FlowInOne: Unifying Multimodal Generation as Image-in, Image-out Flow Matching

Junchao Yi[1*], Rui Zhao[3*], Jiahao Tang[2*], Weixian Lei[3], Linjie Li[5], Qisheng Su[4], Zhengyuan Yang[5], Lijuan Wang[5], Xiaofeng Zhu[1✉], and Alex Jinpeng Wang[2✉]

[1]University of Electronic Science and Technology of China, Sichuan, China
[2]Central South University, Hunan, China
[3]National University of Singapore, Singapore
[4]University of Science and Technology of China, Anhui, China
[5]Microsoft, China
seanzhuxf@gmail.com, jinpengwang@csu.edu.cn

Abstract. Multimodal generation has long been dominated by text-driven pipelines where language dictates vision but cannot reason or create within it. We challenge this paradigm by asking whether all modalities, including textual descriptions, spatial layouts, and editing instructions, can be unified into a single visual representation. We present **FlowInOne**, a framework that reformulates multimodal generation as a purely visual flow, converting all inputs into visual prompts and enabling a clean image-in, image-out pipeline governed by a single flow matching model. This vision-centric formulation naturally eliminates cross-modal alignment bottlenecks, noise scheduling, and task-specific architectural branches, unifying text-to-image generation, layout-guided editing, and visual instruction following under one coherent paradigm. To support this, we introduce **VisPrompt-5M**, a large-scale dataset of 5 million visual prompt pairs spanning diverse tasks including physics-aware force dynamics and trajectory prediction, alongside **VP-Bench**, a rigorously curated benchmark assessing instruction faithfulness, spatial precision, visual realism, and content consistency. Extensive experiments show that FlowInOne achieves state-of-the-art performance across unified generation tasks, outperforming open-source models and rivaling competitive commercial systems. These results establish FlowInOne as a strong foundation for fully vision-centric generative modeling, where perception and creation are unified within a single continuous visual space. Code is available at csu-jpg.github.io/FlowInOne.github.io.

Keywords: Image Generation · Flow Matching · Vision-centric · Visual Prompting · Instruction Following

1 Introduction

Multimodal generation has long operated under a text-dominant assumption: language encodes intent, and vision executes it. Models such as diffusion- and

*Equal contribution, ✉Corresponding author.

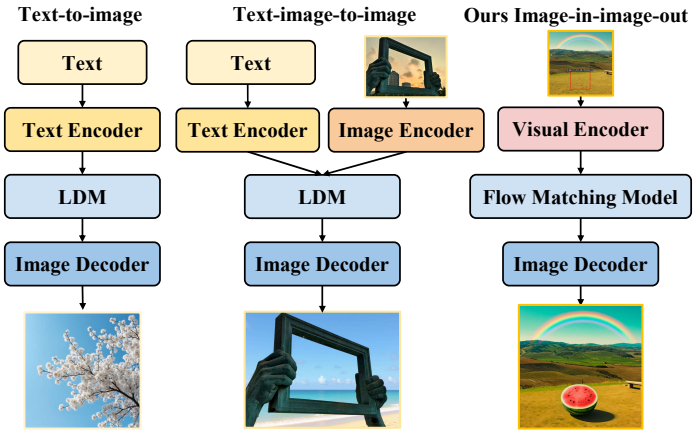

Fig. 1: Comparison of generation paradigms. Left: Traditional T2I only uses the text encoder to condition the Latent Diffusion Model(LDM); Middle: Traditional TI2I requires the joint conditioning of both the text and image encoders; Right: We unify the conditions as visual input and form a simple image-in, image-out framework.

transformer-based text-to-image systems rely on linguistic embeddings as the central conditioning source. While this design achieves impressive visual fidelity, it introduces a fundamental asymmetry: *language controls vision, but vision cannot reason or generate on its own*. This fragmentation of representation space makes it inherently difficult to unify understanding, editing, and generation within a single coherent model.

Recently, a growing trend of vision-centric models has emerged [49, 50, 62, 64], suggesting that textual information can be processed in a purely visual manner by rendering text into pixel space.

These studies collectively suggest that the visual modality itself is expressive enough to serve as the foundation for multimodal understanding. Together, these studies show that representing language visually enables unified perception and alignment within a single modality. Yet they remain fundamentally *perception-oriented*, leaving the generative potential of this vision-first formulation largely unexplored. This raises an important question: **can we build a large model that both reasons and generates entirely within the visual space?**

Flow matching [29] provides a principled answer to this question. Compared with diffusion, it directly learns the underlying velocity field of transformation, offering higher sampling efficiency and stable optimization. By learning visual flows instead of stochastic noise removal, it connects perception and generation under a single deterministic principle.

In this work, we take a decisive step toward this goal and introduce FlowInOne, a framework that redefines multimodal generation as a *purely visual flow*. In FlowInOne, text, layout, and instruction inputs are first transformed into visual prompts, forming the input image state. The model then learns a continuous

transport process that evolves this state into the target visual output using flow matching. This formulation enables a simple and general training pipeline that eliminates noise scheduling, diffusion sampling, and task-specific condition heads.

As shown in Figure 1, FlowInOne departs from the traditional text-conditioned pipeline. Conventional text-to-image models use text encoders (e.g., Flan-T5 [11]) to condition a latent diffusion model, while text-image-to-image setups require two encoders for joint conditioning. In contrast, FlowInOne unifies all input conditions as visual prompts, forming a simple image-in, image-out pipeline with a single model. This design not only simplifies architecture but also ensures consistent alignment between semantic content and spatial control across diverse tasks such as text-to-image generation and visual instruction following.

To support this unified paradigm, we construct VisPrompt-5M, a large-scale Visual Prompt Dataset that spans text-in-image generation, versatile visual editing and physics-aware instruction following.

Each sample pairs a visual prompt canvas with its corresponding target image, providing supervision as continuous visual evolution without task-specific modules or auxiliary channels. We further introduce VP-Bench, a carefully curated evaluation benchmark that assesses model performance across four dimensions: instruction faithfulness, content consistency, visual realism, and spatial precision.

Our main contributions are as follows. *i.* We reformulate multimodal generation into a **vision-centric image-in, image-out paradigm**, eliminating the text encoders and modality-specific bridges. *ii.* We propose FlowInOne, a unified flow matching framework that models multimodal transformation as continuous visual evolution within a shared latent space. *iii.* We build VisPrompt-5M, a comprehensive dataset of visual prompts that enables unified training and strong generalization across text-to-image, image-to-image, and instruction-guided generation tasks. Extensive experiments demonstrate that FlowInOne achieves state-of-the-art performance across unified generation, precise image editing, and physics-aware instruction following, establishing a new foundation for fully vision-centric generative modeling.

2 Related Works

Diffusion and Flow Matching. While diffusion models, from DDPM [23, 52] to LDM [47] and DiT [41], dominate image generation via progressive denoising, Flow Matching [17, 29, 33, 68] learns a continuous transport map between distributions. This approach reduces reliance on complex noise schedules while enhancing sampling efficiency and stability [16]. Building on this, FlowInOne directly models the continuous evolution within a shared latent space, entirely eliminating additional conditioning or noise injection.

Text- and Image-Conditioned Generation. Current T2I models [8, 42, 46, 54, 72] typically inject discrete text tokens via cross-attention [13, 43], leaving control signals disjointed from the visual space. Similarly, conventional image-to-image translation [30, 33, 38, 63, 71], restoration [35, 55], and controllable editing [5, 7, 21, 27, 40, 65, 69] rely heavily on adversarial frameworks [37, 73], diffusion

priors [36], or external control channels. A common limitation across these methods is the dependence on explicit masks or task-specific interfaces for geometric and semantic control. In contrast, FlowInOne entirely bypasses specialized conditioning branches. By rendering heterogeneous constraints—such as text and arrows, we converge multiple forms of control into a single image input, natively aligning semantics and geometry within the visual domain.

Modal mapping. While standard diffusion models map discrete text to images across divergent modalities [2, 6, 8, 10, 14, 25, 53, 59, 67, 70], inherent modality gaps, tokenization artifacts, and reliance on Gaussian noise severely limit spatial precision [24]. Moving beyond generic image-to-image translation [30, 33, 71], we frame generation fundamentally as an **intra-modal transport problem**. By encoding both the visually-instructed input and the target image into a shared, isomorphic latent space, we learn a direct, noise-free flow between them. This pure single-modality mapping resolves structural mismatches at the latent level, demonstrating exceptional scalability across diverse editing and generation tasks.

3 Dataset and Benchmark

In this section, we detail the construction of the **VisPrompt-5M** dataset (Sec. 3.1) and our evaluation benchmark, **VP-Bench** (Sec. 3.2).

3.1 VisPrompt-5M

Illustrated in Figure 2, VisPrompt-5M enables a unified image-in, image-out paradigm. Training pairs $(I_v, I^\star)$ embed all textual and spatial instructions directly into the input canvas I_v. Eliminating auxiliary text channels mitigates ambiguity and enforces strict geometric alignment, empowering a single model to handle diverse tasks within one modality.

Fundamental Generation. We render textual prompts from text-to-image-2M [74] and class labels from an 860K high-quality ImageNet [48] subset directly as input canvases to match their corresponding target images.

Text-in-Image Editing. This category unifies diverse editing and condition-to-image tasks by overlaying textual instructions directly onto the input image canvas. Drawing from GPT-Image-Edit [58], Pico-Banana [44], and UnicEdit [66], we filter out complex or inconsistent pairs to retain approximately 1.6M examples, spanning a diverse range of edits including additions, deletions, and attribute/environment changes. Furthermore, we seamlessly integrate 315K structured image pairs from PixWizard [28], which encompasses a broad spectrum of tasks including Canny-to-image, depth-to-image, inpainting, and image restoration. In total, this yields nearly 1.9M curated pairs where all operational intents are explicitly embedded as visual text prompts.

Text Bounding Box Editing. For precise object insertion with explicit geometric constraints, we extract a high-quality subset of 45K examples from GPT-Image-Edit. Leveraging the combined priors of Qwen Image Edit [60] and

Fig. 2: VisPrompt-5M is a comprehensive dataset that comprises eight distinct data types, including class-to-image generation, text-to-image generation, text-in-image editing, text bounding box editing, visual marker editing, doodles editing, force understanding, and trajectory understanding. The dataset covers a wide spectrum of image-to-image generation, ranging from basic text-in-image generation to compositional editing, and further to physics-aware instruction following.

Qwen3-VL [4] to guide I2I generation, we synthesize pairs where text and bounding boxes jointly dictate the target category, scale, and location. Automated filtering via Qwen3-VL 7B ultimately retains 24K high-quality pairs.

Visual Marker Editing. Leveraging visual understanding capabilities of Qwen3-VL, we generate 250K pairs where arrow annotations serve as salient cues without requiring explicit object names in instruction. This subset supports operations such as deletion, replacement, implicitly representing semantics and spatial relationships through visual markers.

Doodles Editing. Using a two-stage synthesis pipeline with Qwen Image Edit [60] on 5K images crawled from the web, we first add doodles to create input images, then transform them into photo-realistic objects to form target images. Rigorous manual inspection mitigates generative instability, yielding 1.1K high-quality pairs where doodle lines explicitly serve as shape priors.

Force & Trajectory Understanding. VisPrompt-5M supports physics-aware generation. For motion trajectories, we manually annotated the Blender-rendered videos of car and ball movements, yielding 1.5K strictly curated image pairs. For force understanding, we leverage the Force Prompting dataset [18] that covers aerodynamics, oscillations, and linear motion. By extracting keyframes and superimposing text and arrows to denote precise force magnitude and direction, we explicitly visualize object dynamics within the input image.

Across the entire pipeline, we enforce standardized data formats while strictly maintaining task metadata and geometric attributes. We also incorporate automated quality control measures via MLLMs to ensure semantic consistency, visual fidelity, and visual text readability.

3.2 Benchmark and Evaluation

To evaluate our pure Image-in, Image-out paradigm, we curate **VP-Bench**, a comprehensive and manually filtered benchmark covering diverse visual instructions (details provided in Appendix B).

Since conventional metrics like FID [22] struggle with complex visual instructions, we follow recent studies [26, 32] by adopting VLMs as our primary evaluators. A generation is deemed successful only if it simultaneously satisfies four criteria: **(1) Instruction Faithfulness**, **(2) Content Consistency**, **(3) Visual Realism**, and **(4) Spatial Precision**. Alongside the VLM assessment, we conduct rigorous human evaluation based on these exact same criteria to compute the overall pass rate. Since standard VLMs may miss implicit constraints rendered on the image canvas, we manually extract the textual instructions and supply them as supplementary text prompts to ensure fair assessment (refer to Appendix F for VLM evaluation prompts).

To comprehensively evaluate visual quality and editing accuracy, we additionally tailor four quantitative metrics to our paradigm: (1) **CLIP-IQA** [56] to measure overall visual realism; (2) **CLIP Score** [57] to evaluate semantic alignment, for which we directly extract the rendered text instructions from the input image and compute their similarity with the generated image; (3) **Directional CLIP Similarity** [15] to assess semantic consistency in marker-based editing, utilizing an MLLM to generate captions for both the input and generated images, and subsequently computing the directional alignment between the image transition and the corresponding caption pairs; and (4) **DINOv3 Directional Similarity** (DINOv3 Sim) to accurately capture fine-grained spatial and physical structural changes, by directly computing the cosine similarity of the edit displacement vectors among the input, generated, and ground-truth images within the dense DINOv3 [51] feature space.

4 Method

In this section, we first briefly review the preliminaries of Flow Matching (Sec. 4.1) and introduce our core strategy for encoding visual instructions into a unified

visual semantic space (Sec. 4.2). Finally, we present the detailed architecture of FlowInOne (Sec. 4.3), featuring a novel Dual-Path Spatially-Adaptive Modulation mechanism to balance structural preservation and instruction adherence.

4.1 Preliminaries: Flow Matching

Flow Matching (FM) [29, 33, 34] formulates generative modeling as a continuous transport from a source distribution p_0 to a target distribution p_1 over $t \in [0, 1]$. Unlike traditional diffusion models [23, 52], FM does not rely on complex noise scheduling and permits non-Gaussian source distributions, provided they are isomorphic to the target. During training, FM constructs a differentiable probability path z_t between a sample pair (z_0, z_1), which directly yields the ground-truth velocity $v_t^\star$ for supervision:

$$z_t = tz_1 + (1 - (1 - \sigma_{\min})t)\, z_0, \quad v_t^\star = \frac{\partial z_t}{\partial t} = z_1 - (1 - \sigma_{\min})z_0 \tag{1}$$

The network learns a time-dependent velocity field $v_\theta(z_t, t)$ by minimizing the Mean Squared Error (MSE) against $v_t^\star$. In our FlowInOne framework, we explicitly define a non-Gaussian source distribution: z_0 represents the latent state of the unified visual instruction extracted via a visual encoder and text-image VAE, while z_1 represents the target image latent. Since both latents are isomorphic within a shared space, inference is intuitively performed by solving the Ordinary Differential Equation (ODE) $\frac{dz_t}{dt} = v_\theta(z_t, t)$ from $t = 0$ to $t = 1$, deterministically evolving the visual instruction into the final target image.

4.2 Unified processing of text in visual semantic space

The inherent heterogeneity between discrete linguistic symbols and continuous visual textures poses significant alignment challenges in flow matching. To alleviate this, we propose a paradigm shift: rendering textual instructions and diverse visual cues directly onto the image canvas. This explicitly preserves spatial layouts and structural priors without relying on complex cross-modal alignment modules.

By treating text as an integral part of the visual geometry, we circumvent the semantic fragmentation typically introduced by textual tokenizers. To extract robust representations from this unified image I_v, we leverage the visual encoder of Janus-Pro-1B [9]. The input is processed by a SigLIP Vision Transformer to extract patch-level semantic features, which are then mapped into the target embedding space via an MLP projector:

$$X_{\text{fuse}} = \text{MLP}(\text{SigLIP}(I_v)) \in \mathbb{R}^{N \times D} \tag{2}$$

where N is the number of patches and D denotes the embedding dimension. This sequence, X_{fuse}, encapsulates both textual semantics and visual geometry.

To perform continuous flow matching, we map the unified visual tokens to a source latent space via a text-image VAE. Rather than predicting it directly, it parameterizes a distribution to sample the source state $Z_{TI} \sim \mathcal{N}(\bar{\mu}_{z_0}, \text{diag}(\bar{\sigma}^2_{z_0})) \in$

$\mathbb{R}^{H\times W\times C}$. Symmetrically, a frozen image VAE encodes the target image into an isomorphic latent Z_I. Our generative process is thus elegantly formulated as modeling the time-dependent velocity field that continuously transports the source latent Z_{TI} to the target image latent Z_I within the shared latent space.

4.3 FlowInOne

Dual-Path Spatially-Adaptive Modulation. Within the unified Flow Matching framework proposed in this work, we formulate image generation as a deterministic trajectory evolution from a starting distribution p_0 to a target distribution p_1. However, due to the information compression inherent in the visual encoding stage, the initial latent Z_{TI} often fails to fully capture the fine-grained structural features of the source image I_{src}. To address this, we introduce a Dual-Path Spatially-Adaptive Modulation mechanism. This mechanism is designed to dynamically compensate for the missing structural manifold while switching computational paths based on the specific task type. As shown in Figure 3, let $\mathbf{H}^{(l)} \in \mathbb{R}^{N\times D}$ denote the hidden state of the l-th Transformer layer, where N is the token sequence length and D is the feature dimension. This state is first updated via a self-attention $\mathcal{A}_{self}$ to capture the global contextual information:

$$\tilde{\mathbf{H}}^{(l)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \mathcal{A}_{self}(\mathbf{H}^{(l)})) \tag{3}$$

Upon obtaining the features enhanced by self-attention, the model follows a dual-path conditional branch defined by the two different tasks. For text-to-image generation in visual input, where no external structural prior needs to be maintained, the model bypasses the cross-attention layer to prevent the introduction of irrelevant noise. This ensures that the generation trajectory strictly follows the evolution dictated by the text semantics. Conversely, for image editing involving a source image, we employ an Image VAE Encoder to map I_{src} into the latent space, resulting in the latent $\mathbf{z}_{src} \in \mathbb{R}^{H\times W\times C}$. Since $\mathbf{z}_{src}$ captures the visual structure of the image, we reshape it into a reference sequence $\mathbf{S} \in \mathbb{R}^{N\times C}$, where $N = H \times W$. Subsequently, the structural increment $\Delta\mathbf{H}_{struct}$ is computed via a cross-attention mechanism, where the intermediate state $\tilde{\mathbf{H}}^{(l)}$ serves as the Query and the reference sequence $\mathbf{S}$ acts as the Key-Value pair:

$$\Delta\mathbf{H}_{struct} = \text{Softmax}\left(\frac{(\tilde{\mathbf{H}}^{(l)}\mathbf{W}_Q)(\mathbf{S}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)(\mathbf{S}\mathbf{W}_V) \tag{4}$$

To achieve an optimal trade-off between source image fidelity and instruction-following alignment, we design a lightweight adaptive gating network. By concatenating the current denoising state with the extracted source information, the network predicts an anisotropic token-level weight vector $\boldsymbol{\Lambda} \in [0, 1]^N$. This design enables the model to identify spatial heterogeneity at a pixel-level granularity—specifically, distinguishing between regions belonging to the background manifold that require strict preservation and regions targeted for editing that

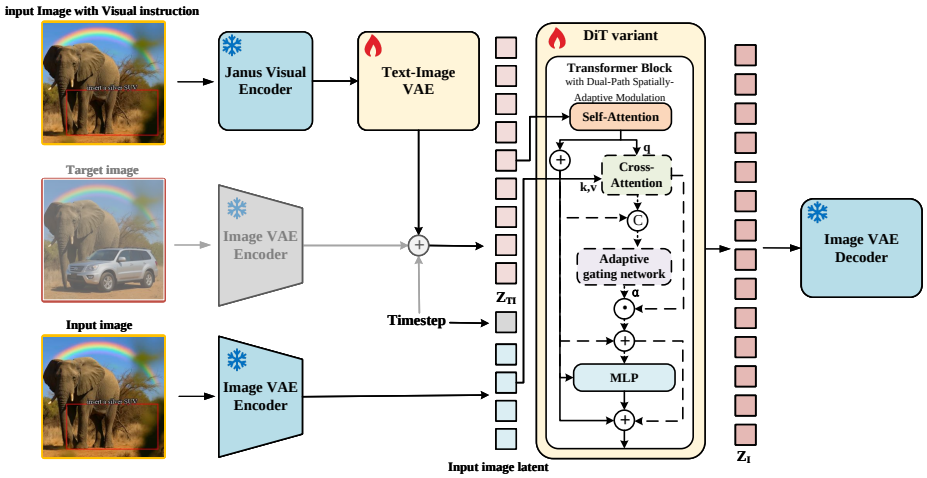

Fig. 3: Overview of the FlowInOne architecture, a general and simple framework using flow matching for continuous evolution in only one modality. FlowInOne employs a Dual-Path Spatially-Adaptive Modulation to adapt computation by modality. For input image rendering with only text, the structural branch is bypassed to strictly follow semantic evolution. Conversely, for image editing, a spatially-adaptive gated network and cross attention activates to selectively inject source priors, dynamically balancing original image preservation with instruction-driven reconstruction.

require reconstruction. The derivation of the gating coefficient matrix is as follows:

$$\boldsymbol{\Lambda} = \sigma\left(\mathrm{MLP}_\theta\left([\tilde{\mathbf{H}}^{(l)} \parallel \Delta\mathbf{H}_{struct}]\right)\right) \tag{5}$$

where $\parallel$ denotes feature concatenation along the channel axis and σ represents the Sigmoid activation function. The final layer output $\mathbf{H}_{out}^{(l)}$ is integrated via a conditional formulation controlled by a task indicator $\mathbb{I}_{edit}$:

$$\mathbf{H}_{out}^{(l)} = \tilde{\mathbf{H}}^{(l)} + \mathbb{I}_{edit} \cdot (\boldsymbol{\Lambda} \odot \Delta\mathbf{H}_{struct}) \tag{6}$$

Here, $\mathbb{I}_{edit} \in \{0, 1\}$ is a binary indicator: for pure text-to-image inputs, $\mathbb{I}_{edit} = 0$, and the modulation term is nullified to sever structural dependencies; for inputs containing visual instruction and source images, $\mathbb{I}_{edit} = 1$, activating the spatially-adaptive refinement. By precisely controlling the infiltration of the structural manifold via $\boldsymbol{\Lambda}$, this method effectively mitigates editing conflicts caused by over-preservation. Furthermore, by leveraging explicit structural compensation, it reduces the evolution error of Flow Matching in complex image editing scenarios.

Flowing in the unified modality. We model visual instruction images to target images as a flow matching process in a shared latent space. Given the instruction sequence $X \in R^{N\times C}$ obtained from a unified visual encoder, a variational

posterior is defined using a Variational Autoencoder.

$$q_\psi(z \mid X) = \mathcal{N}\big(\bar{\mu}_X,\ \mathrm{diag}\big(\bar{\sigma}_X^2\big)\big)\,, \\ z_0 \sim q_\psi(z \mid X), \quad z_0 \in \mathbb{R}^{N\times D}. \tag{7}$$

Image-side training and sampling are conducted in a latent space: the pre-trained and frozen VAE from LDM [47] $\mathrm{Enc}_{\mathrm{img}}$ is used to map the target image $I^\star$ to the target latent variable.

$$z_1 = \mathrm{Enc}_{\mathrm{img}}(I^\star) \in \mathbb{R}^{N\times D}. \tag{8}$$

During training, vanilla flow matching is adopted: time is sampled from $t \sim \mathcal{U}(0,1)$ and linear interpolation is constructed in Equation 1. The instantaneous velocity is represented by the vector field $v_\theta(z,t)$, minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(z_0,z_1),t}\big|\big|v_\theta(z_t,t) - (z_1 - z_0)\big|\big|_2^2. \tag{9}$$

During inference, given only the visual instruction image I_v, first take $z_0 \sim q_\psi(\cdot|E_v(I_v))$, and then solve the ordinary differential equation. Obtaining $\hat{z}_1 = z_{t=1}$, the final image $\hat{I} = \mathrm{Dec}_{\mathrm{img}}(\hat{z}_1)$ is generated through the frozen image VAE decoder. This process achieves continuous transportation from the instruction state to the image state within a unified one-dimensional sequence modality, avoiding additional noise scheduling and conditional branching, and maintaining an isomorphic representation to the visual instructions.

5 Experiment Results

In this section, we first provide the implementation details of FlowInOne (Sec. 5.1), and then present the main results on our carefully curated VP-Bench (Sec. 5.2). Finally, we conduct ablation studies to better understand the design choices of FlowInOne for image generation (Section 5.3).

5.1 Implementation Details

Model architecture. Based on the CrossFlow framework [31], we encode unified image input via Janus-pro-1B [9] and a frozen LDM VAE [47], projecting them through a stacked Transformer text-image VAE. We augment the Transformer blocks with additional cross-attention layers. The concatenated self- and cross-attention outputs are then fed into a lightweight network to predict spatially adaptive weights for input feature modulation.

Training details. The 1.2B FlowInOne is initialized from CrossFlow [31] and trained at 256×256 resolution for 240k steps via a balanced WebDataset [1], optimizing a combined Flow matching, KL divergence, and CLIP contrastive loss (see Appendix G for detailed configurations and more ablations).

Table 1: Evaluation on the VP-Bench visual instruction benchmark using three VLM evaluators and human judges. We use Gemini3 [19], GPT5.2 [39] and Qwen3.5 [45] to evaluate the success ratio of FlowInOne on all kinds of visual instruction tasks. C2I: class to image, T2I: text to image, TIE: text in image edit, FU: force understanding, TBE: text & bbox edit, TU: trajectory understanding, VME: visual marker edit, DE: doodles edit. **Total** denotes the average success rate across all sub-categories.

Method	C2I	T2I	TIE	FU	TBE	TU	VME	DE	Total
Evaluator: ***Gemini3***									
Nano Banana [20]	.810	**.980**	**.521**	.500	**.600**	.020	**.537**	**.740**	**.589**
Omnigen2 [61]	.720	.760	.313	.013	.020	.000	.020	.140	.248
Kontext [26]	.620	.700	.363	.027	.163	.020	.096	.180	.271
Qwen-IE-2509 [60]	.680	.690	.383	.047	.060	.000	.040	.160	.258
FlowInOne (Ours)	**.890**	.700	.355	**.727**	.302	**.520**	.292	.535	.540
Evaluator: ***GPT5.2***									
Nano Banana [20]	.760	**.960**	**.402**	.163	.100	.020	**.227**	**.495**	.391
Omnigen2 [61]	.660	.820	.203	.001	.000	.000	.001	.160	.231
Kontext [26]	.620	.690	.266	.013	.093	.000	.056	.160	.237
Qwen-IE-2509 [60]	.640	.680	.286	.040	.020	.020	.020	.140	.231
FlowInOne (Ours)	**.850**	.800	.079	**.500**	**.116**	**.240**	.083	.465	**.392**
Evaluator: ***Qwen3.5***									
Nano Banana [20]	.780	**.960**	**.446**	.427	.260	.040	**.395**	**.760**	**.508**
Omnigen2 [61]	.740	.790	.257	.027	.020	.000	.010	.160	.251
Kontext [26]	.720	.690	.322	.020	.133	.040	.083	.140	.269
Qwen-IE-2509 [60]	.780	.710	.345	.107	.060	.000	.043	.180	.278
FlowInOne (Ours)	**.859**	.720	.354	**.713**	**.272**	**.320**	.306	.481	.503
Evaluator: ***Human***									
Nano Banana [20]	.790	**.940**	**.372**	.287	.220	.020	**.306**	**.740**	**.459**
Omnigen2 [61]	.720	.710	.268	.013	.020	.000	.010	.120	.233
Kontext [26]	.640	.680	.317	.013	.080	.020	.048	.120	.240
Qwen-IE-2509 [60]	.700	.665	.331	.047	.020	.000	.023	.160	.243
FlowInOne (Ours)	**.800**	.645	.242	**.705**	**.255**	**.280**	.255	.400	.449

Evaluation metrics. We evaluate FlowInOne on a high-quality subset of VP-Bench. Following Section 3.2, we report pass rates assessed by Gemini 3 [19], GPT 5.2 [39], and Qwen3.5 [45] and compute four quantitative metrics. Additionally, for qualitative verification, ten independent evaluators cross-checked 250 stratified random samples (25 each) to ensure judgment reliability.

5.2 Evaluation on Image-to-image generation

State-of-the-art Comparison. We compare FlowInOne against several competitive open-source frameworks such as OmniGen2 [61], Qwen-Image-Edit-2509 [60], and FLUX.1-Kontext-dev [26], as well as the commercial model Nano Banana [20]. To ensure fairness, we evaluate each baseline models through its optimal native interface, using detailed text prompts expanded by Qwen3-VL [3] to supplement

Table 2: Overall four-dimensional breakdown. Each dimension has a maximum score of 5 points. IF: Instruction Faithfulness, CC: Content Consistency, VR: Visual Realism, SP: Spatial Precision.

Method	Gemini3				GPT5.2				Qwen3.5			
	IF	CC	VR	SP	IF	CC	VR	SP	IF	CC	VR	SP
Nano Banana	3.43	3.10	4.43	2.99	3.28	2.78	4.06	2.86	3.30	3.01	4.43	2.94
Omnigen2	2.15	1.39	3.46	1.69	2.29	1.53	3.31	1.92	1.72	0.91	3.13	1.27
Kontext	2.31	1.62	3.15	2.25	2.40	1.71	3.25	2.21	1.97	1.87	3.11	1.76
Qwen-IE-2509	2.96	2.51	3.28	1.91	2.76	2.39	3.58	1.94	2.33	1.98	3.23	1.39
FlowInOne	3.38	2.94	3.12	3.42	3.16	2.81	2.96	3.24	3.31	2.87	3.20	3.30

the image inputs. Quantitative results demonstrate that FlowInOne possesses a significant advantage in unified generation tasks.

As presented in Table 1, FlowInOne consistently achieves the best performance among all open-source baselines across different evaluators. Specifically, FlowInOne obtains total success rates of **54.0%**, **39.2%**, **50.3%**, and **44.9%** under Gemini3, GPT5.2, Qwen3.5, and Human evaluation, respectively, substantially outperforming OmniGen2, FLUX.1-Kontext-dev, and Qwen-IE-2509. Notably, FlowInOne also remains highly competitive with the commercial model Nano Banana, achieving the highest total score under GPT5.2 and only slightly trailing Nano Banana under Gemini3, Qwen3.5, and Human evaluation. These results demonstrate that FlowInOne establishes a strong open-source baseline and approaches commercial model in the image-in, image-out generation paradigm.

Table 2 further provides a fine-grained four-dimensional analysis. FlowInOne achieves consistently strong performance in instruction faithfulness and content consistency, showing that it can effectively follow visual instructions while preserving the required semantic content. More importantly, FlowInOne obtains the best spatial precision scores across all three evaluators, i.e., 3.42 under Gemini3, 3.24 under GPT5.2, and 3.30 under Qwen3.5, demonstrating its advantage in spatially grounded image-in, image-out generation. Although Nano Banana achieves higher visual realism, FlowInOne shows a better balance between instruction following and spatial control, especially compared with existing open-source baselines.

Although minor discrepancies exist between MLLMs and human assessments due to limitations in grounding fine-grained visual markers [12], the overall ranking trends consistently align, verifying the reliability of our automated metrics.

Beyond pass rates, Table 3 further substantiates our findings. Crucially, **FlowInOne excels in fine-grained spatial and physical controls, achieving the highest average DINOv3 Sim score of 48.7% (outperforming Nano Banana's 47.3%)**, with notable margins in force & trajectory understanding and text bbox editing. Furthermore, **in overall visual realism and semantic alignment, FlowInOne significantly surpasses all open-source baselines and performs comparably to the commercial model**, underscoring its robust generation quality and precise instruction-following capabilities.

Table 3: Quantitative evaluation on VP-Bench across five models. CLIP-IQA is averaged over all categories to measure overall visual realism. CLIP Score evaluates semantic alignment for pure generation tasks. Directional CLIP Similarity assesses instruction-driven consistency in marker-based editing. DINOv3 Similarity measures fine-grained spatial and physical structural accuracy.

Method	IQA ↑	CLIP Score ↑			Dir CLIP ↑			DINOv3 Sim ↑				
	Total	C2I	T2I	**Avg.**	TIE	VME	**Avg.**	DE	FU	TBE	TU	**Avg.**
Nano Banana [20]	**0.688**	0.281	**0.302**	**0.291**	**0.106**	**0.103**	**0.105**	**0.430**	0.474	0.501	0.486	0.473
Omnigen2 [61]	0.603	0.173	0.208	0.191	0.001	0.005	0.003	0.224	0.066	0.004	0.207	0.125
Kontext [26]	0.621	0.164	0.172	0.168	0.010	0.008	0.009	0.283	0.240	0.118	0.004	0.161
Qwen-IE-2509 [60]	0.646	0.231	0.216	0.224	0.005	0.011	0.008	0.133	0.215	0.250	0.207	0.201
FlowInOne (Ours)	0.684	**0.290**	0.276	0.283	0.092	0.101	0.097	0.335	**0.536**	**0.506**	**0.570**	**0.487**

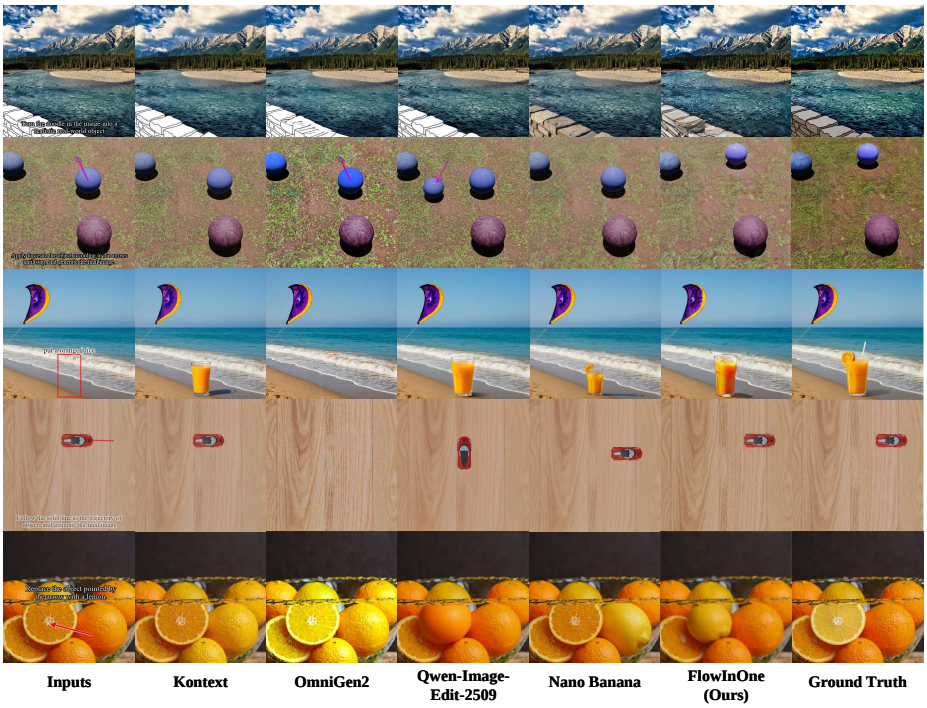

Fig. 4: Visual instruction editing comparison across methods.

Qualitative Comparison. Figure 4 visually compares FlowInOne and baselines across five VP-Bench tasks: force, trajectory, text & bbox, visual marker, and doodle editing. Unlike traditional pipelines, FlowInOne directly processes a unified canvas containing textual instructions, spatial layouts, and visual cues (arrows, markers, doodles). For fairness, we evaluate baselines via their optimal native interfaces: instructions are extracted and expanded into detailed prompts by Qwen3-VL before being provided alongside the processed image.

Despite these enhanced prompts, baselines often fail to translate visual cues into precise edits. In force and trajectory tasks, they struggle to convert arrows into physically plausible motions. For text & bbox editing, they frequently violate spatial constraints or size specifications. In marker and doodle tasks, baselines often misinterpret localized cues as scene elements, resulting in inaccurate synthesis or artifact retention. Conversely, FlowInOne accurately executes modifications while preserving background consistency. This success confirms that our image-in, image-out paradigm enables more reliable grounding of fine-grained spatial and physical intents than conventional text-driven interfaces.

5.3 Ablation Studies

We conduct a series of ablation studies to validate the architectural designs of our model. Due to computational constraints, all models are trained for 100k steps with a batch size of 512 unless otherwise specified. To accurately reflect model performance, the reported Pass Rate averages the Gemini and GPT evaluations. Furthermore, we adopt a progressive strategy, building upon the optimal configuration from each preceding part.

Table 4: Ablation study on compression method, modulation method, and training strategy. The average scores across Gemini3 [19], GPT5.2 [39] and Qwen3.5 [45] are used as the evaluation metric.

Method	Gemini↑	GPT↑	Qwen↑	Avg.
MLP + truncation	0.179	0.153	0.176	0.169
VAE expansion	0.169	0.147	0.155	0.157
MLP + MLP	0.192	0.170	0.185	**0.182**

(a) Compression Method

Cross attention	Gemini↑	GPT↑	Qwen↑	Avg.
Wo CA	0.192	0.170	0.185	0.182
W Dual-Path CA	0.227	0.185	0.229	0.214
Dual-Path SAM	0.242	0.214	0.238	**0.231**

(b) Modulation Method

Train strategy	Gemini↑	GPT↑	Qwen↑	Avg.
2-stage training	0.336	0.256	0.283	0.291
Joint training	0.540	0.392	0.503	**0.478**

(c) Training strategy

Different compression methods. Since our model leverages pre-trained weights, aligning the sequence length and feature dimensions of visual encoder to the pre-trained latent space is critical. We explore three distinct compression and mapping strategies: (1) **MLP + truncation**: The feature dimension is mapped via an MLP, while the sequence length is directly truncated; (2) **VAE expansion**: The number of Transformer layers in the VAE is directly increased to naturally match dimensions within the latent space; (3) **MLP + MLP**: Both the sequence length and feature dimension are projected via MLPs. As shown in Table 4(a), the **MLP + MLP** strategy performs best (18.12%). We hypothesize that simple truncation discards critical edge information, while merely expanding VAE layers

increases optimization difficulty. Conversely, **dual MLP projection effectively preserves the semantic and spatial structure of the input visual prompt while maintaining pre-trained priors.**

Token gated cross attention. We investigate modulation mechanisms to unify generation and editing, which exhibit distinct structural dependencies. We compare: (1) **Wo CA**: self-attention only; (2) **W Dual-Path CA**: dual-path cross-attention without adaptive gating; and (3) **Dual-Path SAM**: our proposed Dual-Path Spatially-Adaptive Modulation. Table 4(b) shows that lacking cross-attention (Wo CA) yields the poorest results (18.12%) due to insufficient utilization of source image structural priors. Adding cross-attention improves performance to 21.40%, while Dual-Path SAM reaches 23.1%. **This demonstrates that the adaptive mechanism balances content consistency and instruction adherence, optimizing performance within a unified framework.**

Joint training vs. two-stage training. Based on the optimal architecture described above, we evaluate data training strategies. We compare: (1) **Two stage training**: 100k steps on 3M samples (T2I, C2I, coarse-grained editing), followed by 140k steps on the remaining 2M samples; and (2) **Joint training**: mixing all 5M samples for 240k steps. As Table 4(c) shows, joint training dominates with a **47.8%** pass rate, far surpassing the two-stage approach (29.1%). Two-stage training likely suffers from catastrophic forgetting across varying tasks. Instead, **joint training forces the model to simultaneously learn semantic generation, geometric transformation, and physical laws within a shared visual flow space, yielding stronger generalization and instruction adherence.**

6 Conclusion

We presented FlowInOne, a unified framework that redefines multimodal generation as a purely visual flow. By embedding all modalities into a shared visual space and learning continuous transport between visual instruction and image states, FlowInOne achieves efficient and consistent generation across diverse tasks. To support this paradigm, we introduced the large-scale VisPrompt-5M dataset, enabling cross-task generalization under a single visual interface. FlowInOne achieves state-of-the-art performance across all evaluated tasks, surpassing existing open-source models and rivaling competitive commercial systems in both automated and human evaluations. Extensive experiments demonstrate that FlowInOne establishes a foundation for future vision-centric multimodal models that unify perception and generation under a single deterministic principle. We believe this work marks a step toward closing the gap between visual understanding and creation within a continuous visual domain.

Acknowledgements

This work is supported by National Natural Science Foundation of China (grant number 62502544).

References

1. Aizman, A., Maltby, G., Breuel, T.: High performance i/o for large scale deep learning. arXiv preprint arXiv:2001.01858 (2020)
2. Bai, J., Ye, T., Chow, W., Song, E., Chen, Q.G., Li, X., Dong, Z., Zhu, L., YAN, S.: Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In: International Conference on Learning Representations. vol. 2025, pp. 74103–74160 (2025)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (June 2023)
6. Chang, Y., Chen, J., Cheng, A., Bogdan, P.: Maskattn-sdxl: Controllable region-level text-to-image generation. arXiv preprint arXiv:2509.15357 (2025)
7. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. arXiv preprint arXiv:2305.10855 (2023)
8. Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-\alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 57611–57640 (2024)
9. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
10. Chen, Z.K., Jiang, J.P., Ye, H.J., Zhan, D.C.: Hawk: Leveraging spatial context for faster autoregressive text-to-image generation. arXiv preprint arXiv:2510.25739 (2025)
11. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024)
12. Dong, Z., Yi, J., Zheng, Z., Han, H., Zheng, X., Wang, A.J., Liu, F., Li, L.: Seeing is not reasoning: Mvpbench for graph-based evaluation of multi-path visual physical cot. arXiv preprint arXiv:2505.24182 (2025)
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. vol. 235, pp. 12606–12633 (2024)
14. Fan, L., Li, T., Qin, S., Li, Y., Sun, C., Rubinstein, M., Sun, D., He, K., Tian, Y.: Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. arXiv preprint arXiv:2410.13863 (2024)

15. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Trans. Graph. **41**(4) (2022)
16. Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R.T.Q., Synnaeve, G., Adi, Y., Lipman, Y.: Discrete flow matching. In: Advances in Neural Information Processing Systems. vol. 37, pp. 133345–133385 (2024)
17. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
18. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. arXiv preprint arXiv:2505.19386 (2025)
19. Google: Gemini. `https://deepmind.google/models/gemini/` (2025), accessed: 2026-01-07
20. Google: Nano banana. `https://aistudio.google.com/models/gemini-2-5-flash-image` (2025), accessed: 2026-01-22
21. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500 (2018)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
24. Jia, J., Gao, J., Xue, B., Wang, J., Cai, Q., Chen, Q., Zhao, X., Jiang, P., Gai, K.: From principles to applications: A comprehensive survey of discrete tokenizers in generation, comprehension, recommendation, and information retrieval. arXiv preprint arXiv:2502.12448 (2025)
25. Kim, D., He, J., Yu, Q., Yang, C., Shen, X., Kwak, S., Chen, L.C.: Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. arXiv preprint arXiv:2501.07730 (2025)
26. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
27. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)
28. Lin, W., Wei, X., Zhang, R., Zhuo, L., Zhao, S., Huang, S., Xie, J., Qiao, Y., Gao, P., Li, H.: Pixwizard: Versatile image-to-image visual assistant with open-language instructions. arXiv preprint arXiv:2409.15278 (2024)
29. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
30. Liu, G.H., Vahdat, A., Huang, D.A., Theodorou, E.A., Nie, W., Anandkumar, A.: I2sb: Image-to-image schrödinger bridge. In: International Conference on Machine Learning (2023)
31. Liu, Q., Yin, X., Yuille, A., Brown, A., Singh, M.: Flowing from words to pixels: A noise-free framework for cross-modality evolution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2755–2765 (2025)

32. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
33. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
34. Liu, X., Zhang, X., Ma, J., Peng, J., Liu, Q.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. arXiv preprint arXiv:2309.06380 (2024)
35. Liu, Y., Fu, X., Huang, J., Xiao, J., Li, D., Zhang, W., Bai, L., Zha, Z.J.: Latent harmony: Synergistic unified uhd image restoration via latent space regularization and controllable refinement. arXiv preprint arXiv:2510.07961 (2025)
36. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
37. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
38. Nobis, G., Springenberg, M., Belova, A., Daems, R., Knochenhauer, C., Opper, M., Birdal, T., Samek, W.: Fractional diffusion bridge models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
39. OpenAI: Chatgpt. `https://chatgpt.com/` (2025), accessed: 2026-01-16
40. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 Conference Proceedings (2023)
41. Peebles, W.S., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4172–4182 (2023)
42. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
43. Polyak, A., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2025)
44. Qian, Y., Bocek-Rivele, E., Song, L., Tong, J., Yang, Y., Lu, J., Hu, W., Gan, Z.: Pico-banana-400k: A large-scale dataset for text-guided image editing. arXiv preprint arXiv:2510.19808 (2025)
45. Qwen Team: Qwen3.5: Towards native multimodal agents (2026), `https://qwen.ai/blog?id=qwen3.5`, accessed: 2026-03-01
46. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8821–8831 (2021)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (June 2022)
48. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**, 211 – 252 (2014)
49. Rust, P., Lotz, J.F., Bugliarello, E., Salesky, E., de Lhoneux, M., Elliott, D.: Language modelling with pixels. arXiv preprint arXiv:2207.06991 (2022)

50. Salesky, E., Etter, D., Post, M.: Robust open-vocabulary translation from visual text representations. arXiv preprint arXiv:2104.08211 (2021)
51. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
52. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
53. Sucheng, R., Qihang, Y., Ju, H., Xiaohui, S., Alan, Y., Liang-Chieh, C.: Beyond next-token: Next-x prediction for autoregressive visual generation. arXiv preprint arXiv:2502.20388 (2025)
54. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
55. Wang, H., Zhang, J., Chen, H., Guo, H., Wang, D., Ma, J., Du, B.: Residual diffusion bridge model for image restoration. arXiv preprint arXiv:2510.23116 (2025)
56. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI Conference on Artificial Intelligence (2022)
57. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring clip for assessing the look and feel of images. arXiv preprint arXiv:2207.12396 (2022)
58. Wang, Y., Yang, S., Zhao, B., Zhang, L., Liu, Q., Zhou, Y., Xie, C.: Gpt-image-edit-1.5m: A million-scale, gpt-generated image dataset. arXiv preprint arXiv:2507.21033 (2025)
59. Weber, M., Yu, L., Yu, Q., Deng, X., Shen, X., Cremers, D., Chen, L.C.: Maskbit: Embedding-free image generation via bit tokens. arXiv preprint arXiv:2409.16211 (2024)
60. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
61. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
62. Xiao, C., Huang, Z., Chen, D., Hudson, G.T., Li, Y., Duan, H., Lin, C., Fu, J., Han, J., Moubayed, N.A.: Pixel sentence representation learning. arXiv preprint arXiv:2402.08183 (2024)
63. Xiao, J., Nayak, R., Zhang, N., Toertei, D., Loianno, G.: Thermalgen: Style-disentangled flow-based generative models for RGB-to-thermal image translation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
64. Xing, L., Wang, A.J., Yan, R., Shu, X., Tang, J.: Vision-centric token compression in large language model. In: Advances in Neural Information Processing Systems. vol. 38, pp. 33080–33110 (2025)
65. Yang, Y., Gui, D., Yuan, Y., Liang, W., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. Advances in Neural Information Processing Systems **36** (2024)

66. Ye, K., Huang, Z., Fu, C., Liu, Q., Cai, J., Lv, Z., Li, C., Lyu, J., Zhao, Z., Zhang, S.: Unicedit-10m: A dataset and benchmark breaking the scale-quality barrier via unified verification for reasoning-enriched edits. arXiv preprint arXiv:2512.02790 (2025)
67. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. arXiv preprint arXiv:2411.00776 (2024)
68. Zhang, H., Siarohin, A., Menapace, W., Vasilkovsky, M., Tulyakov, S., Qu, Q., Skorokhodov, I.: Alphaflow: Understanding and improving meanflow models. arXiv preprint arXiv:2510.20771 (2025)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3813–3824 (2023)
70. Zheng, Y., Ren, Y., Xia, X., Xiao, X., Xie, X.: Dense2moe: Restructuring diffusion transformer to moe for efficient text-to-image generation. arXiv preprint arXiv:2510.09094 (2025)
71. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: The Twelfth International Conference on Learning Representations (2024)
72. Zhou, Y., Liu, B., Zhu, Y., Yang, X., Chen, C., Xu, J.: Shifted diffusion for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10157–10166 (2023)
73. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2020)
74. Zou, K., Zhao, Z., Liu, B., Yu, N.: Advancing aesthetic image generation via composition transfer. International Journal of Computer Vision **134**, 252 (2026)