositions, making it useful for stylized generation though less generalizable to open-world narratives.

Despite these developments, current benchmarks often fall short in capturing long-range consistency, emotional nuance, and group dynamics. Most datasets focus on fixed scenes or short interactions, limiting their usefulness in evaluating open-ended or evolving narratives. Evaluation metrics such as CLIP-I, CLIP-T, DINO, and DreamSim provide coarse and fine-grained visual-textual alignment, but lack sensitivity to narrative continuity and interaction depth. Human studies continue to play a vital role in assessing realism and story fidelity, underscoring the need for richer, human-centric evaluation protocols. As the field matures, future benchmarks must expand to support temporally grounded, interaction-aware, and stylistically diverse stories that align with the goals of scalable, narrative-consistent character generation.

4. CHALLENGES

Multi-character generation remains challenging due to issues in identity blending, spatial alignment, scalability, and generalization. Frameworks like UniPortrait address identity blending using adaptive ID routing but struggle with attribute-based routing and generalization to diverse, non-facial attributes.[27] Similarly, MuDI's Seg-Mix resolves identity blending but suffers fidelity loss with more subjects,[17] while SPDiffusion's SP-Mask-guided attention mitigates attribute confusion but struggles with overlapping regions and nuanced interactions.[15]

Scalability challenges are evident in InstantFamily, which employs masked cross-attention and multimodal embeddings but remains resource-intensive.[28] TaleCrafter[7] improves layout flexibility but cannot automate sketch generation for complex scenes,[7] and CogCartoon addresses data scarcity with lightweight plugins but struggles with stylistic diversity and large-scale narratives.[10]

Generalization remains difficult. SEED-STORY ensures coherence with multimodal attention sinks but is limited to cartoon datasets,[26] while Character-Adapter enhances efficiency without additional training but faces issues with attention map accuracy and diverse datasets.[14] Follow-Your-MultiPose reduces feature overlap but struggles with highly dynamic scenes and low-quality inputs.[16]

Contextual and temporal coherence also pose challenges. DreamStory uses MMSA and MMCA to reduce subject blending but falters in dense scenes and misaligns subjects with textual descriptions.[9] StoryWeaver uses spatial priors for identity blending but relies on curated datasets, limiting novel interactions.[22] Dialogue Director achieves multi-view consistency with structured prompts but struggles with dynamic interactions and rapid transitions.[12]

Integrating techniques like UniPortrait and SPDiffusion with scalable frameworks such as TaleCrafter[7] and CogCartoon could significantly improve multi-character generation. Combining SEED-STORY's coherence strategies with Character-Adapter's modular efficiency may enhance generalization, while addressing scalability and fidelity gaps will enable robust, contextually aware generation across diverse domains.

5. EVALUATION METRICS

Evaluation metrics are critical for evaluating multi-character image generation systems, focusing on identity preservation, text-image alignment, scene coherence, visual quality, and subject consistency.

5.1 Identity Preservation and Subject Consistency

Maintaining character identity is vital in multi-character generation. Metrics like Face Similarity (Face Sim) calculate the cosine similarity between embeddings of reference and generated faces to ensure identity consistency[21,27]:

$$\text{Face Sim} = \cos(\phi(f), \phi(g)), \tag{1}$$

where $\phi(f)$ and $\phi(g)$ are embeddings of the reference and generated images.

For single-ID scenarios, Identity Preservation Metric (SingleSim) evaluates fidelity using cosine similarity between embeddings:[28]

$$\text{SingleSim}(A, A') = \cos(E_f(A), E_f(A')). \tag{2}$$

To assess identity preservation across multiple individuals, Multi-ID Similarity Metric (MultiSim) penalizes identity mixing and measures independence between distinct identities:[28]

$$\text{MultiSim}(A, B) = \frac{\text{SingleSim}(A, A') + \text{SingleSim}(B, B')}{2} + (1 - \text{SingleSim}(A', B')) \tag{3}$$

Metrics like IDA (ArcFace) and DINO-I further evaluate identity consistency in multi-ID setups[22,19,17].

Metrics like DreamSim and D&C-DS measure multi-subject consistency[17,29]. The Expression Coefficient, used in PortraitBooth, evaluates expression accuracy relative to prompts.[9]

5.2 Text-Image Alignment

Text-image alignment metrics assess how well generated visuals adhere to the semantic context of the input text. CLIP Score[30] evaluates the fidelity of text prompts to generated images[31,16] while CLIP Text-Image Alignment (CLIP-T) measures semantic alignment between scenes and text using CLIP embeddings[9,13,12].

To ensure consistency between generated and reference images, CLIP Image Similarity (CLIP-I) is used[22,14]. TextSim directly measures how well the generated output matches textual descriptions:[28]

$$\text{TextSim}(P_t, A') = \cos(E_{ct}(P_t), E_{ci}(A')). \tag{4}$$

5.3 Scene Coherence

Ensuring spatial and temporal coherence is vital for multi-character scenes. Plausibility (P) evaluates scene quality based on factors like motion naturalness, residual errors, and diversity:[6]

$$P = w_1 P_{deform} + w_2 P_{residual} + w_3 P_{col} + w_4 P_{div} + w_5 P_{pref}. \tag{5}$$

Spatiotemporal Error (ξ) measures alignment in translation, rotation, and temporal dimensions:[6]

$$\xi = \|v' - v\| + w_a|\theta' - \theta| + w_t|\delta' - \delta|. \tag{6}$$

The Scene Similarity Metric ensures diversity by comparing paths and events between generated scenes:[6]

$$\text{similarity}(S, S') = \frac{1}{\text{dist}(S, S') + \epsilon}. \tag{7}$$

Metrics such as Frame Consistency (FC) and Alignment Accuracy further measure temporal coherence and spatial correctness across frames[16,8].

5.4 Visual Quality

Image quality metrics assess the aesthetic appeal and technical fidelity of generated outputs. Frechet Inception Distance (FID) compares the distributions of generated and real images, emphasizing quality and diversity[3,21,11,26,15]:

$$\text{FID} = \|\mu_g - \mu_r\|^2 + \text{Tr}(\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{1/2}), \tag{8}$$

where (μ_g, Σ_g) and (μ_r, Σ_r) are the means and covariances of generated and real distributions.

Other metrics include Aesthetic Score (AES) for visual appeal,[9] LAION-Aesthetics (LAION-Aes) for aesthetic quality,[27] and Average Frechet Inception Distance (aFID) for image fidelity and diversity across scenes.[8]

5.5 User Studies

Frameworks such as TaleCrafter,[7] SEED-STORY,[26] Follow-Your-MultiPose,[16] DreamStory,[9] and ConsiStory[31] rely on user studies to assess text-image correspondence, visual quality, and narrative coherence. Subjective evaluations also consider cinematic principles in frameworks like Dialogue Director.

5.6 Qualitative Evaluation

Baseline methods frequently used for benchmarking include Textual Inversion (TI),[33] DreamBooth (DB),[5] LoRA (Low-Rank Adaptation),[18] Custom Diffusion,[34] IP-Adapter,[35] FastComposer,[36] and BLIP-Diffusion (BL).[37] These baselines ensure comparison with state-of-the-art approaches in multi-character generation.

Table 1: Comparison of existing work

Paper	*Technologies*	*Novelty & Contribution*	*Strengths*	*Limitations*	*SOTA Methods Tested*
GR-MCI[6]	Motion Graphs, Probabilistic Synthesis, Zero-shot	Generate-and-Rank framework for diverse multi-character scenes.	Coherent, diverse, high-quality interactions.	Limited to two actors, domain-specific data, high computation.	PageRank-inspired algorithms
TaleCrafter[7]	GPT-4 for S2P, Discrete Diffusion for T2L, Fine-tuning	Multi-modal pipeline for consistent story visualization with layout, sketch, and text conditioning.	Novel characters, multi-character consistency, layout control.	Relies on Stable Diffusion, small-face generation struggles, manual sketches needed.	Make-A-Story, Custom Diffusion, Paint-by-Example
Portrait-Booth[13]	Stable Diffusion (v1.5), ECAC, DIP; Fine-tuning	One-shot text-to-portrait generation with identity fidelity.	Robust identity preservation, scalable multi-subject handling.	Limited to faces, computational overhead, dataset bias.	DreamBooth, Textual Inversion, Subject-Diffusion
Instant-Family[28]	Latent Diffusion, ControlNet, Masked Cross-Attention; Zero-shot	Zero-shot multi-ID generation with spatial control.	Scalable multi-ID consistency, introduced new multi-ID metric.	Prone to pose inaccuracies, edge artifacts, identity mixing in complex cases.	FastComposer, InstantID, IP-Adapter
Theater-Gen[8]	GPT-4, CMIGBench Benchmark; Zero-shot	LLM-based prompt book for consistent multi-turn image generation.	Semantic and contextual consistency, training-free framework.	Pre-trained adapters cause inconsistencies, high computation for multi-turn tasks.	Mini DALL·E 3, MiniGPT-5, SEED-LLaMA
Dream-Story[9]	Multi-subject Diffusion; Training-free	MMSA and MMCA modules for multi-subject consistency.	Consistency across frames in multi-subject scenarios.	Struggles in dense multi-character scenes, LLM dependency.	ConsiStory, MuDI, StoryDiffusion
OMG[19]	Concept Noise Blending, Layout Preservation; Training-free	Two-stage framework for occlusion-friendly multi-concept generation.	Superior identity preservation, harmonious illumination.	Small-face generation struggles, computationally intensive.	DreamBooth, InstantID, Mix-of-Show
ConsiStory[31]	SDSA, Feature Injection, cross-attention maps; Training-free	SDSA and feature injection for consistent subjects and diverse layouts.	State-of-the-art subject consistency, 20× speedup.	Struggles with ambiguous styles, separating appearance from style.	IP-Adapter, Textual Inversion, DreamBooth-LoRA

Paper	*Technologies*	*Novelty & Contribution*	*Strengths*	*Limitations*	*Other SOTA Methods Tested*
Character-Adapter[14]	Prompt-Guided Segmentation, Region Adapters; Training-free	Plug-and-play for precise, region-controlled character customization.	High-fidelity generation, reduced concept fusion.	Intricate clothing details, inaccurate localization in complex scenes.	DreamBooth, Mix-of-Show, InstantID
SP-Diffusion[15]	SP-Mask, SP-Attn; Training-free	Region-specific attention refinement for attribute confusion.	Enhanced text-image consistency, semantic alignment.	Struggles with overlapping regions, computational overhead.	Composable Diffusion, Attend-and-Excite, Divide-and-Bind
Story-Maker[21]	ControlNet, LoRA, SDXL; Fine-tuning	Holistic consistency across multi-character scenes.	Consistent faces, clothing, hairstyles, and poses.	Posture anomalies, incoherent scenes with $>$ 3 characters.	MM-Diff, PhotoMaker-V2, InstantID
Uni-Portrait[27]	ID Embedding/ID Routing, DropToken, DropPath; Two-stage training	Unified framework for single- and multi-ID personalization.	High-fidelity identity preservation, flexible layouts.	Limited to faces, no support for attribute-specific or non-human customization.	FastComposer, FlashFace, IP-Adapter
Cog-Cartoon[10]	Character Plugins, Layout-Guided Inference; Fine-tuning	Lightweight (316 KB) character plugins for story visualization.	Minimal training and storage requirements.	Struggles with dynamic narratives, complex layouts.	StoryGANc, Make-A-Story, Custom Diffusion
MuDI[17]	Segment Anything (SAM); Fine-tuning with segmentation	Seg-Mix data augmentation for identity decoupling.	High multi-subject fidelity, effective identity decoupling.	Reduced performance with similar subjects, preprocessing overhead.	DreamBooth, Cut-Mix, Textual Inversion
SEED-STORY[26]	Train-short-test-long strategy, LoRA	StoryStream dataset for coherent long-sequence generation.	Efficient generation of up to 30 sequences.	High computational requirement, annotated dataset dependency.	StoryGen, LDM, SEED-LLaMA
OneActor[20]	ResNet, AdaIN; One-shot tuning	Cluster-conditioned model with semantic interpolation.	Faster tuning, efficient subject consistency.	Struggles with $>$ 4 subjects per image, latent biases.	DreamBooth, Textual Inversion, ConsiStory
Dialogue Director[12]	Multi-View Diffusion Models; Training-free	Framework for dialogue-driven storytelling.	Enhances storytelling with cinematic principles.	Dynamic storyboards and complex poses pose challenges.	OmniGen, MIP-Adapter, StoryMaker
DiffSensei[11]	Manga Pretraining, MangaZero Dataset	MLLM-based identity adapter for manga generation.	Precise layout control, dynamic character adaptation.	Limited flexibility, struggles with style control.	MS-Diffusion, AR-LDM, StoryDiffusion

Paper	*Technologies*	*Novelty & Contribution*	*Strengths*	*Limitations*	*Other SOTA Methods Tested*
Follow-Your-MultiPose[16]	Pose Guidance; Tuning-free	Pose-guided multi-character video generation.	High temporal and spatial consistency.	Struggles with fine-grained interactions, dynamic backgrounds.	ControlVideo, Text2Video-Zero, MasaCtrl
StoryWeaver[22]	Knowledge Graphs, KE-SG	Character-Graph for improved visualization.	Enhanced identity preservation, text alignment.	Computationally intensive, complex character interactions.	StoryGen, IP-Adapter, DreamBooth

6. APPLICATIONS AND USE CASES

The advancements in multi-character image generation have broad applications across storytelling, gaming, and animation. These models enable coherent narratives by maintaining character consistency across scenes, making them invaluable for visual storytelling in books, comics, and interactive media.[38] In gaming, they facilitate dynamic character generation[39] and immersive environments,[40] while in animation, they support the efficient creation of consistent characters across multiple frames.[41] These versatile applications highlight the potential of these methods in creative and commercial domains[42][43].[44]

7. FUTURE DIRECTIONS

The field of multi-character image generation is undergoing transformative advancements, with innovations addressing scalability, consistency, and generalization. After the first notable work in 2014,[45] progress stagnated until 2023, likely due to the computational demands of multi-character generation and the research focus shifting toward Large Language Models (LLMs). The emergence of diffusion models also required time for researchers to understand and innovate within this architecture. However, 2024 has seen significant advancements, signaling renewed interest in the field.

Innovations like TaleCrafter's sketch automation streamline storytelling workflows,[7] while Follow-Your-MultiPose[16] and InstantFamily[28] improve pose and spatial coherence for realistic interactions. Video-based storytelling advancements from DreamStory[9] and ConsiStory[31] address temporal consistency, and physics-based simulations in PortraitBooth[13] and Generating and Ranking Diverse Multi-Character Interactions[6] enhance realism but face computational challenges.

Most research has focused on single-character consistency, with multi-character challenges gaining attention recently due to improved computational power and diffusion model maturity. LLMs now play a vital role in enabling advanced semantic understanding, as seen in DreamStory[9] and Dialogue Director.[12]

Future improvements include incorporating reinforcement learning[46] into frameworks like MuDI[17] and InstantFamily[28] for dynamic adaptation, and self-supervised learning in DiffSensei[11] and StoryWeaver[22] to reduce biases and enhance generalization. Lightweight physics models are needed for real-time simulations, while latent space optimization in OneActor[20] and UniPortrait[27] could maintain fidelity in dense multi-subject scenes.

Existing computer vision techniques could have a transformative impact. Motion prediction models can improve temporal coherence, self-supervised learning enables robust representations, CLIP enhances semantic alignment, and SAM[47] addresses spatial separation in complex scenes.

Many frameworks lack integration, such as combining DreamStory's multimodal anchors with SPDiffusion's semantic refinement,[15] which could address consistency challenges. Computational overhead remains a barrier for methods like PortraitBooth,[13] requiring techniques like knowledge distillation and model pruning[48] for scalability. Expanding to diverse datasets is critical for methods like StoryWeaver,[22] and balancing realism and efficiency in physics simulations is essential.[49]

The advancements in 2024 highlight a growing focus on solving multi-character challenges. Integrating techniques like UniPortrait and SPDiffusion with scalable frameworks such as TaleCrafter and CogCartoon could create adaptive, robust solutions, redefining applications in gaming, virtual reality, and digital storytelling.

8. CONCLUSION

Consistent multi-character image generation remains a challenging yet essential goal in the application of diffusion models, particularly for storytelling, gaming, animation, and personalized content creation. While significant progress has been made through techniques such as iterative refinement, identity clustering, advanced attention mechanisms, and multi-modal alignment, challenges like identity drift, temporal coherence, and computational scalability persist. Addressing these issues requires innovations in memory systems for character recall, improved generalization across styles, and scalable methods for real-time, multi-character interactions. By consolidating current advancements and identifying key limitations, this review provides a comprehensive foundation for future research, paving the way for diffusion models to unlock new possibilities in creative and interactive domains through consistent, visually coherent multi-character outputs.

REFERENCES

[1] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1125–1134 (2017).

[2] Bansal, G., Nawal, A., Chamola, V., and Herencsar, N., "Revolutionizing visuals: the role of generative ai in modern image generation," *ACM Transactions on Multimedia Computing, Communications and Applications* **20**(11), 1–22 (2024).

[3] Ho, J., Jain, A., and Abbeel, P., "Denoising diffusion probabilistic models," *Advances in neural information processing systems* **33**, 6840–6851 (2020).

[4] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I., "Zero-shot text-to-image generation," in [*International conference on machine learning*], 8821–8831, Pmlr (2021).

[5] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K., "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 22500–22510 (2023).

[6] Won, J., Lee, K., O'Sullivan, C., Hodgins, J. K., and Lee, J., "Generating and ranking diverse multi-character interactions," *ACM Transactions on Graphics (TOG)* **33**(6), 1–12 (2014).

[7] Gong, Y., Pang, Y., Cun, X., Xia, M., He, Y., Chen, H., Wang, L., Zhang, Y., Wang, X., Shan, Y., et al., "Talecrafter: Interactive story visualization with multiple characters," *arXiv preprint arXiv:2305.18247* (2023).

[8] Cheng, J., Yin, B., Cai, K., Huang, M., Li, H., He, Y., Lu, X., Li, Y., Li, Y., Cheng, Y., et al., "Theatergen: Character management with llm for consistent multi-turn image generation," *arXiv preprint arXiv:2404.18919* (2024).

[9] He, H., Yang, H., Tuo, Z., Zhou, Y., Wang, Q., Zhang, Y., Liu, Z., Huang, W., Chao, H., and Yin, J., "Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion," *arXiv preprint arXiv:2407.12899* (2024).

[10] Zhu, Z. and Tang, J., "Cogcartoon: towards practical story visualization," *International Journal of Computer Vision* , 1–26 (2024).

[11] Wu, J., Tang, C., Wang, J., Zeng, Y., Li, X., and Tong, Y., "Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation," *arXiv preprint arXiv:2412.07589* (2024).

[12] Zhang, M., Wang, Z., Chen, L., Liu, K., and Lin, J., "Dialogue director: Bridging the gap in dialogue visualization for multimodal storytelling," *arXiv preprint arXiv:2412.20725* (2024).

[13] Peng, X., Zhu, J., Jiang, B., Tai, Y., Luo, D., Zhang, J., Lin, W., Jin, T., Wang, C., and Ji, R., "Portraitbooth: A versatile portrait model for fast identity-preserved personalization," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 27080–27090 (2024).

[14] Ma, Y., Xu, W., Tang, J., Jin, Q., Zhang, R., Zhao, Z., Fan, C., and Hu, Z., "Character-adapter: Prompt-guided region control for high-fidelity character customization," *arXiv preprint arXiv:2406.16537* (2024).

[15] Zhang, Y., Zhang, R., Nie, X., Li, H., Chen, J., Hao, Y., Zhang, X., Liu, L., and Li, L., "Spdiffusion: Semantic protection diffusion for multi-concept text-to-image generation," *arXiv preprint arXiv:2409.01327* (2024).

[16] Zhang, B., Ma, Y., Fu, C., Song, X., Sun, Z., and Li, Z., "Follow-your-multipose: Tuning-free multi-character text-to-video generation via pose guidance," *arXiv preprint arXiv:2412.16495* (2024).

[17] Jang, S., Jo, J., Lee, K., and Hwang, S. J., "Identity decoupling for multi-subject personalization of text-to-image models," *arXiv preprint arXiv:2404.04243* (2024).

[18] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al., "Lora: Low-rank adaptation of large language models.," *ICLR* **1**(2), 3 (2022).

[19] Kong, Z., Zhang, Y., Yang, T., Wang, T., Zhang, K., Wu, B., Chen, G., Liu, W., and Luo, W., "Omg: Occlusion-friendly personalized multi-concept generation in diffusion models," in [*European Conference on Computer Vision*], 253–270, Springer (2024).

[20] Wang, J., Yan, C., Lin, H., Zhang, W., Wang, M., Gong, T., Dai, G., and Sun, H., "Oneactor: Consistent subject generation via cluster-conditioned guidance," *Advances in Neural Information Processing Systems* **37**, 21502–21536 (2024).

[21] Zhou, Z., Li, J., Li, H., Chen, N., and Tang, X., "Storymaker: Towards holistic consistent characters in text-to-image generation," *arXiv preprint arXiv:2409.12576* (2024).

[22] Zhang, J., Tang, J., Zhang, R., Lv, T., and Sun, X., "Storyweaver: A unified world model for knowledge-enhanced story character customization," *arXiv preprint arXiv:2412.07375* (2024).

[23] Gu, Y., Wang, X., Wu, J. Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al., "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models," *Advances in Neural Information Processing Systems* **36**, 15890–15902 (2023).

[24] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H., "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 15954–15964 (2023).

[25] Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., and Chen, Q., "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **38**(5), 4117–4125 (2024).

[26] Yang, S., Ge, Y., Li, Y., Chen, Y., Ge, Y., Shan, Y., and Chen, Y., "Seed-story: Multimodal long story generation with large language model," *arXiv preprint arXiv:2407.08683* (2024).

[27] He, J., Geng, Y., and Bo, L., "Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization," *arXiv preprint arXiv:2408.05939* (2024).

[28] Kim, C., Lee, J., Joung, S., Kim, B., and Baek, Y.-M., "Instantfamily: Masked attention for zero-shot multi-id image generation," *arXiv preprint arXiv:2404.19427* (2024).

[29] Fu, S., *Learning New Dimensions of Human Visual Similarity using Synthetic Data*, PhD thesis, Massachusetts Institute of Technology (2023).

[30] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y., "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718* (2021).

[31] Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., and Atzmon, Y., "Training-free consistent text-to-image generation," *ACM Transactions on Graphics (TOG)* **43**(4), 1–18 (2024).

[32] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems* **30** (2017).

[33] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D., "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618* (2022).

[34] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y., "Multi-concept customization of text-to-image diffusion," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 1931–1941 (2023).

[35] Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W., "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721* (2023).

[36] Xiao, G., Yin, T., Freeman, W. T., Durand, F., and Han, S., "Fastcomposer: Tuning-free multi-subject image generation with localized attention," *International Journal of Computer Vision* , 1–20 (2024).

[37] Li, D., Li, J., and Hoi, S., "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *Advances in Neural Information Processing Systems* **36**, 30146–30166 (2023).

[38] Kapadia, M., Frey, S., Shoulson, A., Sumner, R. W., and Gross, M. H., "Canvas: computer-assisted narrative animation synthesis.," in [*Symposium on computer animation*], 199–209 (2016).

[39] Charles, F., Mead, S. J., and Cavazza, M., "From computer games to interactive stories: interactive storytelling," *The Electronic Library* **20**(2), 103–112 (2002).

[40] Yu, Z., Wu, X., Wang, H., Katsaggelos, A. K., and Ren, J., "Automated adaptive cinematography for user interaction in open world," *IEEE Transactions on Multimedia* **26**, 6178–6190 (2023).

[41] Ma, Y., Xu, W., Zhao, C., Sun, K., Jin, Q., Zhao, Z., Fan, C., and Hu, Z., "Storynizor: Consistent story generation via inter-frame synchronized and shuffled id injection," *arXiv preprint arXiv:2409.19624* (2024).

[42] Ronfard, R., "Univ. grenoble alpes, inria, cnrs, grenoble inp, ljk domain: Perception, cognition, interaction. theme: Interaction & visualization. february 2, 2021," (2021).

[43] Pan, Y., Zhang, R., Wang, J., Chen, N., Qiu, Y., Ding, Y., and Mitchell, K., "Miencap: Performance-based facial animation with live mood dynamics," in [*2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*], 654–655, IEEE (2022).

[44] Smed, J., Skult, N., Skult, P., et al., [*Handbook on interactive storytelling*], John Wiley & Sons (2021).

[45] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial networks," *Communications of the ACM* **63**(11), 139–144 (2020).

[46] Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., and He, X., "Hierarchically structured reinforcement learning for topically coherent visual story generation," in [*Proceedings of the AAAI Conference on Artificial Intelligence*], **33**(01), 8465–8472 (2019).

[47] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., "Segment anything," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 4015–4026 (2023).

[48] Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.-C., and Li, H., "Avatar digitization from a single image for real-time rendering," *ACM Transactions on Graphics (ToG)* **36**(6), 1–14 (2017).

[49] Lin, J., Wang, Z., Jiang, S., Hou, Y., and Jiang, M., "Phys4dgen: A physics-driven framework for controllable and efficient 4d content generation from a single image," *arXiv preprint arXiv:2411.16800* (2024).