Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2025)

Xanadu: Generative Media Pipelines for Immersive Participatory Theater

Chiheb Boussema[1], Naisha Agarwal[2], Camilo Vargas[1], Malcolm Wilson[1], Anthony Doolan[3], Andrew Browning[3], Mira Winick[1,4], Jeff Burke[1,4]

[1]Center for Research, Engineering, Media and Performance (REMAP), University of California Los Angeles
[2]Department of Computer Science, University of California Los Angeles
[3]Office of Advanced Research Computing, University of California Los Angeles
[4]Department of Theater, University of California Los Angeles
{chiheb, naishaa, cjvargas, malcolmkw}@g.ucla.edu, {adoolan, abrowning}@oarc.ucla.edu, mirawinick@ucla.edu, jburke@remap.ucla.edu

Abstract

We present *Xanadu*, a full-scale participatory theater production that integrated generative AI into a live musical. Over two weeks, 500 audience members contributed sketches, movements, and sounds that were transformed in real-time —via pipelines combining vision-language models, diffusion models, and moderation— into images, 3D objects, poetry, and choreography rendered within an extended reality environment. Audience inputs were framed as ritual "offerings" to the Muses, with performers guiding participation and AI serving as interpretive intermediary. A hybrid hosting architecture combined controllable, research-driven models with fast foundation models, enabling 30–60 second generations while preserving stylistic and narrative coherence. We discuss design strategies and trade-offs that supported large-scale, group-level human–AI collaboration. This work contributes practical insights into deploying generative AI in live performance and highlights opportunities for designing AI systems that facilitate collective rather than individual creativity.

Introduction

Live performances sometimes employ techniques to involve audiences, from sensory immersion and sing-alongs to, less often, direct audience participation in crafting the plot or playing characters. Such techniques, which require flexibility, can be difficult to apply with modern theatrical technology.Off-the-shelf video, sound, lighting and control systems typically assume fixed timing, rely on brittle handcrafted cues, and (until recently) depend on pre-rendered rather than real-time media.

On the other hand, the rise of generative AI promises new avenues for participation and audience engagement. Video games, as computationally driven interactive entertainment, have led AI adoption. Large Language Models (LLMs), valued for their ease of integration, multimodal versatility, and broad generalization, have been integrated into video game generation (Maleki and Zhao 2024; Buongiorno et al. 2024) and storytelling-based applications (Kumaran, Rowe, and Lester 2024; Sun et al. 2023). Beyond language models, researchers have also impressively leveraged diffusion models, despite their computational demands and significant engineering efforts to tailor them to specific applications and interactive settings. For example, (Feng et al. 2024) trained a diffusion model to stream AAA-level game scenes responsive to player controls, producing high-quality, infinite-length video in real time. Closer to our work, (Xu et al. 2024) introduced a pipeline that generates explorable 3D game scenes from user sketches within three minutes, extending interaction modes beyond text and enhancing opportunities for unique experience creation.

While these advances are impressive, most remain single-user focused (Lee, Hwang, and Lee 2025), largely LLM-dominated (Maleki and Zhao 2024), and oriented toward primarily digital experiences such as games.Our case study instead explores how generative AI can support novel group-level audience participation in media-rich live performances that use contemporary audio-visual systems, blending physical and digital worlds. Specifically, we developed an immersive, participatory XR staging of the Broadway musical *Xanadu* (Beane, Lynne, and Farrar 2007), presented at UCLA in May 2025 to nearly 500 audience members over a two weeks. Spectators were seated on stage, surrounded by seven thirteen-foot high, movable LED displays – "shrines" celebrating the Greek muses of the story – and invited to participate in the show by making music, drawing sketches, and dancing – interactions the AI transformed into elements of the evolving digital world.

This case study focuses on the generative AI components of the audience-led media synthesis: audience members used a custom WebXR app to sketch on the 13-foot shrines during three key moments of the show. Within a minute, our AI pipelines converted all these sketches into meaningful, design-aligned images and 3D objects that populated a persistent Unreal-rendered world displayed on the large shrines surrounding the audience and playing area.

Although earlier theater productions have used AI, to our knowledge, they were mainly LLM-driven text generation for improvisation (Mathewson and Mirowski 2018) or script writing (Rosa et al. 2020; Mirowski et al. 2022). Few full-length, full-scale productions have integrated generative media (images, 3D, choreography) into a scripted performance.

Our production was made possible by (a) a careful, iterative design of the AI pipelines, employing Vision-Language Models (VLMs) to help interpret often abstract user drawings, and cascading diffusion models in a hybrid hosting architecture balancing quality, controllability, and latency, and

(b) a modular software stack supporting flexible development and large-scale deployment.

We believe *Xanadu* makes for a valuable real-world case where non-experts engaged AI together in a live theater setting, demonstrating scalable collective creativity and group play, and offering lessons for designing AI for groups and publics rather than single users.

Implementation

Audience members were split into seven groups, one per muse (the story's main characters). During three key moments of the show, they were invited to contribute to their muse's shrine, and thus to the stage and virtual world surrounding them, by drawing sketches. They were guided in this by "acolytes", actors who led audience engagement by example and direct interaction. Each of the three interaction moments called for a different creative task: (a) designing a background image for their muse, (b) drawing their muse in a pose of their choice, and (c) creating 3D objects to populate the shrine. In each task, phones acted as "wands", tracking users' spatial movements and translating them into sketch strokes visible on the LED shrines. These sketches served as input for our generative AI pipelines.

Additionally, samples of the resulting media objects were analyzed by a VLM to generate poetry and choreographic choices for subsequent performance segments.

Model Cascade and Hybrid Hosting

Media generation in live theater faces strict constraints of budget, latency, quality, and design alignment. To meet these, we combined vision-language models (VLMs) and diffusion models in a hybrid hosting setup: small research models ran on AWS SageMaker endpoints for controllability, while large state-of-the-art models ran on AWS Bedrock for fast ($<$8s) high-quality inference. Open-source models (mostly $\leq$3.5B parameters) on SageMaker produced fast (10-20s), controllable, medium-quality outputs, which were then refined, via image variation generation, by larger (e.g., 8.1B params) Bedrock-hosted models into high-resolution (e.g., 1024×1024) visually pleasing images with minimal defects. Medium quality refers to lower-resolution (e.g., 512×512) and/or images with visible roughness. This hybrid model cascading allowed us to deploy our pipelines on modest GPUs (1-4 L4s instead of A100s), cutting costs while keeping latency low ($\sim$20s per model call, $<$1 minute end-to-end) for high quality designer-aligned media.

This choice was also motivated by the fact that many controllable generation frameworks target older (and smaller), UNet-based model architectures —which can struggle with overall quality—, while research support for the latest diffusion transformers (Peebles and Xie 2023) is still emerging.

Generative AI Pipelines

Three generative tasks that called for audience creativity were serviced by distinct AI pipelines, all following a similar basic pattern: VLMs helped in creative input interpretation and prompt generation, local diffusion models along with state-of-the-art control techniques generated task-specific media outputs, occasionally enhanced by Bedrock-hosted diffusion models. From audience sketch to shrine manifestation, each generative pipeline took 30–60 seconds, meeting the production's objectives for immersion and agency. A human-in-the-loop moderation layer verified outputs before rendering to prevent substandard or inappropriate content. We expand on each pipeline below; see Fig. 1 for an overview.

Scenes Celebrating the Muses Attendees sketched environments for their muses that would appear in a "frieze" at the bottom of the digital shrine. Show designers provided a per-muse scene reference image to steer color palette and background type. A VLM (Deepseek, (Lu et al. 2024)) interpreted sketches into detailed descriptions that prompted a diffusion model (Stable Diffusion 3.5 Large, SD3.5L, (Stability AI 2024)). The latter was further guided by the sketches via a ControlNet (Zhang, Rao, and Agrawala 2023) and designer-aligned aesthetics via IP-Adapter (Ye et al. 2023) to produce fast, medium-quality images, which served as reference to a Bedrock-hosted model (Amazon Nova Canvas, (AGI et al. 2025)) for fast high-quality variation generation. The muse was then composited at a random location along the bottom axis of the final image.

Note that although SD3.5L is a large, low-resolution generation (512×384) kept computation manageable and inference fast since self-attention scales quadratically with image area.

Texturing Objects with the Muses in Custom Poses & Costumes Audience-drawn poses were combined with designer-selected fabrics to generate clothed muse characters, later applied as textures to shrine objects. Acolytes guided audiences, while muses inspired them with pose suggestions. We designed a multi-agent system, using Claude 3.5 Sonnet (Anthropic PBC 2024), to handle both audience and designer inputs. A first agent described the muse's clothing based on the designer reference image, capturing textures, colors, and motifs. A second agent translated the audience's pose sketches into textual descriptions, which a third agent converted to numerical joint positions that were automatically adjusted to match the proportions of the real actor using a reference photo. Few-shot learning improved pose interpretation and position estimation: we collected images of various poses and generated a set of template pose descriptions (for the second agent), and generated a $\langle$pose description, numerical pose$\rangle$ set (for the third agent). When at times the multi-agent system failed due to irrecoverably bad sketches, we defaulted to a predefined pose library.

For facial identity preservation, we modified InstantID (Wang et al. 2024) to accept open-pose images, adjust them to the actor's proportions, and differentially schedule guidance from the IdentityNet and Pose ControlNet. A Stable Diffusion XL (SDXL)[1] checkpoint then generate a fully clothed, accurately posed muse that retained facial identity and realistic body proportions, dressed in garments aligned with designer aesthetics.

Because both the Pose ControlNet and the IdentityNet op-

[1]All Stable Diffusion XL use cases evoked in this manuscript employed the 1.0 base model without the refiner.

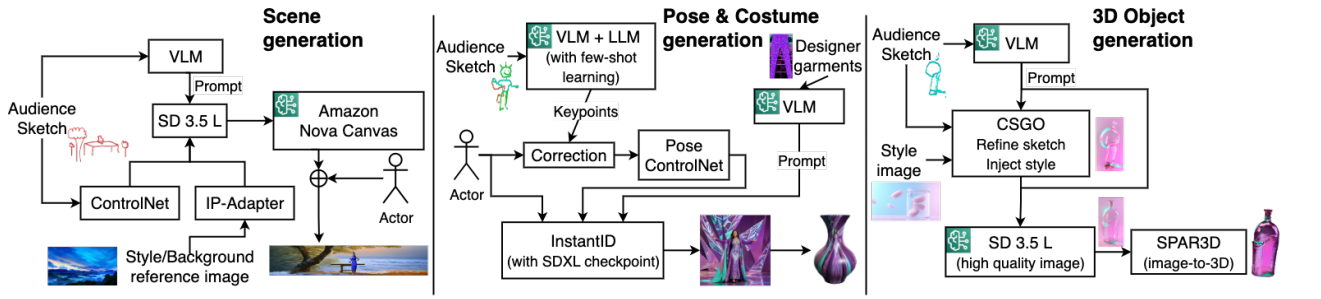

Figure 1: Sketch-based generative AI pipelines. ⊕: simple compositing. Boxes with a green symbol represent Bedrock-hosted models. All other models were hosted within our SageMaker endpoints. See text for additional details.

erate on facial keypoints, facial features were at times deformed. To mitigate this, IdentityNet ran across all inference steps, while Pose ControlNet started at a later step and was active only through mid-generation. This was enough to capture the essence of the pose while minimizing deleterious effects on facial features.

Note that while recent models such as Flux (Labs 2024) can offer excellent quality, text alignment, and ID preservation (Guo et al. 2024), we found that simultaneous pose control and ID preservation to be challenging and inconsistent. Although more recent research (Tao et al. 2025) has shown promise in this direction, it came too late in our development cycle to include. This is in addition to the fact that these models are very large (12B params) and slow, and would have required more costly AWS instances to deploy.

Generating 3D Objects A crowd favorite, audience scribbles were transformed into 3D objects populating the shrines. A VLM (Claude 3.5 Sonnet) first interpreted the sketches to provide guiding prompts. Hosted on a SageMaker endpoint, an SDXL-based model for style-content disentanglement (Xing et al. 2024) combined the content of the sketch with the style of a reference image provided by the show designers to quickly generate an intermediate, more refined sketch image. This image then guided Bedrock-hosted SD3.5L to generate a high-quality image variation. Finally, the latter was converted into a 3D mesh thanks to SPAR3D (Huang et al. 2025), a fast single image-to-3D model. Note that the intermediate refinement step, leveraging the lighter SDXL base, was crucial for aesthetic alignment, outperforming direct scribble-to-SD3.5 while remaining cheap (memory- and latency-wise).

Generating Poems & Dance Moves In addition to media generation, we used AI to generate poetry and choreography GPT-4o (OpenAI et al. 2024) analyzed a random subset of the audience-driven media generated during each evening. It then selected three dance moves from the show's set of moves, and generated a short poem that included dance-related vocabulary. This was interpreted live by an actor, the Oracle, who taught the choreography to the muses and audience. Real-time audience interaction was monitored via computer vision-based pose estimation, where movement accuracy and timing were analyzed by Object Keypoint Similarity and Dynamic Time Warping.

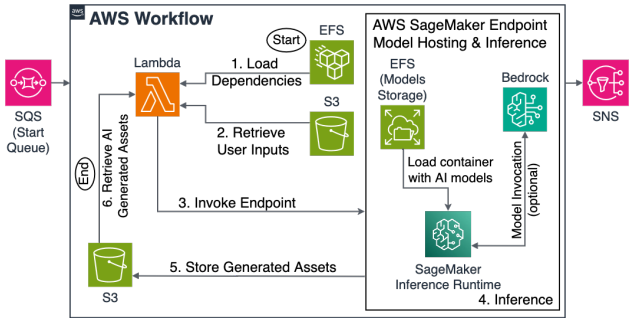

Figure 2: System architecture using AWS Lambda for orchestration, EFS/S3 for storage, and SageMaker endpoints for inference.

System Architecture

The core of our implementation is an event-driven, modular, flexible, and scalable software system architecture.

Audience drawings, made through our custom WebAR app, were sent via Firebase to Unreal Engine and exported to AWS S3 for storage. This triggered AWS Lambda, which coordinated the workflow without the need for dedicated servers. We had orchestrator Lambda functions differentially dedicated to development and rehearsal/production. This allowed us to continue development cycles and explorations while cementing working versions in production pipelines. Incoming sketches, queued via AWS messaging services (SNS and SQS), were routed through AWS Lambda to one of the three media generation modules. Each module in turn invoked its corresponding AWS SageMaker endpoint, where inference was serviced by the endpoint-hosted custom AI models and, where applicable, API calls to Bedrock-hosted vanilla models were made. The outputs were stored back in S3 (with completion messages delivered by SNS), reviewed by humans, then loaded into Unreal Engine and displayed on the shrines. This architecture is summarized in Fig. 2.

Inference requests were handled by 24 SageMaker AI Endpoints, eight for each of the three media generation tasks. We used 8 g6.12xlarge (48 vCPUs, 4x24GB Nvidia L4 GPUs) for the first and third generation tasks, and 16 g6.4xlarge (16 vCPUs, one 24GB Nvidia L4 GPU) instances for the second generation task. Average inference times were 20-30

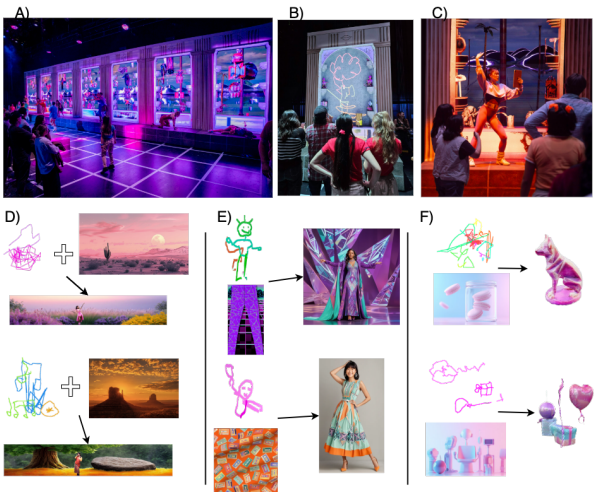

Figure 3: Audience interaction and sample results. A) Shrines populated with 3D assets. B) Audience drawing on a shrine. C) The Oracle performing the generated choreography. D–F) sample audience inputs and output generations.

seconds for the first and third pipelines, and 40-60 seconds for the second pipeline. The AWS Lambda sorted inference requests to match open resources to avoid clogging a small number of endpoints.

Takeaways and Impact

The successful development and deployment of this unique XR experience required addressing a number of human and technical challenges.

Audience engagement had to be solicited without disrupting story or performance flow. This was solved dramaturgically by framing audience drawings as offerings to the show's Muses and Gods, inspired by Greek ritual (Parker 2011). Moreover, having "acolytes", actors who guided audience interactions by example, encouraged organic engagement, learning by imitation, and reduced "performance anxiety" when audiences saw even rough sketches accepted.

Managing non-expert user inputs was another challenge. Audience drawings were often poorly defined. Employing VLMs with interpretive latitude over strict description as an intermediate layer between audience inputs and diffusion models was essential to transform drawings into visually pleasing, design-aligned outputs. A sense of agency arose when audiences recognized their creations, supported by both VLM prompts and sketch-guided diffusion models.

Given the highly curated style and atmosphere of the show, ensuring the stylistic alignment of the generated media was crucial. While prompts included custom terms (e.g., "vaporwave"), style injection was best achieved by directly targeting the denoising process. As the saying goes, "an image is worth a thousand words", designer-selected reference style images provided rich cues for style transfer models, such as IP-Adapter (Ye et al. 2023) and CSGO (Xing et al. 2024), to affect the diffusion model's cross-attention layers.

Another major challenge was the preservation of the actors' identity, in particular facial, when generating images. We explored several models and frameworks, but few were satisfactory. This was complicated by our choice of using older UNet-based models, which tend to struggle with subject consistency. We encourage published research in this field to avoid showcasing results with celebrity faces, as these likely appear in training data and misrepresent true performance. Fine-tuning individual models for character fidelity was infeasible in our case due to time and resource constraints, in addition to initial actors' concerns.

Despite the development efforts to make the AI pipelines robust against bad or inappropriate content, we found that a moderation layer was essential to intercept such cases, infrequent as they were.

Finally, all our development choices were motivated by complex trade-offs among quality, controllability, latency, and budget constraints. This led to our choice of using mainly older, UNet-based diffusion models (e.g., SDXL) in our SageMaker endpoints, since they have a lower memory and latency footprint, yield moderate quality media, and benefit from a large amount of open-source research providing various control capabilities, and community-trained checkpoints offering various aesthetic qualities. Newer models, while impressive, are significantly larger, requiring costlier deployment, and benefit from fewer research expansions. Model cascading within a hybrid hosting architecture allowed us to capture the best of both worlds: controllable generations with older models within SageMaker endpoints, and quick enhancement with newer models hosted on Bedrock.

Together, these points reflect the broader difficulties of integrating AI into live and interactive entertainment, making our work a valuable real-world reference.

Conclusion

To the authors' knowledge, there are very few full-scale productions employing generative AI in live theater, particularly for media synthesis rather than language. As a result, we believe our case study offers a unique perspective. Contributions that we expect to be of interest to readers concerned with practical deployment include: (a) creative generative workflows transforming sketches into production-ready media aligned with the show's artistic direction, (b) practical real-time application of media generation via model cascading and hybrid cloud deployment, (c) modular, scalable software infrastructure, and (d) demonstrating a real-world generative XR experience at scale.

Acknowledgments

We would like to thank all of the contributors to the production, particularly directors Mira Winick and Corey Wright. Full show credits can be found in https://xanadu.remap.ucla.edu. Sponsors providing funding and in-kind equipment included Boxx, Mo-Sys, 4Wall Entertainment, and Amazon MGM Studios' Innovative Storytellers Initiative.

References

AGI, A.; et al. 2025. The Amazon Nova Family of Models: Technical Report and Model Card. arXiv:2506.12103.

Anthropic PBC. 2024. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3.5-sonnet.

Beane, D. C.; Lynne, J.; and Farrar, J. 2007. Xanadu. Directed by Mira Winick and Corey Wright. Stage musical premiered on Broadway at Helen Hayes Theatre, New York, July 10 2007. Music and Lyrics by Jeff Lynne & John Farrar; Book by Douglas Carter Beane.

Buongiorno, S.; Klinkert, L.; Zhuang, Z.; Chawla, T.; and Clark, C. 2024. PANGeA: procedural artificial narrative using generative AI for turn-based, role-playing video games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, 156–166.

Feng, R.; Zhang, H.; Yang, Z.; Xiao, J.; Shu, Z.; Liu, Z.; Zheng, A.; Huang, Y.; Liu, Y.; and Zhang, H. 2024. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*.

Guo, Z.; Wu, Y.; Zhuowei, C.; Zhang, P.; He, Q.; et al. 2024. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37: 36777–36804.

Huang, Z.; Boss, M.; Vasishta, A.; Rehg, J. M.; and Jampani, V. 2025. SPAR3D: Stable Point-Aware Reconstruction of 3D Objects from Single Images. arXiv:2501.04689.

Kumaran, V.; Rowe, J.; and Lester, J. 2024. NARRATIVE-GENIE: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, 76–86.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lee, S.; Hwang, S.; and Lee, K. 2025. Beyond Individual UX: Defining Group Experience (GX) as a New Paradigm for Group-centered AI. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*, 357–362.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Ruan, C. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525.

Maleki, M. F.; and Zhao, R. 2024. Procedural content generation in games: A survey with insights on emerging llm integration. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, 167–178.

Mathewson, K.; and Mirowski, P. 2018. Improbotics: Exploring the imitation game using machine intelligence in improvised theatre. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 14, 59–66.

Mirowski, P. W.; Mathewson, K. W.; Pittman, J.; and Evans, R. 2022. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

OpenAI; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Parker, R. C. 2011. *On Greek Religion*. Cornell University Press.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Rosa, R.; Dusek, O.; Kocmi, T.; Marevcek, D.; Musil, T.; Schmidtov'a, P.; Jurko, D.; Bojar, O.; Hrbek, D.; Kovsvt'ak, D.; Kinsk'a, M.; Dolevzal, J.; and Voseck'a, K. 2020. THEaiTRE: Artificial Intelligence to Write a Theatre Play. In *AI4Narratives@IJCAI*.

Stability AI. 2024. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5.

Sun, Y.; Li, Z.; Fang, K.; Lee, C. H.; and Asadipour, A. 2023. Language as reality: a co-creative storytelling game experience in 1001 nights using generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19, 425–434.

Tao, J.; Zhang, Y.; Wang, Q.; Cheng, Y.; Wang, H.; Bai, X.; Zhou, Z.; Li, R.; Wang, L.; Wang, C.; Lin, Q.; and Lu, Q. 2025. InstantCharacter: Personalize Any Characters with a Scalable Diffusion Transformer Framework. arXiv:2504.12395.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; and Hu, Y. 2024. InstantID: Zero-shot Identity-Preserving Generation in Seconds. arXiv:2401.07519.

Xing, P.; Wang, H.; Sun, Y.; Wang, Q.; Bai, X.; Ai, H.; Huang, R.; and Li, Z. 2024. CSGO: Content-Style Composition in Text-to-Image Generation. arXiv:2408.16766.

Xu, Y.; Ng, Y.; Wang, Y.; Sa, I.; Duan, Y.; Li, Y.; Ji, P.; and Li, H. 2024. Sketch2Scene: Automatic Generation of Interactive 3D Game Scenes from User's Casual Sketches. *arXiv preprint arXiv:2408.04567*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv:2308.06721.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.