The Role of Diffusion Models in Modern Visual Content Creation

□ **Samuel Teodoro, Munchurl Kim** / KAIST

Abstract

Diffusion models are reshaping visual media creation by turning complex editing workflows into intuitive, text-guided processes. Moving beyond adversarial training, these models generate and modify content through probabilistic denoising, offering greater stability, flexibility, and control. They enable both localized and global edits in images and videos while maintaining temporal and structural coherence. As these systems become more integrated into creative pipelines, they raise ethical challenges involving bias, authenticity, and misuse. Continued development points toward a future where human intention and AI generation work together seamlessly, redefining how visual content is produced, understood, and experienced.

Ⅰ. Introduction

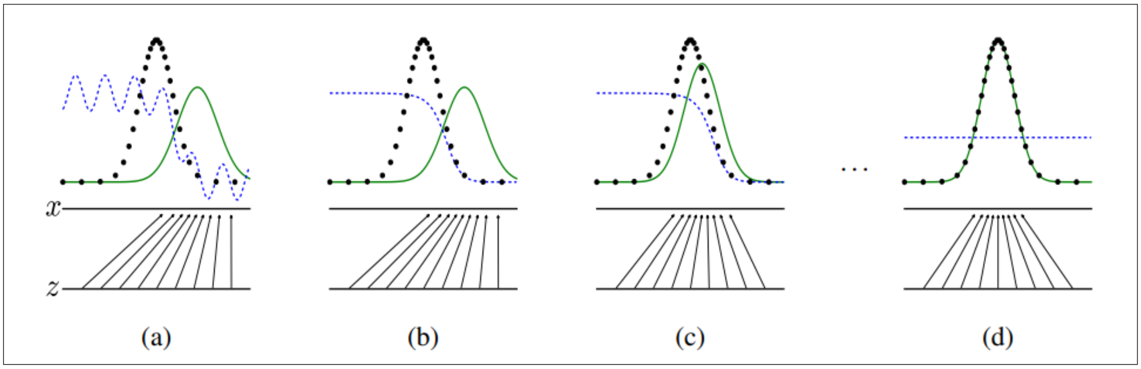
Imagine you have captured a quiet, perfect moment on video: someone close to you sitting by the window, a cup of coffee in hand, warm sunlight spilling across the table. As you replay the video, your eye catches the mug. It is bright and yellow, but you find yourself wishing it were a cool, deep blue instead, something that contrasts with the warmth of the sunlight. You open your editing tool, load the video, pause on the first frame, and type a simple prompt: "make the mug blue." Almost instantly, the change happens. The artificial intelligence (AI) within the tool tracks and edits the mug throughout every second of the clip, preserving its reflections, shadows, and subtle movements. It looks effortless, as if the video had always been that way. No masks, no tracking, no hours spent adjusting frame by frame, just your words, quietly translated into reality.

This transformation is now possible through diffusion models, a breakthrough in AI-based image and video editing. Unlike earlier approaches that required manual adjustments, diffusion models enable editing with remarkable accuracy by gradually adding and removing noise from images and videos. These models can reinterpret an image or video according to the user's intention, modifying only the desired elements while leaving the rest untouched. What once required expert skill can now be accomplished in seconds, opening new creative possibilities for anyone who works with visual media.

This article explores diffusion-based editing for images and videos from multiple aspects. Chapter II explains the core principles behind diffusion models and why they have become a major development in generative AI. Chapter III examines how these models are applied to both images and videos, highlighting the techniques that enable consistent and realistic edits. Chapter IV presents real-world applications in photography, filmmaking, and content creation. Chapter V discusses the ethical and creative questions that this technology raises. Finally, Chapter VI concludes the article by offering a glimpse into the future and considering how diffusion models may continue to shape the way people create, perceive, and interact with visual media.

II. Evolution of Generative Architectures for Image and Video Synthesis

Generative Adversarial Networks (GANs) [1] were among the first deep generative models to produce highly realistic images. They consist of a generator that maps random noise (or conditional inputs) to images and a discriminator that distinguishes generated images from real ones. Through adversarial training, the generator gradually improves so that the discriminator can no longer reliably tell fakes from reals. Although GANs offer



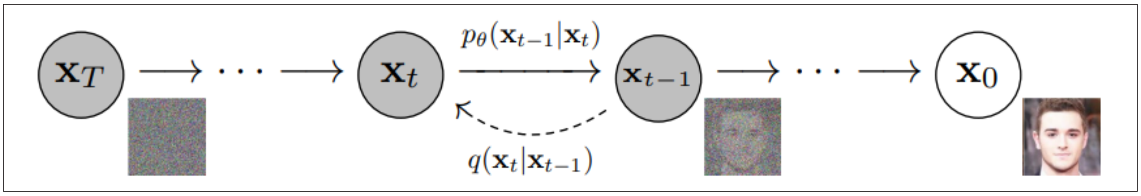
<Figure 1> Overview of the adversarial training process in Generative Adversarial Networks (GANs). The generator (green solid line) gradually learns to produce samples that the discriminator (blue dashed line) can no longer distinguish from real data (black dotted line), achieving convergence when the generated distribution matches the true data distribution. Reproduced from Goodfellow et al., "Generative adversarial networks," NeurIPS, 2014 [1].

fast inference once trained and sharp image quality, they suffer from several fundamental shortcomings: training instability, mode collapse (i.e. producing limited diversity), and difficulties in robust conditioning or control. These drawbacks have motivated the rise of alternative generative frameworks such as diffusion models that trade off some speed for greater stability and flexibility.

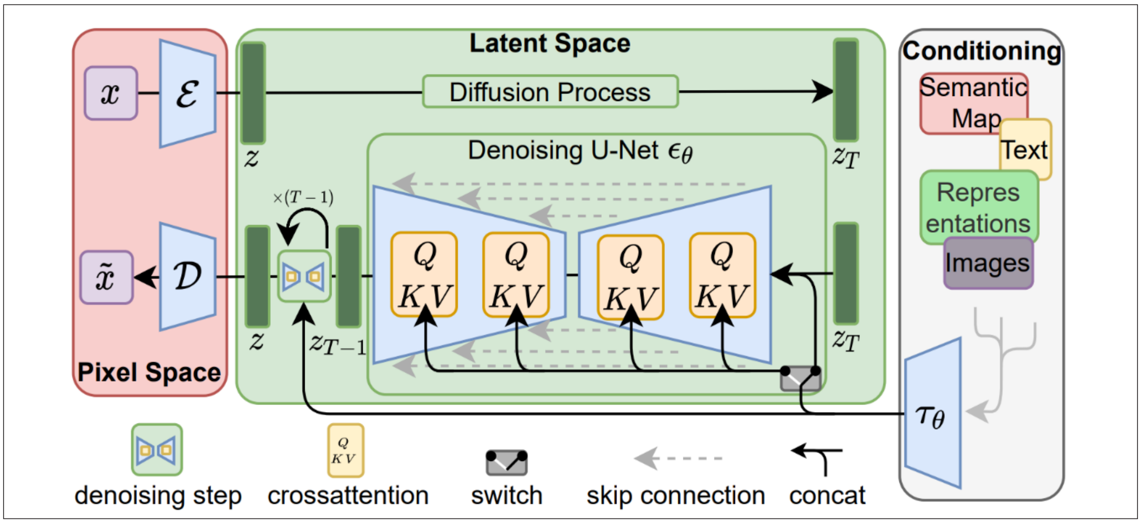
Denoising Diffusion Probabilistic Models (DDPMs) [2] offer a fundamentally different paradigm from GANs by framing generation as a learned denoising process rather than an adversarial game. In DDPMs, shown in Figure 2, a forward process gradually corrupts clean data by adding noise, and the model learns to reverse this process step by step, recovering the original image by predicting the noise at each timestep. The backbone of many DDPM implementations is a U-Net architecture, which takes as input the noisy image plus a time embedding and outputs an estimate of the added noise. Its multi-scale encoder-decoder design with skip connections enables capturing both coarse structure and fine detail. Because DDPMs do not rely on adversarial training, they tend to offer greater stability, reduced mode collapse, and more reliable conditioning (text prompts or other inputs) control compared to GANs.

Although DDPMs can generate high-fidelity images in pixel space, they are computationally expensive, especially when operating at high resolutions or processing videos. Latent Diffusion Models (LDMs) [3] were introduced to overcome this limitation by performing diffusion in a lower -dimensional latent space rather than directly in pixel space. In this approach, an image is first encoded into a latent representation using a variational autoencoder (VAE). The diffusion process, comprising both the noising and denoising steps, then takes place within this latent space, and the result is decoded back into pixel space to reconstruct the final image. This design greatly reduces memory and computational costs while maintaining perceptual quality. Furthermore, LDMs naturally support conditioning mechanisms, such as text prompts through cross-attention layers, allowing more flexible and efficient control over the generation process. Figure 3 shows an overview of the LDM architecture.

To control conditional generation, such as ensuring that a prompt like “make the mug blue” is accurately followed, classifier-free guidance (CFG) [4] is one of the key techniques. CFG involves training a diffusion model both with and without conditioning information, such as text or class la-


$\mathbf{x}_T$
$\mathbf{x}_t$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_{t-1}$
$\mathbf{x}_0$

<Figure 2> Overview of the diffusion process showing the forward (noising) and reverse (denoising) stages. Reproduced from Ho et al., “Denoising diffusion probabilistic models,” NeurIPS, 2020 [2].

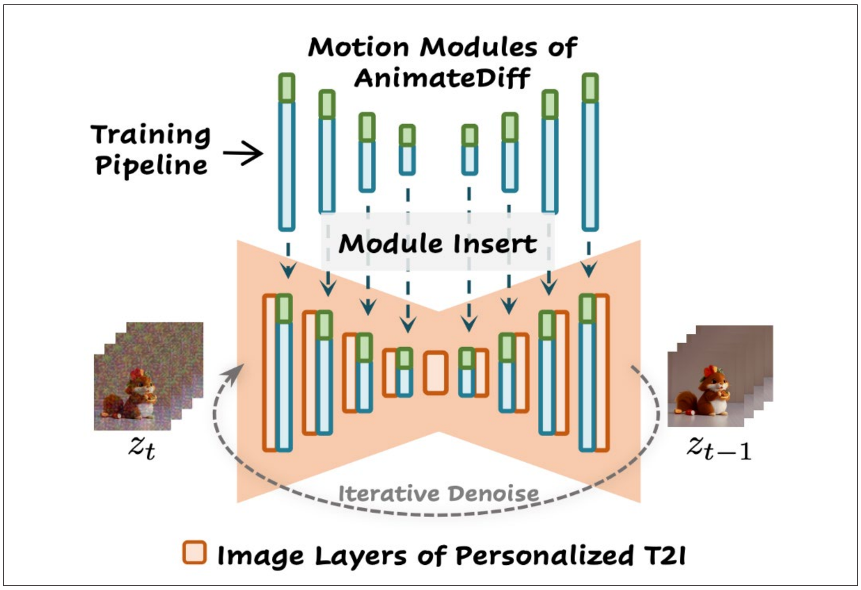


<Figure 3> Overview of the Latent Diffusion Model (LDM) architecture, which performs the diffusion process in a compressed latent space instead of pixel space. Reproduced from Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR, 2022 [3].

bels, by randomly removing the condition during training. During sampling, the model combines the conditional and unconditional predictions through a weighted sum or difference controlled by a guidance scale parameter, which determines how strongly the conditioning influences the output. This approach allows users to balance fidelity to the prompt with diversity of generated results, providing an adjustable level of control over the final output.

Text condition alone is often not sufficient for fine-grained control over spatial layout, edges, depth, pose, or other structural properties. ControlNet [5] introduces a way to augment pretrained text-to-image diffusion models with control signals like edge maps, depth maps, segmentation masks, pose skeletons, etc., via additional control layers while keeping the main model intact. Variants such as ControlNet++ [6] further refine this by explicitly enforcing consistency between the input condition (e.g. depth map) and the output image, via cycle consistency loss. Moreover, Uni-ControlNet [7] allows combining multiple local and global controls in a flexible composable framework under one model.

Researchers have extended LDMs to video synthesis by adding temporal modeling on top of image LDMs. One approach is to pretrain an LDM on images only and then fine-tune or adapt it on sequences of encoded frames to capture motion and temporal consistency. For example, Align Your Latents [8] introduces a temporal dimension into the latent diffusion process and fine-tunes temporal layers over video datasets, converting a pre-trained image LDM into a video LDM. Another example is AnimateDiff [9], shown in Figure 4, which adds a motion module to a frozen image diffusion model to animate images into videos without per-model fine-tuning. These works demonstrate how latent diffusion with temporal layers or motion modules


Motion Modules of AnimateDiff
Training Pipeline
Module Insert
z_t
z_{t-1}
Iterative Denoise
Image Layers of Personalized T2I

<Figure 4> Overview of the AnimateDiff pipeline, which extends a pre-trained image diffusion model with a motion module to synthesize temporally consistent videos. Reproduced from Guo et al., "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning", ICLR, 2024 [9].

provide an efficient path from still-image generation toward full video generation under the same modeling framework.

A growing direction in diffusion research is the replacement of traditional convolutional U-Net denoisers with Transformer-based architectures, known as Diffusion Transformers (DiTs), designed to better capture global context and scale model capacity. For instance, DiT [10] replaces the U-Net backbone with a transformer that processes latent patches, achieving strong results on large-scale image generation benchmarks. CogVideoX [11] extends this idea to video generation, employing a 3D VAE and expert transformer layers to integrate spatial semantics with temporal motion cues. These transformer-based diffusion models excel in long-range attention, global coherence, and flexible conditioning, though they often require greater computational resources and exhibit higher architectural complexity compared to convolution-based approaches.

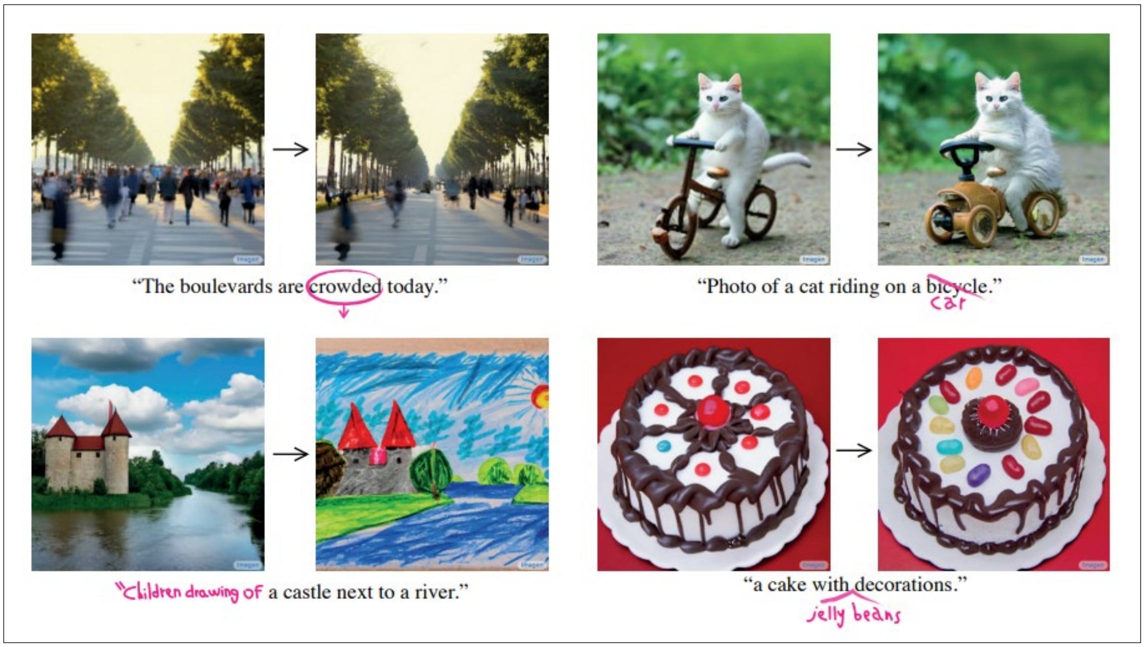
Ⅲ. Diffusion-Based Editing for Images and Videos

Image and video editing, in the context of diffusion-based approaches, involve identifying which visual elements should remain unchanged and which should be modified according to a given prompt or guide. The goal is to generate a new image or video that preserves the essential content and structure of the original while applying the desired edits in a natural and coherent way. These edits can range from low-level changes, such as adjusting colors, textures, or lighting, to high-level modifications, such as altering object identity, expression, or style.

In diffusion-based editing, both image and video tasks can be categorized as local editing or global editing. Local editing focuses on modifying specific areas or objects, such as changing a color, adjusting texture, or replacing an element, while keeping the rest of the scene unchanged. Methods like LIME [12] perform local editing in images by automatically identifying regions of interest using clustering and cross-attention maps, and by applying attention regularization that suppresses unwanted tokens within the target region. In the video domain, VideoGrain [13] modulates spatial and temporal (cross- and self-) attention, to achieve fine-grained control over video content. Global editing, on the other hand, modifies broader characteristics such as scene layout, lighting, or style. InstructPix2Pix [14] exemplifies this approach by following natural language instructions to alter an image globally (for example, “make the image look vintage”). In videos, FRESCO [15] enforces both intra-frame (spatial) and inter-frame (temporal) constraints through correspondence maps and explicit feature updates, enabling coherent global edits, such as style transfer, across all frames without retraining.

Editing with diffusion models generally involves two main stages: inversion and editing. Inversion methods like DDIM inversion [16] encodes a source image or video into a latent representation that preserves its structure and meaning, essentially mapping it into a form the model can later modify. The editing stage then uses this latent representation, along with the user’s prompt, to guide the backward diffusion process and generate the edited result. For videos, the same principle applies but with added challenges: each frame must remain consistent with the others, requiring temporal consistency and motion coherence. Overall, most diffusion-based editing techniques differ in how they perform inversion, how they intervene during denoising, and how they balance preserving the original content with following the editing prompt.

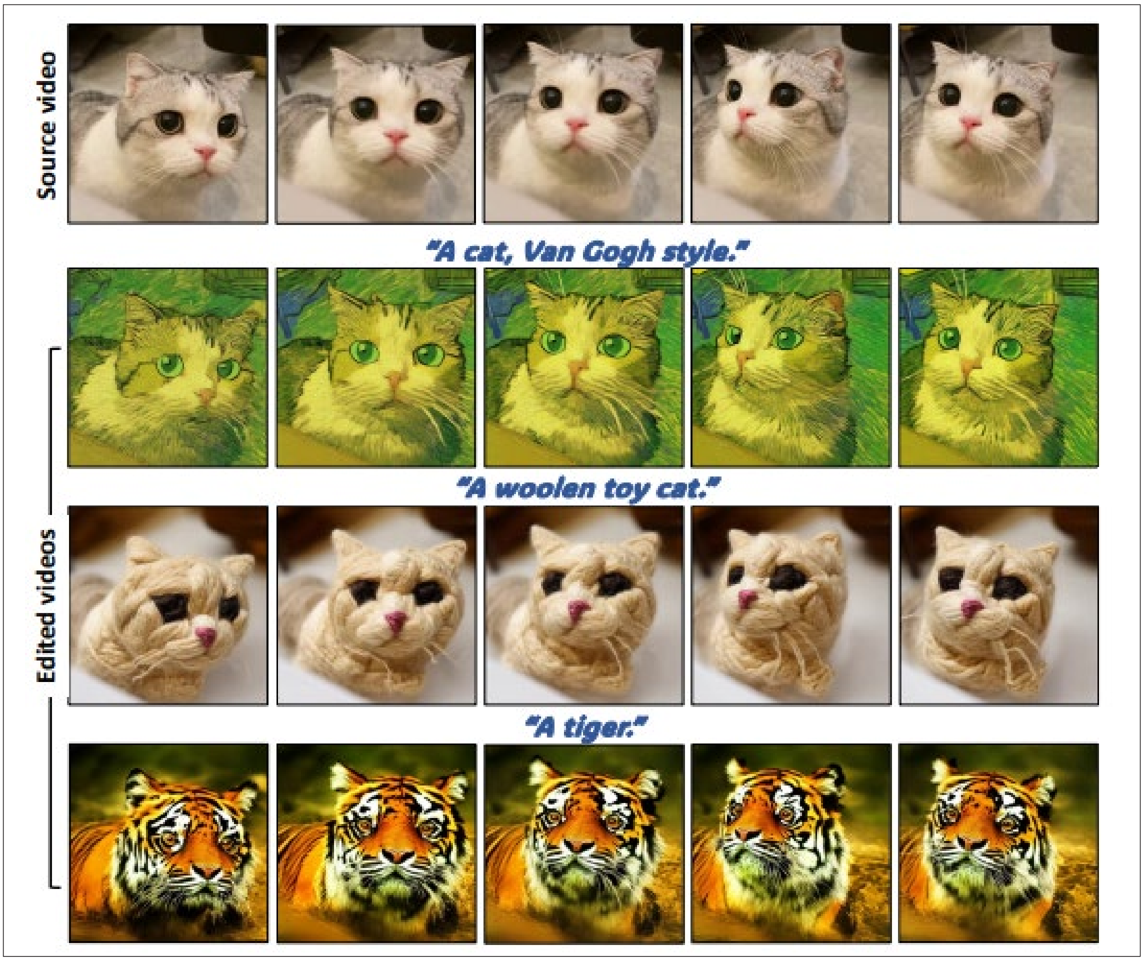
Several algorithms for diffusion-based image editing have already been proposed. These methods are commonly classified as training-based, optimization-based, or attention-based, although other approaches that do not fit neatly into these categories have also emerged. Training-based methods involve explicitly fine-tuning the diffusion model (or adding modules) on a curated editing dataset, allowing the model to learn editing behavior directly. For instance, InsightEdit [17] introduces a pipeline that constructs training pairs of input and edited images, then trains on these image pairs to improve the model’s ability to follow natural language editing instructions. Optimization-based methods approach editing as a test-time optimization problem in the diffusion latent space. For example, Diffusion-Based Conditional Image Editing through Optimized Inference with Guidance [18] adjusts the latent trajectory using representation guidance that combines CLIP score and structural distance losses. Similarly, MAG-Edit [19] refines the noise latent features by maximizing mask-based cross-attention constraints tied to the edit token, achieving localized and precise modifications


"The boulevards are crowded today."
"Photo of a cat riding on a bicycle."
car
"Children drawing of a castle next to a river."
"a cake with decorations."
jelly beans

<Figure 5> Example edits using Prompt-to-Prompt. Reproduced from Hertz et al. "Prompt-to-prompt image editing with cross attention control", ICLR, 2023 [20].

without retraining. Attention-based methods operate primarily through manipulating internal attention maps (cross-attention or self-attention) during inference without full retraining. For example, Prompt-to-Prompt [20] (example results shown in Figure 5) preserves the spatial layout and geometry of a generated image by reusing the attention maps from a source image. This allows users to edit an image simply by changing the text prompt, such as swapping or emphasizing words, while keeping the original structure intact.

Video editing with diffusion models generally falls into three categories based on how temporal consistency is maintained. First, training-based methods usually start from a text-to-image diffusion model and add motion modules that are trained on video datasets. For example, AnimateDiff [9] integrates a motion module into a pre-trained image diffusion model and retrains it on video data to capture temporal dynamics. Second, single-shot or one-shot methods, such as Tune-A-Video [21], adapt a pre-trained image diffusion model to a specific video by fine-tuning spatio-temporal attention blocks. While these methods can achieve high temporal consistency, they are time-consuming because each video requires separate fine-tuning. Lastly, zero-shot methods eliminate the need for additional training or fine-tuning on the input video. For instance, FLATTEN [22] (example results shown in Figure 6) enforces temporal alignment without retraining by using optical flow-guided attention to link similar patches across frames, ensuring spatial and temporal coherence throughout the edited video.

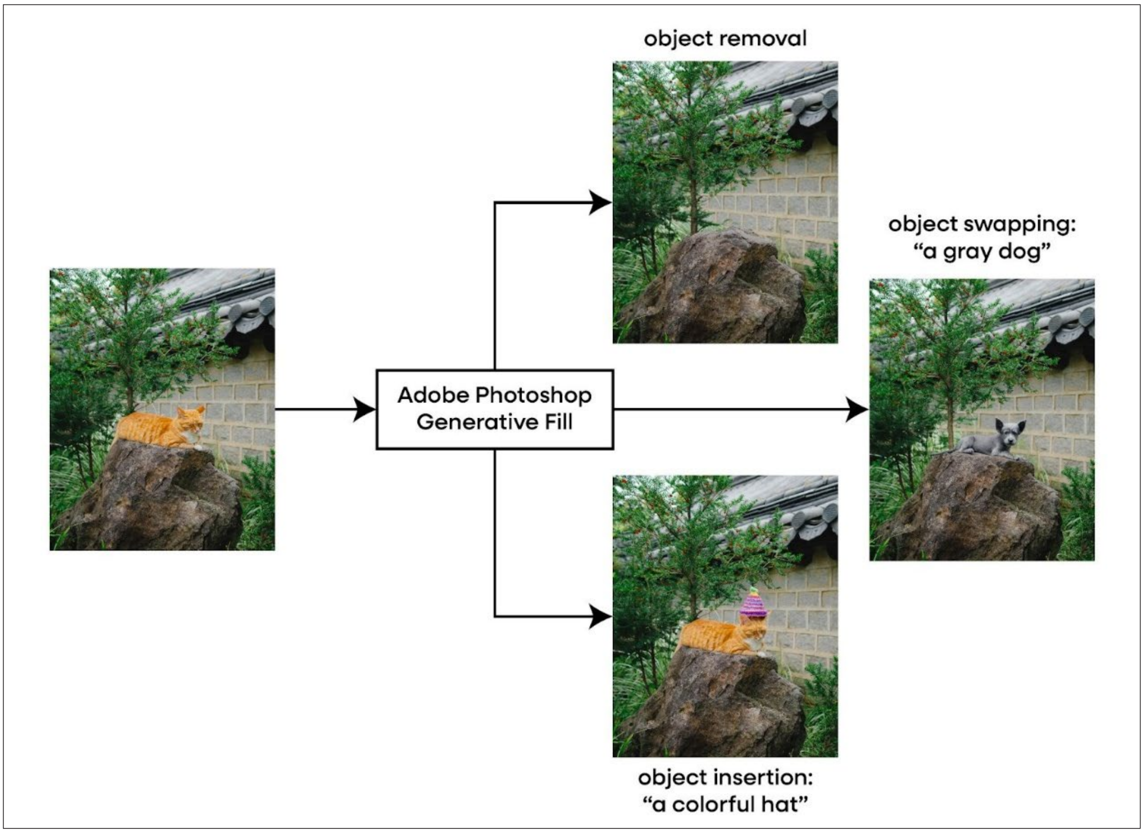


<Figure 6> Sample results from FLATTEN, a zero-shot diffusion-based video editing method that enforces temporal alignment using optical flow-guided attention. Reproduced from Cong et al., "FLATTEN: Optical flow-guided attention for consistent text-to-video editing", ICLR, 2024 [22].

Ⅳ. Diffusion Models in Creative Workflows

1. Photography and Interactive Image Editing

Diffusion models are redefining how photographers and digital creators edit images by turning what were once time-consuming, technical processes into fast and intuitive experiences. A leading example is Adobe Photoshop's Generative Fill (example results shown in Figure 7), which lets users select any region and enter a short text prompt to add, remove, or modify content, such as removing a tree, adding clouds to the sky, or changing an object's color. Another powerful tool is Gemini 2.5 Flash Image or Nano Banana, which supports targeted transformations and fine-grained local edits like erasing an object from the background or adjusting a subject's pose using only simple text com-

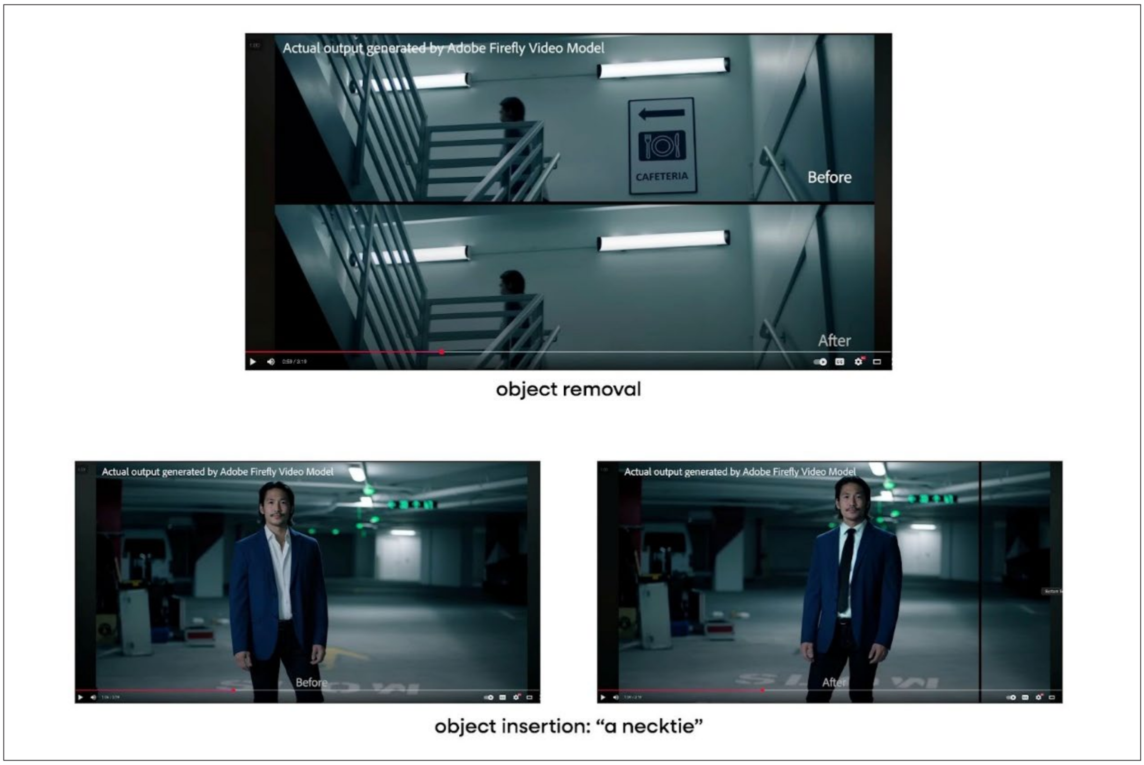

object removal
Adobe Photoshop Generative Fill
object swapping: "a gray dog"
object insertion: "a colorful hat"

<Figure 7> Examples of diffusion-based editing using Adobe Photoshop's Generative Fill, demonstrating object removal, swapping, and insertion guided by text prompts while preserving lighting, texture, and perspective.

mands. These tools perform edits while preserving lighting, texture, and perspective across the surrounding image. For photographers and designers, this means distractions can vanish in seconds, compositions can expand beyond their original borders, and creative ideas can be explored freely without labor-intensive retouching.

2. Video Editing

In the video domain, several commercially available products already support video editing powered by diffusion-based technology. Adobe Premiere Pro (demo shown in Figure 8) has begun integrating generative video tools that use AI to perform tasks such as object addition, object removal, and generative scene extension. Runway's Aleph, a generative video model, allows users to edit existing footage using text prompts, such as adding or removing objects or generating new camera angles, while maintaining coherence across frames. These tools enable creators to prototype cinematic ideas quickly, explore new visual directions, and make continuity adjustments before full post-production.


Actual output generated by Adobe Firefly Video Model
CAFETERIA
Before
After
object removal
Actual output generated by Adobe Firefly Video Model
Before
Actual output generated by Adobe Firefly Video Model
After
object insertion: “a necktie”

<Figure 8> Demonstration of diffusion-based video editing in Adobe Premiere Pro, showing object removal and insertion through generative tools. Screenshots from https://www.youtube.com/watch?v=6de4akFiNYM.

V. Ethical and Societal Implications of Diffusion-Based Editing

Diffusion-based editing is powerful, but it also carries risks, especially when models reproduce or amplify social biases. Studies show that popular diffusion models can reflect gender and racial imbalances beyond what is written in the prompt. For example, Stable Diffusion Exposed: Gender Bias from Prompt to Image [23] found that even neutral prompts often generate images that lean toward masculine traits, exposing bias in the output’s layout and object representation. Another study, DiffLens [24], examines this issue more deeply by identifying the internal components of diffusion models that contribute to biased behavior and suggesting ways to adjust these “bias features” to produce fairer and more balanced outputs. There are also works like InvDiff: Invariant Guidance for Bias Mitigation in Diffusion Models [25] that introduce a lightweight module that help steer the sampling process toward fairer and less biased outputs. These works raise the question: who is responsible for fairness? The model developer, the training dataset curators, or the user?

Authenticity, misuse, and trust are also critical ethical concerns. Diffusion-based editing makes it easier than ever to create realistic yet manipulated media, where subtle changes can distort memory or mislead audiences in journalism, advertising, or political communication. To detect and counter such misuse, works like DIRE [26] develop detectors that distinguish real images from those generated by diffusion models. Another defense approach, DiffusionGuard [27], aims to prevent unauthorized or malicious editing through techniques such as mask augmentation and perturbation loss. These issues extend to questions of consent, copyright, and transparency, raising the debate over whether edited content should be disclosed, particularly in domains where visual accuracy and truthfulness are essential, such as news, documentary production, and legal evidence.

Ⅵ. Conclusion

Diffusion models have already transformed how we generate, edit, and experience visual media and their influence is only expected to grow. In the near future, these tools will become increasingly integrated into creative workflows, enabling real-time, interactive editing where adjustments propagate naturally across images and videos, allowing artists, filmmakers, and everyday users to iterate and experiment at unprecedented speed. Beyond efficiency, diffusion models will reshape how we perceive authenticity, style, and authorship, challenging notions of originality while offering unprecedented creative control. Looking ahead, diffusion models promise a future where human intention and AI capabilities collaborate to redefine storytelling, communication, and the very way we experience visual media.

References

[1] Goodfellow, I., et al., "Generative adversarial networks", NeurIPS, 2014.

[2] Ho, J., et al., "Denoising diffusion probabilistic models", NeurIPS, 2020.

[3] Rombach, R., et al., "High-resolution image synthesis with latent diffusion models", CVPR, 2022.

[4] Ho, J., and Salimans, T., "Classifier-free diffusion guidance", NeurIPS Workshop on Deep Generative Models and Downstream Applications, 2021.

[5] Zhang, L., et al., "Adding conditional control to text-to-image diffusion models", ICCV, 2023.

[6] Li, M., et al., "ControlNet++: Improving conditional controls with efficient consistency feedback", ECCV, 2024.

[7] Zhao, S., et al., "Uni-ControlNet: All-in-one control to text-to-image diffusion models", NeurIPS, 2023.

[8] Blattmann, A., et al., "Align your latents: High-resolution video synthesis with latent diffusion models", CVPR, 2023.

[9] Guo, Y., et al., "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning", ICLR, 2024.

[10] Peebles, W. and Xie, S., "Scalable diffusion models with transformers", ICCV, 2023.

[11] Yang, Z., et al., "CogVideoX: Text-to-video diffusion models with an expert transformer", ICLR, 2025.

[12] Simsar, E., et al., "LIME: Localized image editing via attention regularization in diffusion models", WACV, 2025.

[13] Yang, X., et al., "VideoGrain: Modulating space-time attention for multi-grained video editing", ICLR, 2025.

[14] Brooks, T., et al., "InstructPix2Pix: Learning to follow image editing instructions", CVPR, 2023.

References

[15] Yang, S., et al., “FRESCO: Spatial-temporal correspondence for zero-shot video translation”, CVPR, 2024.
[16] Mokady, R., et al., “Null-text inversion for editing real images using guided diffusion models”, CVPR, 2023.
[17] Xu, Y., et al., “InsightEdit: Towards better instruction following for image editing”, CVPR, 2025.
[18] Lee, H., et al., “Diffusion-based conditional image editing through optimized inference with guidance”, WACV, 2025.
[19] Mao, Q., et al., “MAG-Edit: Localized image editing in complex scenarios via mask-based attention-adjusted guidance”, ACM MM, 2024.
[20] Hertz, A., et al., “Prompt-to-prompt image editing with cross attention control”, ICLR, 2023.
[21] Wu, J., et al., “Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation”, ICCV, 2023.
[22] Cong, Y., et al., “FLATTEN: Optical flow-guided attention for consistent text-to-video editing”, ICLR, 2024.
[23] Wu, Y., et al., “Stable diffusion exposed: Gender bias from prompt to image”, AIES, 2024.
[24] Shi, Y., et al., “Dissecting and mitigating diffusion bias via mechanistic interpretability”, CVPR, 2025.
[25] Hou, M., et al., “InvDiff: Invariant guidance for bias mitigation in diffusion models”, KDD, 2025.
[26] Wang, Z., et al., “DIRE for diffusion-generated image detection”, ICCV, 2023.
[27] Choi, J., et al., “DiffusionGuard: A robust defense against malicious diffusion-based image editing”, ICLR, 2025.

Authors

Samuel Teodoro

- 2018 : BS in Electrical Engineering, University of the Philippines Los Baños, Laguna, Philippines
- 2021 ~ Present : PhD in Electrical Engineering, KAIST, Dajeon, South Korea
- ORCID : https://orcid.org/0009-0001-1893-0551
- Areas of interest : Generative AI, Image Editing, Video Editing

Munchurl Kim

- 1989년 2월 : 공학사, 경북대학교 전자공학과
- 1993년 12월 : 공학석사, Elec. & Comp. Engr., Univ. of Florida, Gainesville, USA
- 1996년 8월 : 공학박사, Elec. & Comp. Engr., Univ. of Florida, Gainesville, USA
- 1997년 1월 ~ 2001년 2월 : 한국전자통신연구원 선임연구원 (실감/방송미디어연구팀 팀장)
- 2000년 1월 ~ 2000년 12월 : ISO/IEC JTC1 SC29 WG11 (MPEG) 한국 대표단 단장
- 2001년 2월 ~ 2009년 2월 : 한국정보통신대학교 공학부 조교수/부교수
- 2009년 3월 ~ 현재 : KAIST 전기및전자공학부 부교수/정교수
- 2021년 3월 ~ 2024년 2월 : KAIST ICT 석좌교수