Flow Matching for Augmented Reality Applications

Christian Sandor*
Team ARAI
Université Paris-Saclay, CNRS, LISN

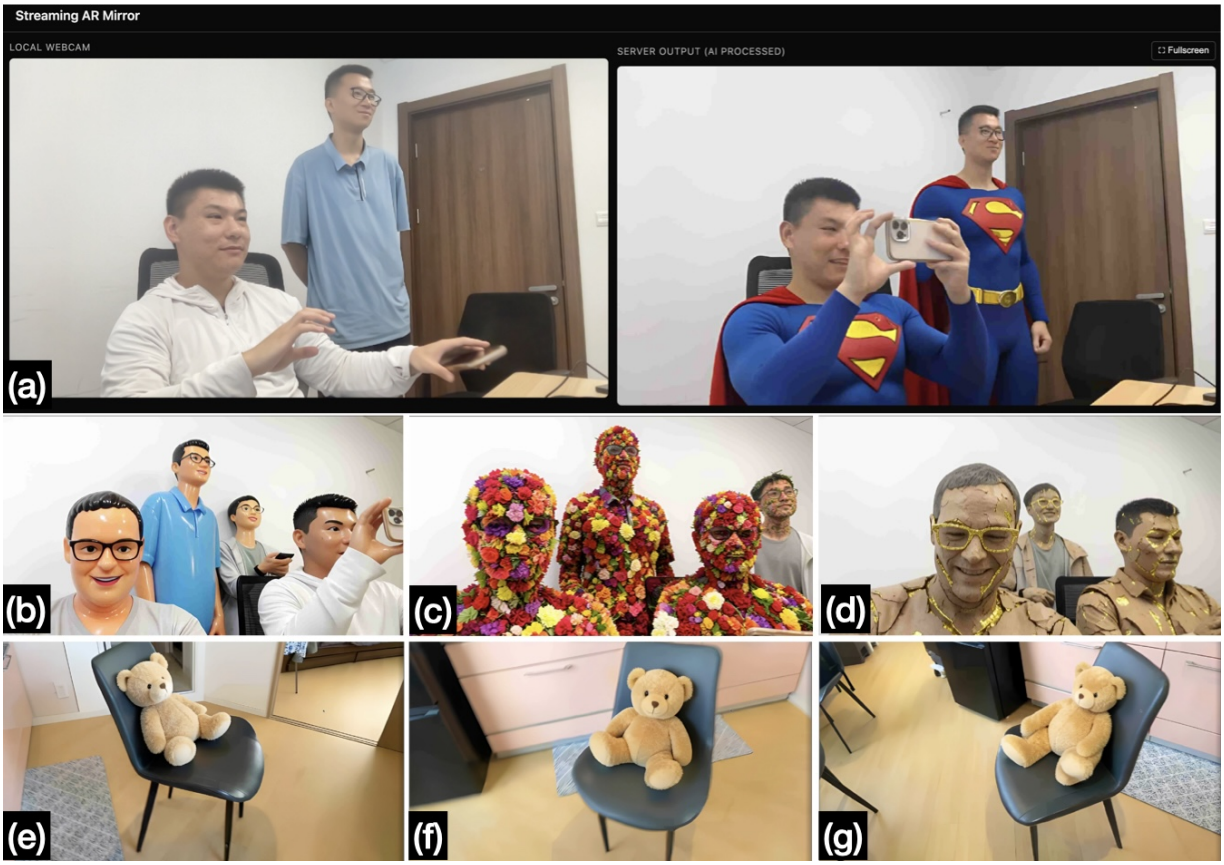


Figure 1: Our demo enables prompt-based editing of a video stream in real-time and in high-fidelity (e.g., 1080p at 46 FPS with 87 ms per-frame pipeline latency). (a–d) modify existing real-world people: (a) input & output pair; example prompts (b) inflatable plastic dolls, (c) the men are made out of flowers, (d) the men are made out of broken pottery with gold-filled seams. (e–g) add a new, spatially registered object for the prompt "Make a teddy sit in the chair".

Abstract

We present a demonstration of real-time, prompt-based video editing for augmented reality (AR). Our system uses a lightweight Flow Matching image generation model and runtime optimizations to transform live camera streams while preserving interactive frame rates. Unlike scene-reconstruction methods, which add virtual objects but cannot modify existing real-world content, our pipeline unifies both operations. At the conference, visitors will experience two setups: a large-screen Magic Mirror that alters their appearance, and AR glasses that modify their egocentric view. At 1080p, the prototype achieves 46 FPS with 87 ms pipeline latency, and supports output resolutions up to 4K.

Index Terms: Augmented Reality, Generative Artificial Intelligence, Flow Matching.

*e-mail: christian.sandor@universite-paris-saclay.fr

1 Motivation

With the recent explosion of Artificial Intelligence, Augmented Reality (AR) and Virtual Reality research has been heavily influenced by it. Gaussian Splatting [4] and Neural Radiance Fields [8] are widely used; in their standard form, they reconstruct a scene and possibly add new virtual objects to it, but cannot modify existing real-world content; a capability that prompt-based generative editing provides. Image (e.g., [1, 10]) and video generation models have also found wide adoption, mainly for offline high-fidelity generation; their underlying architectures require heavy computations that make them basically infeasible for real-time applications. The fundamental question that led to this demo is: *How far can we push real-time Flow Matching to make it usable for a wide range of AR applications?*

2 Related Work

The most similar works are EgoEdit [6], StreamDiffusion (SD1) [5], and StreamDiffusionV2 (SD2) [2]. These modify video streams at high frame rates, but each has a limitation for AR: EgoEdit ignores motion-to-photon latency (almost 1 s, unusable for

AR); SD1, an image model, has severe temporal inconsistency in egocentric AR; and SD2, a video model with impressive speed (42 FPS at 480p, 360 ms latency on 4×H100), produces severe artifacts with high inertia; fast moving objects need up to 1 s to update.

By contrast, our single-GB200 system natively generates 1080p (5× the pixels per frame of SD2's 480p output) at 15 FPS; with ESRGAN upscaling it delivers 46 FPS at 1080p with 87 ms per-frame pipeline latency, higher image quality, and lower inertia than SD2.

3 Demo Description

Our demonstration lets attendees directly experience prompt-based editing of a live video stream: a visitor chooses a text prompt and observes the transformation while moving in front of the camera or viewing the environment through AR glasses. The live nature of both setups makes the tradeoff between editability, visual quality, and response latency directly visible.

Our prototype runs on NVIDIA GB200 GPUs in a data center and streams the result over a wired network; the planned configuration for the conference. At 1080p, it achieves 46 FPS with an 87 ms per-frame pipeline latency (ESRGAN-upscaled from 540p; native 1080p generation reaches 15 FPS at 270 ms per frame; see Table 1); the supplementary video shows detailed measurements for row 1).

Generated Resolution	Display	FPS	Cycle (ms)
540p	1080p	46	87
720p	1440p	25	157
1080p	4K	15	270
1152p	2304p	13	300
1504p	2976p	5	740

Table 1: Performance measurements. FPS is the output (temporally upsampled) frame rate; cycle is per-frame inference time. The 540p and 720p rows use a second GPU for ESRGAN 2× upscaling [11] to the displayed resolution.

We plan to show our system in two different setups at the ISMAR conference:

- **Magic Mirror.** On a large portrait display, visitors see a live, prompt-modified view of themselves, making changes in appearance and response latency immediately visible.
- **AR Glasses.** Wearing AR glasses, a visitor experiences a prompt-modified egocentric view of the environment; this view is mirrored to a separate landscape display so bystanders can share it.

4 Requests to Organizers

We will bring almost all hardware required for the demonstration. We kindly ask the conference organizers to provide us with the following items:

- A high-speed internet connection through an ethernet cable: bandwidth above 10 Mbps and roundtrip latency around 30 ms. Test with `https://speedtest.net` (bandwidth) and `ping lisn.fr` (latency).
- 2 large screens or projectors: one **portrait** display for the Magic Mirror ($\geq 90 \times 160$ cm), and one **landscape** display ($\geq 160 \times 90$ cm) to share the AR-glasses egocentric view with bystanders.

5 Conclusions

This demonstration shows prompt-based generative editing running live in AR at interactive frame rates and measured end-to-end latency, but it relies on data center GPUs and a wired network, so it is not yet suitable for untethered mobile AR.

The two setups have different latency requirements. The Magic Mirror, a non-head-coupled display, is analogous to video conferencing, where ITU-T G.114 recommends one-way delays below 400 ms; our 87 ms is well within range [3]. The AR glasses are head-coupled video see-through; a single latency-matched pipeline avoids the optical see-through misregistration ("swimming"), but their motion-to-photon tolerance is about 20 ms [9], so closing the gap from 87 ms is future work.

As the system uses an image generation model, rapidly moving surfaces show temporal inconsistencies; the demo exposes this tradeoff. Improving temporal coherence and the computational footprint are next steps; a local deployment on non-data-center GPUs is under investigation but not needed for the demo. Future neural-network hardware [7] could eventually integrate such computations into mobile AR.

A key insight this demonstration surfaces is the distinction between adding to and altering the real world. Scene-reconstruction methods such as Gaussian Splatting [4] and Neural Radiance Fields [8] excel at adding new virtual objects to a captured scene and keeping them registered across viewpoints (Figure 1(e–g)), but in their standard form cannot alter the appearance of existing real-world objects (Figure 1(a–d)). Prompt-based generative editing performs both within one pipeline, unifying adding and modifying; something reconstruction-based AR keeps separate.

Acknowledgments

The author thanks Hanxi Zhu, Jia-Hong Liu, and Jun Yang (Tsinghua University) for appearing in the teaser image, un très grand merci à Clément Jambon for a last-minute review, and Alibaba Inc. and Z.ai for releasing the weights of their GLM5.2 and Qwen3.6 models, used here for code generation and pipeline optimization. This work used HPC resources of DALIA (allocation 2026AD010617776, GENCI).

References

[1] P. Esser, S. Kulal, A. Blattmann, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 1

[2] T. Feng, Z. Li, S. Yang, et al. StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation. *arXiv preprint arXiv:2511.07399*, 2025. 1

[3] ITU-T. G.114: One-way transmission time. `https://www.itu.int/rec/T-REC-G.114`, 2003. Recommendation G.114 (05/2003). 2

[4] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 1, 2

[5] A. Kodaira, C. Xu, T. Hazama, et al. Streamdiffusion: A Pipeline-Level Solution for Real-Time Interactive Generation. *arXiv preprint arXiv:2312.12491*, 2023. 1

[6] R. Li, M. Haji-Ali, A. Mirzaei, et al. EgoEdit: Dataset, Real-Time Streaming Model, and Benchmark for Egocentric Video Editing. In *Proc. IEEE/CVF CVPR*, 2026. 1

[7] Ljubisa Bajic. The Path to Ubiquitous AI. `https://taalas.com/the-path-to-ubiquitous-ai/`, 2026. Accessed: 2026-07-13. 2

[8] T. Müller, A. Evans, C. Schied, and A. Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics*, 41(4):102:1–102:15, July 2022. 1, 2

[9] M. Nabiyouni, S. Scerbo, D. A. Bowman, and T. Höllerer. Relative Effects of Real-world and Virtual-World Latency on an Augmented Reality Training Task: An AR Simulation Experiment. *Frontiers in ICT*, 3, 2016. 2

[10] R. Rombach, A. Blattmann, D. Lorenz, et al. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proc. IEEE/CVF CVPR*, 2022. 1

[11] X. Wang, K. Yu, S. Wu, et al. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proc. ECCV Workshops*, 2018. 2