Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30066

A Review of Human-Centric Generative Image Editing: From Latent Space Control to Interactive Manipulation

Yuezhe Yu

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China
yyzbill1106@gmail.com

Abstract. Over the past ten years, the image editing field has experienced a paradigm evolution driven by pixel-wise control, such as Adobe Photoshop, to generative models. With the rise of Generative Adversarial Networks(GANs) and diffusion models, it is now possible to control images through a higher-level semantic understanding. This core vision aims to bridge the gap between human intention and model performance. To address this challenge, the research has shifted from improving generative quality to editing methods, which puts "human-in-the-loop" at the core. The evolution of control reflects the changes in user communities: from code-based abstract latent space manipulation used by early researchers, to the later natural language-based text-to-image image editing (such as InstructPix2Pix), and finally developed to the direct drag-and-drop interaction represented by DragGAN for creators without background mechanisms. The alternative from technical mechanisms. "model-centered" to "user-centered" means the democratization of content creation tools, implying more focus on human-computer interaction principles in future research. To clearly outline the development of this field, this review categorizes existing methods into three paradigms based on the discrepancy in human control modalities: latent spatial navigation, language-guided manipulation, and direct spatial and structural control. This paper's unique contribution is that, systematically analyzes and reviews groundbreaking research that has been conducted since 2018, based on GANs and diffusion models, focusing on "human-control". This paper reveals the inner revolutionary logic of different types of methods, aiming to provide a unique perspective for understanding the future development trend of controllable generative technology.

Keywords: Controllable Generation, Generative Image Editing, Human-Computer Interaction

1. Introduction

Digital image editing has shifted from traditional pixel-level operations to generative models in the past decade. With GANs and Diffusion Models, visual content can now be manipulated through semantic understanding rather than pixel-level operation. The purpose is to bridge human creativity with the capabilities of generative models for a "what you think is what you get" experience. As image quality has developed, research direction has shifted from chasing realism to providing better

DOI: 10.54254/2755-2721/2025.LD30066

user controllability, which highlights the "human-in-the-loop" and evolution from "model-centric" to "user-centric" methods, democratizing content creation tools for creators without specialized expertise. The evolution of this field demonstrates a clear progression from abstract to concrete and professional to accessible. Control mechanisms evolved from code-based latent space navigation to text-prompt editing and finally, to drag-based interactive editing, such as DragGAN [1], demonstrating how technology has become increasingly accessible to general users.

Therefore, this survey aims to systematically review and analyze the pioneering research since 2018, within the frameworks of GANs and Diffusion Models, focusing on the central theme of "human control". Based on the fundamental differences in human control modalities, we will categorize existing techniques into three core paradigms: Latent Space Navigation and Semantic Discovery; Language-Guided, Instruction-Based Manipulation; Direct Spatial Layout and Interactive Control. Through an analysis of the groundbreaking principles underlying these approaches, this paper offers a unique and insightful perspective into the revolution and future of controllable generative technology.

2. Literature review

Contemporary generative image editing is predicated on two principal architectural frameworks: Generative Adversarial Networks(GANs) and Diffusion Models. GANs use an adversarial mechanism involving a Generator and a Discriminator [2]. The former generates images from a latent space, and the latter distinguishes between authentic and artificial images. Through this adversarial process, the Generator is capable of producing highly realistic images, with its latent space playing a vital role in enabling controllable semantic manipulation. In parallel, Diffusion Models (DDPMs) have emerged, which function by reversing a noise-adding process. They initially incrementally degraded an image into pure noise and trained a neural network to denoise it iteratively, ultimately reconstructing a high-quality image from random noise. Although potent, standard DDPMs are computationally intensive because they operate directly within the high-resolution pixel space. To address this issue, Latent Diffusion Models(LDMs) have been designed, significantly improving efficiency by conducting the diffusion process within a compressed, low-dimensional autoencoder-generated latent space. This method substantially reduces computational costs, rendering high-resolution image synthesis more achievable.

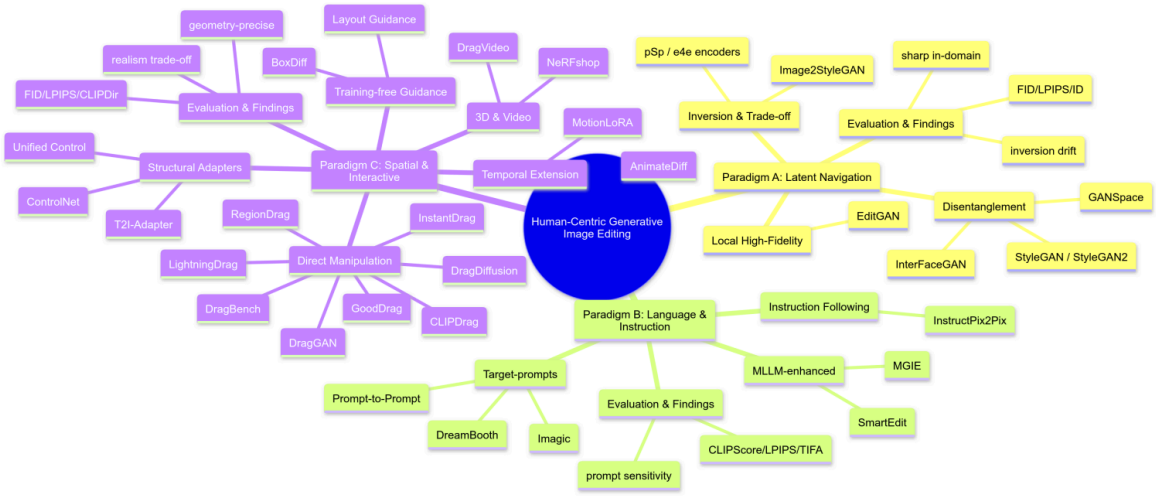

Figure 1. Framework

DOI: 10.54254/2755-2721/2025.LD30066

2.1. GAN latent space manipulation

Gnerative Adversarial Networks (GANs) cannot infer latent codes from existing images, which restrains their use to generate images from random noise. GAN Inversion mapping real images onto the latent space of pretrained GANs to address this problem. This inversion operation links the generator capabilities with real-world data manipulation, enabling image editing, repair, and reconstruction. The development of this field focuses on decoupling and inversion through pioneering work in infrastructure and applications.

2.1.1. Pioneering exploration and paradigm shift

In the generation model field, achieving precise decoupling and control of the generated content attributes remains a core challenge. InfoGAN [3] pioneered controllability force models to learn decoupling by adding extra information during training. By introducing latent code C with explicit meaning into the input noise, the approach demonstrated a correlation by maximizing the mutual information between the codes and image properties, which enabled unsupervised learning. However, the relationship between latent codes and their properties remains unclear, resulting in limited control over the visual trends. StyleGAN [4] uses a network structure design that makes decoupling emerge naturally, marking a turning point in addressing these limitations. Its style-based generator maps the latent vector z into the middle latent space W, which is more linear and "unentangleable", and is suitable for direct semantic editing. Through adaptive instance normalization(AdaIN), vectors in space W enter the network as "style codes", allowing image feature control at different scales, which enhances latent space characteristics, supporting various editing tasks.

2.1.2. GAN inversion and the core trade-off

After StyleGAN demonstrated its semantic generation capabilities, applying it to real image editing became the next focus of academics, leading to GAN Inversion research. Image2StyleGAN [1] pioneered the embedding of real images into the latent space of StyleGAN, establishing a route for applying the editing capabilities of StyleGAN to real images. It introduces an optimization-based inversion method that adjusts the latent code to minimize perceptual loss differences andlocates the code for real images in the W space. However, this study revealed challenges in achieving precise reconstruction while maintaining editability. While improving the image quality, StyleGAN2 [5] examines the expanded W+ Space. It provides independent style vectors for generator layers to expand the latent space, enabling the precise reconstruction of input image details. Research has revealed a fundamental trade-off: while achieving high reconstruction precision, the latent code often deviates from semantically meaningful manifolds, which affects editability. Conversely, codes inverted in Space W, although less precise in reconstruction, are better suited for semantic editing owing to their linear manifold position.

2.1.3. Discovering and applying semantic directions

After GAN inversion enabled real image editing, research focused on discovering controllable semantic editing directions within the latent space. GANSpace [6]presented an unsupervised direction discovery tool that applied Principal Component Analysis to StyleGAN latent vectors to identify principal components corresponding to semantic properties such as age and expression. While unsupervised methods identify the main variations, targeted approaches are necessary for

DOI: 10.54254/2755-2721/2025.LD30066

specific attributes of interest. AdvStyle [7] introduced an adversarial method that requires only a few positive target samples and trains a discriminator to discover the control directions for semantic attributes. The editing directions commonly face semantic "entanglement"—unintended changes to other attributes during the modification. To address this, a study on disentangling the latent space of GANs for semantic face editing [8] presented an unsupervised strategy with penalties for direction decoupling, achieving more independent facial attribute editing.

2.1.4. High-fidelity and application-driven editing

As theoretical refinement continued, research shifted from identifying editing directions to developing practical tools that balance freedom with high-fidelity results. EditGAN [9] illustrates this shift from theoretical exploration to practical tools, achieving unprecedented local editing with rich detail. It introduces a precision-editing framework based on semantic segmentation masks. The system embeds images into the GAN's latent space, allowing users to modify the segmentation masks and optimize the latent code accordingly. Its key innovation is learning editing operations as transferable "editing vectors" for real-time interaction across images. The emergence of EditGAN indicates that GAN-based facial attribute editing technology has matured. GAN-Based Facial Attribute Manipulation [10]provides an authoritative reference summarizing this field's motivations, technical details, model classifications, and future challenges.

By transforming images into a well-structured and comprehensible latent space, one can achieve targeted and increasingly precise semantic editing by manipulating the resulting latent vector.

2.2. Text instruction-based editing

Text-instruction-based image editing represents a significant advance in human-computer creative interaction, enabling the understanding of natural language commands for image alteration. This paradigm evolves from "describing targets" to "instructing processes" to "deep understanding," with key models marking different stages.

2.2.1. Exploring editing via "target prompts"(~2023)

Before genuine "instruction" editing models, comprehensive descriptions of target images by users are required in mainstream text-driven editing. These methods function more like condition-restricted image generation than instruction execution. Representative works in this field reflect this philosophy. Imagic [11], a milestone in the semantic editing of real images demonstrated remarkable capabilities, such as altering animal postures. Nevertheless, explicit descriptions of the desired outcomes is needed to support its workflow. Additionally, in 2023, Prompt-to-Prompt (P2P) [12] is a training-free method that manipulates cross-attention maps, but complete target prompt for modification is required. DreamBooth [13] impacted personalized editing by allowing users to integrate specific subjects into new scenes, with generation guided by prompts describing the complete scene. Although powerful, these models require users to "draw the target" precisely (provide target descriptions). The goal of the model was to "hit the target" (generate this description), essentially performing conditional generation.

2.2.2. A paradigm shift—understanding "transformation instructions" (2023)

The fundamental leap of this field happens when the model can start to understand commands that describe the transformation process itself, which marks the transition from generation tasks to

DOI: 10.54254/2755-2721/2025.LD30066

inference tasks.

The milestone work that pioneered this new paradigm is InstructPix2Pix [14]. Unlike previous work, this work pioneered the instruction-based image editing category, providing a clear definition and solution. Significant advancement in the intelligence needed for the model, which must comprehend verbs, actions, and intentions. So this is an algorithmic reasoning task. To achieve this goal, the paper innovatively utilizes large language models (GPT-3) and diffusion models (Stable Diffusion) to automatically generate triple-tuple large-scale datasets containing 450,000 triplets (input image, editing instruction, output image) for supervised learning.

2.2.3. Deep integration—enhancing comprehension with MLLMs (2024-present)

Although InstructPix2Pix pioneered a new paradigm, the ability to understand complex instructions is still restrained by its reliance on the CLIP text encoder. The researchers soon realized that more powerful multimodal large language models (MLLMs) must be integrated to handle more complicated logic, spatial relations, and world knowledge.

In this new stage, two distinct strategies for MLLM integration emerged. MGIE [15], released by Apple, suggests that brief human instructions (eg, "make the sky bluer") lack information for the diffusion model. It therefore uses an MLLM as an "instruction optimizer," rewriting the simple command into a highly explicit and detailed prompt to better guide the edit. In contrast, SmartEdit [16] is more direct, which explicitly identifies the CLIP [17] encoder as the bottleneck and replacesit completely with an MLLM, which allows the model to directly understand instructions involving complex reasoning, such as edits that require understanding the relationship between the inside and outside of a mirror.

Text-guided image editing technology has quickly developed in just a few years, from generation to inference, and now to cognition. This journey started with the "target prompts" depended on by models like Imagic, to "procedural instructions" that InstructPix2Pix achieves, and has excelled in the "deep semantic understanding" pursued by works like MGIE and SmartEdit through the integration of MLLMs.

2.3. Spatial and conditional control

Text-to-Image generation models allow creating images through text descriptions. However, text prompts alone cannot precisely control spatial composition and object layout because text is a low-dimensional input that is inadequate for describing high-dimensional visual spaces. For example, "a dog to the left of a cat" may result in semantic errors. Research has change direction from style personalization to multimodal control systems to achieve precise control. ControlNet [18] marked a breakthrough by achieving spatial control without compromising the pretrained knowledge, establishing a new paradigm for controllable generation.

2.3.1. Early personalization and style customization (before spatial control)

The core challenge in personalizing diffusion models is balancing the generation quality with the training efficiency and storage costs. Various techniques have addressed this trade-off. Dreambooth [13], known for its high-quality results, fine-tunes the entire model on a small image set, associating subjects with unique identifiers. However, time-consuming training, large model files, and the risk of "catastrophic forgetting" of general knowledge are required to support training quality. Textual Inversion offers a lightweight alternative by freezing the main model and

DOI: 10.54254/2755-2721/2025.LD30066

optimizing only a new "pseudo-word" embedding. Although highly efficient with minimal storage requirements, it provides less control than Dreambooth. Low-Rank Adaptation (LoRA) [19], a Parameter-Efficient Fine-Tuning technique, balances these approaches by freezing the main model while injecting trainable low-rank matrices into key layers. LoRA achieves fast training and small file sizes without significant performance compromise. Hypernetworks extend LoRA by using a smaller network to predict the weights for the main model, allowing flexible concept management, but with more complex training and less stable output.

2.3.2. Spatial control adapters (dominant paradigm)

These methods introduce powerful spatial control capabilities to a frozen foundational model by adding external modules, representing the current dominant solution. A prevalent strategy in modern controllable generation merges external, trainable modules into frozen foundational models for precise spatial control. ControlNet [18], a pioneer in this field, learns new control conditions by creating a trainable "copy" of the encoder blocks from a pretrained model. By reconnecting to the frozen model through"zero-convolution" layers, the original model's knowledge is preserved during training. ControlNet supports various control types, including edge detection (Canny), human poses (OpenPose), and depth maps, with each condition managed by an independent model. As a lightweight alternative, the T2I-Adapter [20] uses a compact adapter network to achieve similar results with fewer parameters, which processes the conditional signal and injects features into a frozen text-to-image model. Its advantage is a faster inference speed, as the adapter runs once during generation. This efficiency is suitable for scenarios in which multiple control conditions are combined.

2.3.3. Unified frameworks and training-free control (exploring efficiency and flexibility)

To overcome the inefficiency of the "one condition, one model" paradigm, subsequent research has explored two primary approaches: consolidating multiple control types into a unified model and eliminating the need for training. The ControlNet [18] framework addresses the scalability issue of training and storing numerous models. It improves deployment efficiency by categorizing control conditions into two types: "local conditions" (like Canny edges or poses) and "global conditions" (such as depth or color palettes) [21]. This enables the handling of various control inputs with only two specialized adapters. A different approach, seen in training-free methods such as Layout Guidance and BoxDiff, bypasses dedicated training. These techniques achieve control by intervening in the internal workings of the model during inference. BoxDiff [22] allows users to define bounding boxes and manipulate cross-attention maps by calculating a loss based on spatial constraints and back-propagating gradients to guide latent updates. This ensures that objects are generated atspecified locations. Although this method offers flexibility, it typically requires more computation and a slower inference speed.

2.3.4. Temporal dimension extension

The concept of spatial control has been successfully extended to video and animation generation, solving the problem of temporal consistency between frames. A seminal framework in this area is AnimateDiff [23]. This framework addresses the challenge of creating animations from personalized T2I models by training an independent and universal "motion modeling module." This pretrained model can be injected into any compatible static T2I model during inference. This

DOI: 10.54254/2755-2721/2025.LD30066

method addresses the problem of generating temporally coherent animations by achieving a decoupling between appearance and motion. Additionally, it introduces the MotionLoRA technique, allowing users to fine-tune and customize the types of motion generated.

2.3.5. Future outlook: from current challenges to a new era of creation

Despite progress in controllable generation, key challenges in achieving seamless composability when combining control signals and balancing the fidelity with the control strength remain. Current methods excel at low-level control (pose, depth, edges), whereas high-level semantic control remains a frontier challenge, including object interactions and emotional expressions. Research has focused on language-driven specifications, using LLMs to convert high-level commands into control maps. Composability and fidelity improve through Interactive Generation, where users instantly update elements. The field advances toward 3D-Native Control using meshes and point clouds for gaming and virtual reality applications. These developments aim to create a Unified Model integratesating text, images, sketches, 3D models, and audio inputs. The text prompt has evolved from "director" to "collaborator," enabling an intuitive multimodal visual creation.

2.4. Direct manipulation & interactive editing

2.4.1. The origin and generalization of generative dragging

The pioneering DragGAN [1] first redefined this "dragging" concept on a GAN's image manifold as an iterative optimizationproblem. Through its core steps of "motion supervision" and "point tracking", it achieved precise control over the generated images. However, its reliance on GANs creates a significant bottleneck: the "GAN inversion" process required to apply it to real-world images often leads to distortion, hindering its practical use.

To overcome this generalization challenge, research focus quickly shifted to more powerful and flexible diffusion models. DragDiffusion [24] was the foundational work in this transition, successfully extending the dragging paradigm to real-image editing. Its core innovations were crucial for this success: it proposed latent code optimization at a single denoising timestep to balance efficiency and effectiveness; it introduced LoRA [19] fine-tuning and a reference latent control technique to solve the critical identity preservation challenge; and it released the first standardized evaluation benchmark, DragBench [24], which established a foundational framework for all subsequent research in the field.

2.4.2. The technical explosion and exploration of core trade-offs

The success of DragDiffusion spurred a wave of subsequent research, primarily revolving around the inherent trade-offs between its three core aspects—fidelity, speed, and controllability—and giving rise to a tension between two major technical routes: "test-time optimization" and "feed-forward inferenceTo enhance fidelity, researchers have addressed the artifact and distortion issues caused by error accumulation during iterative optimization. Works like GoodDrag [25] introduced the "Alternating Dragging and Denoising" framework, which cleverly inserts denoising steps during theoptimization process to maintain image quality.

The demand for speed and real-time interaction has driven a major paradigm shift away from slow optimization methods. To solve this, methods such as LightningDrag [26] and InstantDrag [27] adopted a "feed-forward inference" approach. Instead of optimizing at edit time, they pre-train a

DOI: 10.54254/2755-2721/2025.LD30066

universal network that learns to generate the edited result in a single pass, reducing the editing time from minutes to under a second.

Finally, to improve controllability and resolve the ambiguity of sparse point inputs (e.g., does dragging a mouth corner mean "smile" or "open mouth"?), Researchers have explored richer interaction methods. CLIPDrag [28], for the first time, integrate drag points for precise local controls with text instructions offering global, semantic guidance. Meanwhile, RegionDrag [29] enhanced the interaction element from a "point" to a "region", allowing the model to utilize more contextual information, which significantly reduces operational ambiguity.

2.4.3. Expansion to higher dimensions

Direct manipulation principles are applied to complex data types, indicating future development beyond static images. In 3D scene editing, Neural Radiance Fields (NeRFs) face a hurdle in which the scene geometry and appearance are "entangled" in the network weights. To solve this, frameworks such as NeRFshop [30] reintroduce classic computer graphics primitives. They allow users to select 3D objects via "scribble-based selection" and "sculpt" the implicit scene by dragging vertices of an automatically generated "control mesh." In video editing, drag-based control faces the challenge of maintaining this temporal consistency. Independent frame editing causes a flickering artifact. DragVideo [31]addresses this by optimizing a unified "video latent space," using "mutual self-attention denoising" to ensure smooth editing effects across the video.

3. Discussion

3.1. The core tension between identity preservation and editing freedom

All editing paradigms face an inherent conflict: how to accurately execute users' editing intentions while maintaining the subject's identity. This tension is ubiquitous because this problem is evident in the StyleGAN latent space, as there is still a trade-off between reconstruction fidelity and editability. For instance, to balance identity "drift" during editing, technologies like DragDiffusion [24] are needed to employ additional mechanisms, such as LoRA [19] fine-tuning, which indicatesthat disentangling identity features from other editable attributes remains a significant, open question.

3.2. The cognitive divide between pattern matching and logical reasoning

A significant evolutionary step in this field is the leap from executing "target descriptions" to comprehending "process instructions". InstructPix2Pix [14] marked the first shift from a conditional generation task to an action-reasoning task, enabling the model to understand "what to do" instructions. However, to cross the true cognitive divide—that is, to understand instructions involving complex spatial relationships, common sense, and logic—the semantic matching capabilities of encoders such as CLIP [17] are insufficient. Works such as MGIE [15] and SmartEdit [16] demonstrate that introducing Multimodal Large Language Models (MLLMs) is necessary to overcome this "semantic ceiling". Future editing systems must possess deeper world knowledge and reasoning capabilities, moving beyond mere visual pattern recognition.

3.3. The inversion bottleneck and the multi-axis trade-off space

Nearly all editing techniques for real images are constrained by the fundamental bottleneck of "model inversion". Mapping an arbitrary real image losslessly into a model's latent space is

DOI: 10.54254/2755-2721/2025.LD30066

extremely difficult, and the "fidelity-editability paradox" inherent in this process is the root cause of many other technical trade-offs in this field. The development of this entire domain is not a linear pursuit of a single optimal solution but rather an exploration within a multidimensional trade-off space, seeking the best balance for different application scenarios.

Direct manipulation exposes two recurring trade-offs. On the one hand, optimization-based methods, such as DragDiffusion [24], maximize output quality and editing flexibility but slow down interaction; on the other hand, feedforward inference-based methods, including InstantDrag [32], deliver real-time interaction at the cost of some precision. Similarly, controllability improves under strong spatial guidance—exemplified by ControlNet [18]—yet excessive conditioning tends to reduce naturalness and erode fine detail in the generated images.

Therefore, the value of any new technology underlying complex trade-off space evaluation is essential. Future breakthroughs may emerge not only from enhancing model capabilities but also from entirely new interactive frameworks that enable users to dynamically adjust their trade-off strategies according to their needs.

DOI: 10.54254/2755-2721/2025.LD30066

Table 1. Comparison of the three major paradigms

Paradigm	Core mechanism	Representative methods / models	Key assumptions	Training & data needs	Compute & engineering cost (typical)	Performance advantages	Limitations
Latent space control (GAN latent navigation & inversion)	Manipulate interpretable directions/vectors in GAN latent spaces; optionally invert real images to latent codes for editing.	StyleGAN/StyleGAN2; InterFaceGAN; GANSpace; pSp/e4e encoders; EditGAN.	Pretrained GAN manifold captures semantics of target domain; (near)linear directions exist; good inversion possible.	Often no additionaltraining for edits; requires pretrained GAN; optional encoder training; limited paired data for EditGAN.	Inference fast (feedforward); inversion can be seconds–minutes; commodity GPU (≈8–16 GB) typically sufficient.	High fidelity and sharpness in domains well covered by the GAN (e.g., faces); responsive tuning; interactive sliders possible.	Domain coverage narrow; attribute entanglement; reconstruction–editability tradeoff; poor generalization to arbitrary real images.
Languageguided/ instructionbased (diffusion)	Modify/generated images via text prompts or naturallanguage instructions; rely on crossattention control and/or inversion.	SDEdit (trainingfree); PrompttoPrompt; Nulltext inversion; InstructPix2Pix (instruction triples); PlugandPlay Diffusion.	Text encoder reliably maps intent; diffusion prior preserves realism; for real photos, accurate inversion or instructionfine tune is available.	From none (SDEdit/P2P) to large synthetic triplets (InstructPix2Pix ~454k) for instruction supervision.	Inference slower than GANs; VRAM varies; InstructPix2Pix typical GPU > 18 GB; training optional.	Broad semantic coverage; natural interface; strong realism via diffusion prior; supports both global/local edits.	Lifestyle/product edits; marketing creatives; photoreal “whatif” changes with concise language.
Direct spatial & interactive control (layout/handles/drag)	User supplies explicit structure (edges/pose/boxes) or direct handles (points/regions) to constrain generation; optimization or feedforward.	ControlNet / T2IAdapter / GLIGEN (structureconditioned); DragGAN (GAN); DragDiffusion (diffusion + DragBench).	Structural constraints (maps, boxes) or sparse handles are informative; pretrained diffusion/GAN can follow strong conditioning.	Usually reuse pretrained backbones; ControlNet/T2IAdapter need extra supervised training on (structure, image) pairs; drag methods are trainingfree or light LoRA.	Peredit optimization (Drag*) seconds–tens of seconds; ControlNet inference comparable to SD; DragGAN “few seconds on RTX 3090”.	Precise geometry/layout control; high user agency (“what you see is what you get”); good locality.	Pose retargeting; object relocation; layoutfaithful ads/design; UI/UXoriented creative tools.

4. Conclusion

This paper has systematically reviewed the evolutionary journey of the generative image editing field since 2018, centered on the theme of "human control." This review clearly reveals that the development in this area constitutes a profound paradigm revolution, with control methods progressing along three major paths: latent space navigation, language-guided manipulation, and direct spatial and interactive control. This progression has evolved from abstract code-based operations exclusive to researchers to natural language instructions for a broader creator base and

DOI: 10.54254/2755-2721/2025.LD30066

finally to "what you see is what you get" direct-drag interactions that even novice users can easily master. The essence of this journey is the "democratization" of creative tools, aimed at completely bridging the gap between human creative intent and the generative capabilities of the models.

In tracing this developmental trajectory, evidence indicates that the evolution is not linear but rather a continuous exploration and advancement within several core tradeoff spaces. Fundamental challenges that drive continuous exploration in the field include three conflicts: 1) the inherent conflict between "reconstruction fidelity" and "semantic editability" in GAN inversion, 2) the difficult choice between "interaction speed" and "generation quality" in direct manipulation paradigms, and 3) the balance between "adherence to guidance" and "maintaining realism" under strong conditional controls. It is even more important that recent work observes a decisive cognitive leap: models are shifting from relying on encoders such as CLIP for semantic pattern matching to integrating Multimodal Large Language Models (MLLMs) to achieve deeper logical reasoning and world knowledge understanding. This transition is key to breaking through the current "semantic ceiling" and comprehending complex spatial relationships and deeper user intent.

Looking ahead, the field stands at the threshold of a new era—a move from "Text-to-Image" to an age of truly controllable, intuitive, and multimodal visual content creation. Future editing systems will develop into a "Universal Visual Canvas", enabling the seamless integration of conditional inputs in various forms, such as text, audio, images, and 3D models. In this new approach, the text prompt is no longer the sole "director," but will serve as a "collaborator" alongside various intuitive control modalities. Eventually, the maximum vision for generative models is to evolve from a passive "tool" only executing instructions to an intelligent "creative partner" capable of understanding, collaborating with, and predicting human intent. It truly realizes a seamless "what you think is what you get" creative experience.

References

[1] Pan, X., Chen, C., Liu, S., & Li, B. (2023). Drag your GAN: Interactive point-based manipulation on the generative image manifold. ACM SIGGRAPH 2023 Conference Proceedings , 32 (2), 1–12.

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems , 27 .

[3] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. Advances in Neural Information Processing Systems , 29 .

[4] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 4401–4410.

[5] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 8110–8119.

[6] Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). GANSpace: Discovering interpretable GAN controls. Advances in Neural Information Processing Systems , 33 , 9841–9850.

[7] Terayama, K., Iwata, H., & Sakuma, J. (2021). AdvStyle: Adversarial style search for style-mixing GANs. Proceedings of the AAAI Conference on Artificial Intelligence , 35 (3), 2636–2644.

[8] Chen, X., Zirui, W., Bing-Kun, L., & Chang-Jie, F. (2023). Disentangling the latent space of GANs for semantic face editing. Journal of Image and Graphics , 28 (8), 2411–2422.

[9] Ling, H., Liu, S., & Le, T. (2021). EditGAN: High-precision semantic image editing. Advances in Neural Information Processing Systems , 34 , 16491–16503.

[10] Wang, Z., Chen, K., & Li, C. (2023). GAN-based facial attribute manipulation. arXiv preprint arXiv: 2303.01428

[11] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Tarcai, N., ... & Irani, M. (2023). Imagic: Text-based real image editing with diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 6007–6017.

DOI: 10.54254/2755-2721/2025.LD30066

[12] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross-attention control. arXiv preprint arXiv: 2208.01626 .
[13] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2255–2265.
[14] Brooks, T., Holynski, A., & Efros, AA (2023). InstructPix2Pix: Learning to follow image editing instructions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , 18392–18402.
[15] Cui, Y., Wu, Z., Xu, C., Li, C., & Yu, J. (2024). MGIE: MLLM-guided image editing. arXiv preprint arXiv: 2312.13558
[16] Huang, Y., He, Y., Chen, Z., Yuan, Z., Li, J., & Wu, J. (2024). SmartEdit: A multi-modal language model for instruction-based image editing. arXiv preprint arXiv: 2404.08749 .
[17] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the International Conference on Machine Learning .
[18] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. Proceedings of the IEEE/CVF International Conference on Computer Vision , 3836–3847.
[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv: 2106.09685 .
[20] Mou, C., Wang, X., Xie, L., Zhang, J., Zhao, Z., & Zhou, M. (2023). T2I-Adapter: Learning adapters to inject human craftsmanship in text-to-image models. arXiv preprint arXiv: 2302.08453 .
[21] Zhao, Z., Zhang, J., & Zhou, M. (2024). Uni-ControlNet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv: 2305.16322 .
[22] Xie, Z., Zhang, H., Wang, Z., Huang, Z., Wang, Z., & Li, M. (2023). BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion guidance. arXiv preprint arXiv: 2307.10816 .
[23] Gao, X., Zhang, Y., Zhang, R., Han, X., Chen, W., Liu, Y., ... & Kwok, JT (2024). AnimateDiff: Animate your personalized text-to-image models without specific tuning. arXiv preprint arXiv: 2307.04725 .
[24] Shi, K., Yin, H., Wang, Z., Zhang, S., Yang, K., Wang, Z., & Chen, T. (2023). DragDiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv: 2306.14435
[25] Yin, Z., Liang, Z., Cui, Z., Liu, S., & Zhang, C. (2023). GoodDrag: Towards good drag-style image manipulation. arXiv preprint arXiv: 2312.15342 .
[26] Xie, Z., Zhang, H., Wang, Z., Huang, Z., Wang, Z., & Li, M. (2023). BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion guidance. arXiv preprint arXiv: 2307.10816 .
[27] Xie, W., Jiang, Z., Li, Z., Zhang, J., & Zhang, Y. (2024). InstantDrag: Fast and high-fidelity drag-style image editing. arXiv preprint arXiv: 2405.05346 .
[28] Li, S., Zhang, C., Xu, Y., & Chen, Q. (2023). CLIP-Driven Image Editing via Interactive Dragging. arXiv preprint arXiv: 2307.02035.
[29] Xu, J., Fang, J., Liu, X., & Song, L. (2023). RegionDrag: Precise Region-Based Interactive Image Manipulation with Diffusion Models. arXiv preprint arXiv: 2310.12345.
[30] Lyu, Z., Zhang, Z., Wu, J., & Xu, K. (2023). NeRFshop: Interactive editing of neural radiance fields. ACM Transactions on Graphics (TOG) , 42 (6), 1–16.
[31] Wang, Z., Lin, J., Shi, Y., & Zhou, B. (2023). DragVideo: Interactive Point-based Manipulation on Video Diffusion Models. arXiv preprint arXiv: 2311.18834.
[32] Xie, W., Jiang, Z., Li, Z., Zhang, J., & Zhang, Y. (2024). InstantDrag: Fast and high-fidelity drag-style image editing. arXiv preprint arXiv: 2405.05346 .