JoLT: Joint Latent Trajectories for Context-Guided High-Resolution Tiled Generation

Mathis Koroglu[1,2], Guillaume Jeanneret[2], Hugo Caselles-Dupré[1], Matthieu Cord[2,3], and Arnaud Dapogny[2]

[1] Obvious Research, Paris, France
[2] Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[3] Valeo.ai, Paris, France

Abstract. Although text-to-image generative models produce impressive results, they struggle to generate densely detailed, high-resolution (HR) images. Current literature addresses this issue with a low-to-high-resolution approach. First, a low-resolution (LR) image is generated. Then, an upsampled version is generated using the LR image as an additional cue. In this paper, we present Joint Latent Trajectories (JoLT). To generate an image, JoLT uses two streams that jointly denoise LR and HR latent images at each sampling step. The LR latent controls the overall layout, while the HR latent controls the details. We interconnect both branches to jointly integrate their information. We extensively validate our method, demonstrating its advantages over competing baselines. The resulting images are not only richly detailed but also visually pleasing, opening new avenues for artistic creation.

Keywords: High-resolution image generation · Tiled diffusion · Text-to-Image

1 Introduction

The introduction of novel generative technologies, such as text-to-image (T2I) models, has sparked interest in adopting these tools for artistic creation. However, using vanilla T2I models can still feel limited, as even the best-performing models, such as FLUX.2, produce relatively little detail. This shortfall is most consequential for large-format works meant for physical display, such as prints or murals, which are read at two distances: the composition from afar, the fine detail up close. Artists would like to freely control intricate local details, including through natural language, without sacrificing the coherence and aesthetics of the global composition.

To set a frame of reference, let us clearly state what we refer to as a "complex" or "detailed" image. We consider an image to be complex or detailed if the density of semantic information is high locally, yet the scene can be understood globally.

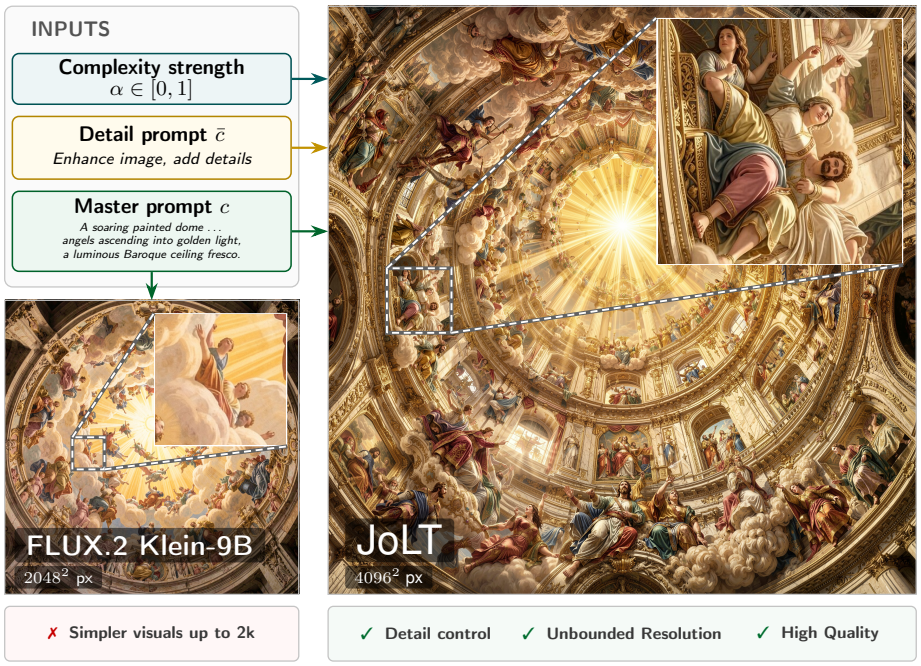


Fig. 1: Controllable high-resolution generation with JoLT. The master prompt c conditions both the FLUX.2 Klein-9B base generation (left) and our result (right), while the detail prompt $\bar{c}$ and complexity strength α control the added local detail. Dashed boxes identify matched groups of figures near the image center, enlarged in the overlaid crops. The base image is displayed at half the width and height of our result; the bands summarize their qualitative capabilities.

However, it is not complex if it only contains texturized or recurring patterns. For example, increasing the sharpness of an image is perceived by the human eye as more detailed, yet it does not fit into our definition. Fig. 1 shows an example of a generation made with FLUX.2 Klein-9B and another made with our proposed approach, both using the same input prompt. As can be seen here, FLUX.2 Klein-9B's generation is simple, even with a detailed prompt. In contrast, our image is locally complex yet globally coherent.

Current training-free methods typically generate large, aesthetically pleasing images in two steps. First, they create an LR image and, second, they upsample it to generate details or refine textures. This strategy has two drawbacks. First, since the LR image is bound by the inherent complexity limitations of the underlying model, the refining step barely enhances the richness of the textures. Second, since the LR image has already been generated, it cannot be influenced by HR information, making it difficult to integrate fine-grained elements seamlessly into the global layout. Existing approaches therefore face an inherent

trade-off between detail and global coherence. This motivates a different strategy for generating these images.

In this paper, we focus on generating densely detailed HR images in a training-free manner. Instead of using the conventional low-to-high-resolution generation paradigm, we propose a novel two-stream approach that jointly generates LR and HR latent features. Our method, called Joint Latent Trajectories (JoLT), transfers information from the LR stream to the HR stream via the context input of FLUX.2, then feeds the evolving HR representation back into the LR stream. Consequently, the detail prompt does more than populate local regions: together with the master prompt, it helps shape the style and rendering of the final result. This bidirectional exchange integrates both resolutions into a coherent scene while giving users control over image complexity and stylization.

To summarize, our contributions are as follows:

- A Novel Joint-Latent Paradigm: We propose JoLT, a training-free pipeline for generating high-resolution, complex images. JoLT transforms the standard low-to-high generation paradigm into a simultaneous low-and-high generation paradigm.
- Bidirectional Latent Information Exchange: To transfer information between the HR and LR branches, we propose a cross-resolution feedback mechanism that leverages the in-context capabilities of FLUX.2 and variance-matched latent subsampling.
- Controllable High-Density Synthesis: We enable users to explicitly control visual complexity (α), as well as the style and local content of the final result through the detail prompt ($\bar{c}$), without requiring fine-tuning.

With a series of strong empirical studies, we show the advantages of JoLT over recent methods and baselines.

Project resources. The project page, source code, and evaluation prompts are available at https://obvious-research.github.io/jolt/.

2 Related Work

Diffusion and rectified-flow generative backbones. Modern image generation is largely built on diffusion and score-based models that reverse a gradual noising process [17,40], with samplers such as DDIM [39] and classifier-free guidance [18]. Since pixel-space denoising scales poorly with resolution, latent diffusion moves the process into a compressed autoencoder space and makes high-resolution text-to-image synthesis practical [37]. Newer systems shift toward transformer backbones and rectified-flow objectives that reshape the generative trajectory: Diffusion Transformers [32] denoise over latent patches, making capacity and token resolution central scaling variables, Stable Diffusion 3 [13] couples a multimodal diffusion transformer with a rectified-flow formulation, and families such as FLUX [4] and HunyuanImage [7] carry the trend into large-scale democratization in the creative community. A distinct paradigm is image-conditioned,

in-context generation, where a reference image is optionally fed as additional conditioning. In-Context LoRA [20] adapts a diffusion transformer to jointly generate related images from concatenated in-context examples, and OminiControl [41] injects image conditions by appending their tokens to the generation sequence with minimal added parameters. FLUX.1 Kontext [6] unifies generation and editing in one latent flow-matching model. Our method builds on this in-context view, conditioning on a lower-resolution image as it is being denoised.

High-detail and high-resolution generative synthesis. High-resolution generative synthesis attempts to create large, coherent, semantically rich images from text and/or reference context, outside the generator's native training resolution. MultiDiffusion [2] fuses diffusion paths over spatial windows to enable spatially controlled generation. DiffCollage [52] similarly composes large content through parallel diffusion generation, while SyncDiffusion [25] synchronizes overlapping diffusion windows to decrease window incoherence for separate patches. Subsequent panorama and synchronized-generation methods further develop this idea for wide fields of view and semantically coherent compositions [23,33,43,49,50]. These works show that increasing canvas size or detail density is not only a matter of producing more pixels: local generation must be coordinated so that fine detail does not fragment the global scene.

Prompt- and strength-based controls are also well established in research and artistic tools. Mixture of Diffusers [3] assigns prompts to spatial regions and harmonizes their diffusion processes across a larger canvas, while ControlNet Tile [51] and Ultimate SD Upscale [11] combine controlled tiled refinement with denoising-strength controls. In contrast to these sequential refinement approaches, JoLT jointly denoises coupled LR and HR trajectories, allowing the evolving representations to interact bidirectionally at every step rather than treating a completed LR image as a fixed input to HR refinement.

A second branch studies tuning-free high-resolution generation with pretrained generative models. DemoFusion [12] builds on pretrained latent diffusion and combines progressive upscaling, skip residuals, and dilated sampling to produce high-resolution outputs without retraining. ScaleCrafter [15] analyzes direct high-resolution generation with pretrained diffusion models, attributing failures such as repetition and structural artifacts partly to limited effective receptive fields and introducing inference-time changes such as re-dilation. AccDiffusion [28] targets semantic inconsistency and object repetition in higher-resolution generation. More recent work moves this problem into Diffusion Transformers: ResDiT [31] studies positional-embedding behavior, local enhancement, patch-level fusion, and layout preservation for resolution extrapolation, while SEGA [36] proposes spectral-energy guided attention to balance global structure and fine-detail fidelity in DiT-based extrapolation. They reveal a recurring tension between local detail, global coherence, computational cost, and adherence to conditioning signals, which is precisely the regime in which high-complexity artistic image generation becomes difficult.

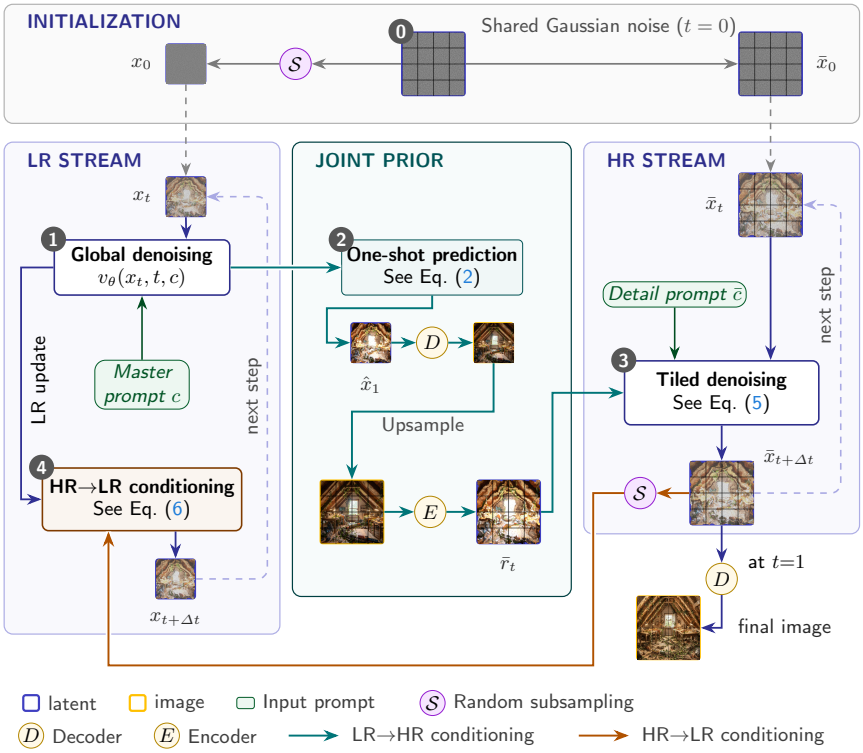


Fig. 2: Overview of JoLT. At $t = 0$, a shared Gaussian draw initializes the HR canvas $\bar{x}_0$ and is randomly subsampled to initialize the LR latent x_0. Two denoising trajectories then run jointly over the same 4 steps of a distilled flow-matching model. The LR stream denoises the whole scene from the master prompt; the HR stream denoises overlapping windows and adds local detail. At every step, the LR stream's clean estimate $\hat{x}_1$ is decoded, upsampled, and re-encoded into a reference latent $\bar{r}_t$ whose aligned crops $F_i(\bar{r}_t)$ condition the HR stream. The updated HR canvas is randomly subsampled and blended into the LR update with weight α before the next step.

3 Method

JoLT couples a global LR denoising trajectory with a tiled HR trajectory, as seen in Fig. 2. Our method uses LR and HR branches: the former maintains coherent scene composition, while the latter synthesizes local detail. Importantly, these two branches do not operate separately. Instead, they exchange information to synchronize high- and low-level image semantics.

3.1 Preliminaries

In-context Flow-Matching Sampling. Modern T2I generators use the flow matching formulation to iteratively generate an image. Let v_θ be the T2I model and

$x_0 \sim \mathcal{N}(0, I) \in \mathbb{R}^{C\times H\times W}$ the starting image latent, where C, H, and W denote its number of channels, height, and width. Let t be the current timestep, Δt the timestep size, and c the text prompt. Sampling follows:

$$x_{t+\Delta t} = x_t + \Delta t \cdot v_\theta(x_t, t, c), \tag{1}$$

until $t = 1$. We can also compute a one-shot estimate of the clean latents with:

$$\hat{x}_1(x_t, t, c) = x_t + (1 - t) \cdot v_\theta(x_t, t, c). \tag{2}$$

The formulation in Eq. (2) is equivalent to Eq. (1) by choosing $\Delta t = 1 - t$. Let D and E be the decoder and encoder of a variational autoencoder (VAE). When $t = 1$ (or $\hat{x}_1$ is created), the image is decoded into pixel space with $D(x_1)$. Optionally, v_θ can also take a reference image r (encoded with E), *i.e.*, v_θ takes as input (x_t, t, c) or (x_t, t, c, r).

MultiDiffusion. MultiDiffusion [2] synthesizes spatially larger latents than the native denoising field by running the base model on overlapping spatial windows and reconciling their local flow predictions. We refer to this higher-resolution latent as a canvas. Formally, suppose that we have a noised latent canvas $\bar{x}_t \in \mathbb{R}^{C\times \bar{H}\times \bar{W}}$ at timestep t, with $\bar{H} > H$ and $\bar{W} > W$ as its height and width. Let $F_i(\cdot)$ be a cropping operation and $F_i^\top(\cdot)$ a zero-padding function, both operating in the i^{th} window (out of N). MultiDiffusion combines the outputs of all N windows as follows:

$$\bar{v}_{\text{MD}}(\bar{x}_t, t, c) = \frac{\sum_{i=1}^{N} F_i^\top\big(w \odot v_\theta(F_i(\bar{x}_t), t, c)\big)}{\sum_{i=1}^{N} F_i^\top(w)}, \tag{3}$$

where $w \in \mathbb{R}^{H\times W}$ are cosine border-ramp weights that smooth transitions between regions to avoid any visible seams. For the rest of the paper, a bar $\bar{(\cdot)}$ denotes an HR variable.

3.2 Joint LR and HR trajectories

We propose a two-stream approach: each stream transfers information to the other at every iteration. Adopting the variables from the previous section, let x_t and $\bar{x}_t$ be the LR and HR latents, and let c and $\bar{c}$ be the master and detail prompts. The former describes the overall content of the scene, while the latter describes the details to add. Each iteration begins by computing the LR velocity $v_\theta(x_t, t, c)$ and updating x_t using Eq. (1); the resulting global estimate then conditions the HR branch as described below.

LR→HR conditioning.[*] We produce a one-shot approximation $\hat{x}_1$ using Eq. (2). Using D, we decode the LR one-shot latents to pixel space, upsample them, and then encode them back to latent space using E, that is

$$\bar{r}_t = E\Big(\texttt{upsample}\big(D(\hat{x}_1)\big)\Big). \tag{4}$$

[*] These colors match the color coding used in the method overview (Fig. 2).

We call $\bar{r}_t$ our joint prior. We deliberately resize in pixel space because conventional latent-space upsampling cannot simultaneously preserve the marginal distribution [8]. The joint prior thus captures a coarse representation of the content in x_t.

To update the canvas $\bar{x}_t$ with $\bar{r}_t$, we adapt MultiDiffusion's tiled-denoising strategy to leverage in-context T2I conditioning using $\bar{r}_t$. For each window i, we crop the reference image, $F_i(\bar{r}_t)$, and forward it alongside the canvas crop $F_i(\bar{x}_t)$. Consequently, we modify Eq. (3) to include $\bar{r}_t$ as follows:

$$\bar{v}_\theta(\bar{x}_t, t, \bar{c}, \bar{r}_t) = \frac{\sum_{i=1}^{N} F_i^\top \left(w \odot v_\theta(F_i(\bar{x}_t), t, \bar{c}, F_i(\bar{r}_t)) \right)}{\sum_{i=1}^{N} F_i^\top (w)}. \tag{5}$$

Finally, we update $\bar{x}_t$ with $\bar{v}_\theta(\bar{x}_t, t, \bar{c}, \bar{r}_t)$ using the standard update schedule (Eq. (1)). This sequence produces a consistent trajectory and prevents divergence among the different windows.

HR→LR conditioning.[*] In the reverse direction, we update the LR branch to reflect the details already generated in the HR branch. Once both branches have taken their step, we randomly subsample the canvas $\bar{x}_{t+\Delta t}$ to preserve its marginal latent distribution and inject it into $x_{t+\Delta t}$. More specifically, we linearly mix both branches and then correct the variance to match that of the LR latents using

$$\begin{aligned} x^{\text{temp}} &= (1-\alpha)\, x_{t+\Delta t} + \alpha\, \mathcal{S}(\bar{x}_{t+\Delta t}) \\ x_{t+\Delta t} &\leftarrow \mu(x_{t+\Delta t}) + \frac{\sigma(x_{t+\Delta t})}{\sigma(x^{\text{temp}})}(x^{\text{temp}} - \mu(x^{\text{temp}})), \end{aligned} \tag{6}$$

where $\mathcal{S}$ is a subsampling mechanism, $\alpha \in [0, 1]$ is the HR→LR conditioning strength, and μ and σ denote per-channel spatial mean and standard deviation, broadcast back to the latent grid. To initialize the two latents x_0 and $\bar{x}_0$, we do not sample them independently. Instead, we sample $\bar{x}_0$ and use $\mathcal{S}$ to create $x_0 = \mathcal{S}(\bar{x}_0)$. The additional LR trajectory requires one model evaluation for every N HR window evaluations, yielding a relative overhead of $1/N$ that becomes negligible for large images.

4 Experiments

4.1 Implementation Details and Baselines

We use a four-step-distilled FLUX.2 Klein-9B backbone [5]. Unless stated otherwise, we generate a 4096^2 px canvas using 1024^2 px HR windows with a 512 px stride in both dimensions, *i.e.*, 50% overlap. The LR stream operates at 1024^2 px, corresponding to a downsampling factor $s = 4$. Every run uses a single constant α throughout the four sampling steps; unless stated otherwise, $\alpha = 0.25$.

Our full method recomputes the LR-derived in-context prior at every denoising step (joint-prior conditioning), uses $\alpha = 0.25$ for HR→LR conditioning, and

applies MultiDiffusion (MD) aggregation in the HR stream, producing 4096^2 px images. We choose the most representative approaches from the literature as our baselines: DemoFusion [12], SEGA [36], and ResDiT [31]. For fairness, we re-implement every baseline with the same FLUX.2 Klein-9B four-step backbone, isolating the HR strategy from the base model.

Evaluation Dataset We use GPT-5.6 Sol-high to generate 200 semantically dense evaluation prompts. Each prompt is 30 to 80 words long (36.3 on average) and combines at least four otherwise disparate objects, activities, or settings through explicit spatial, functional, or visual relationships. We release the prompt set in the project repository.

Metrics We assess three complementary axes:

Complexity metrics measure the degree of image detail according to our definition. We report IC9600 [14], Discrete Cosine Transform energy (DCT) [1], and JPEG ratio [42]; higher values indicate greater complexity rather than universally better quality. DCT and JPEG ratio are frequency- and compression-based complexity proxies, respectively; additional measures are reported in Sec. C.2.

Text alignment metrics assess correspondence between the generated image and its prompt. We use LMM4LMM-Correspondence [45] (LMM corr.): it is well suited to detailed correspondence judgments on our long prompts, and it is the alignment judge that agrees best with the human prompt-match preferences of our user study (Sec. 4.4). Additional alignment metrics and their agreement analysis appear in Secs. C.3 and C.4.

Finally, *aesthetics and quality* metrics assess visual appeal, composition, and the absence of artifacts. We report LMM4LMM-Perception (LMM percept.) and the Qwen-Image-Bench (QIB) Quality score [27], which aggregates its three level-2 dimensions: Realism, Detail, and Resolution. All reported metrics are higher-is-better (or higher-is-more-complex). Metrics excluded from our primary claims because of resolution or context limitations are reported in Sec. C.6. We complement these axes with a global–local coherence comparison; full protocol details are reported in Sec. C.1.

As an additional evaluation, we conduct a user study, described in Sec. 4.4.

4.2 Quantitative Comparison

As shown in Tab. 1, several points are evident. (i) Our full method outperforms all baselines on the three complexity metrics shown, with particularly large gains in DCT and JPEG ratio. (ii) Its LMM4LMM correspondence score remains competitive at 0.367, only 0.004 below SEGA, and this difference is not significant after metric-family correction. The detail prompt deliberately introduces plausible content beyond the master prompt, which can be penalized by strict correspondence judgments even when it increases local visual complexity. (iii) JoLT has the highest mean QIB Quality and LMM4LMM Perception scores, although neither top-two difference is significant after correction. Thus, its added complexity

creates a distinct aesthetic without measured quality degradation. Section C.5 summarizes the metric families.

Table 1: Quantitative comparison on 200 prompts. Our full method leads all three displayed complexity metrics. Compared with the strongest baseline, the gains are 5% on IC9600, 50% on DCT energy, and 26% on JPEG ratio. It also has the highest means on LMM4LMM perception and QIB Quality. Its LMM4LMM correspondence is within 0.004 of the leading baseline. Entries are mean ± sample standard deviation. Boldface and underlining mark the best and second-best means only when their paired difference is significant after metric-family correction ($p_{\text{adj}} < .05$, two-sided Wilcoxon). Latency reports seconds per image on one H100 80 GB

	Complexity			Text alignment	Aesthetics & quality		Latency
Method	IC9600 (↑)	DCT (↑)	JPEG (↑)	LMM corr. (↑)	LMM percept. (↑)	QIB Quality (↑)	Time (s) (↓)
SEGA	0.615±0.070	483±236	0.096±0.022	0.371±0.038	0.423±0.160	47.2±10.4	75±9
ResDiT	<u>0.714±0.073</u>	490±206	0.095±0.019	0.360±0.043	0.448±0.152	47.3±10.0	86±22
DemoFusion	0.689±0.068	<u>613±219</u>	<u>0.104±0.017</u>	0.356±0.044	0.411±0.167	50.2±9.2	125±10
Ours	**0.753±0.055**	**920±140**	**0.131±0.011**	0.367±0.047	0.462±0.164	50.9±9.4	127±15

To assess whether added detail belongs to the surrounding scene, Qwen3.6-27B [34] compares each full image and four corresponding crops without seeing the generation prompt or method names. Table 2 shows that the judge rates JoLT in the same range as DemoFusion and significantly prefers it to SEGA and ResDiT ($p_{\text{Holm}} < .001$).

Table 2: Global–local coherence comparison on 200 paired prompts per baseline. Order agreement is the percentage of pairs for which the decision is unchanged after reversing the A/B presentation order. Intervals are 95% percentile-bootstrap CIs and p_{Holm} corrects exact sign tests. More details are provided in Sec. C.1.

Baseline	Preference [95% CI]	Order agree.	p_{Holm}	Outcome
DemoFusion	.458 [.398, .520]	79.5%	.204	Same range
SEGA	.748 [.693, .800]	83.5%	$< .001$	JoLT preferred
ResDiT	.940 [.910, .965]	93.0%	$< .001$	JoLT preferred

4.3 Qualitative Comparison

Next, we qualitatively compare JoLT with the baselines in Fig. 3; additional examples are provided in Fig. 8. The full-frame views show global composition, while the center crops highlight local detail. Across these examples, ResDiT tends toward a darker ambience, SEGA often exhibits repetitive textures, and DemoFusion uses simpler layouts. In contrast, JoLT produces denser and more varied local structure while retaining a coherent global composition.

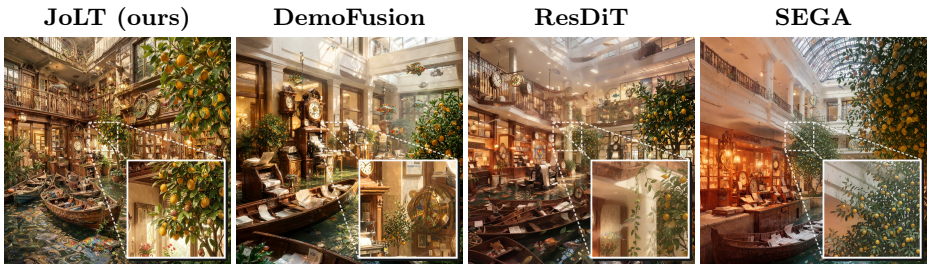


Fig. 3: Qualitative comparison with high-resolution generation baselines. Our result is shown in the leftmost panel. All methods use the same backbone and output resolution. Each lower-right inset magnifies the square center crop spanning one eighth of the source image width and height. Among the compared outputs, ours presents the densest and most varied fine-grained detail in both the full-frame view and the magnified inset.

Overall, JoLT produces striking visuals with substantially greater detail and visual complexity than the compared state-of-the-art methods. Its denser local structure and richer compositions introduce a distinct aesthetic while preserving coherent layouts and strong visual quality.

4.4 Human Preference

Following standard practice in text-to-image generation, we conducted a user study on Amazon Mechanical Turk (AMT) to quantify user preferences against the baselines.[4] For each prompt, a participant evaluated a randomized left/right pair containing our output alongside one of the baselines: DemoFusion, SEGA, or ResDiT. Participants answered three independent questions evaluating which image was (i) more detailed and visually complex, (ii) a better match to the prompt, and (iii) more artistically appealing. In total, the study collected 2,700 pairwise preference judgments across a diverse set of 120 prompts.

Table 3 summarizes the results, with light-green cells marking significant preferences. (i) Users prefer JoLT for detail and complexity against every baseline. (ii) For prompt adherence, JoLT is preferred over SEGA and ResDiT, with no significant preference against DemoFusion, consistent with our focus on complexity rather than maximal prompt adherence. (iii) Artistic appeal follows the same pattern. As seen in Fig. 3, JoLT nevertheless produces striking visuals with denser detail and more complex structure than DemoFusion. Together with Tab. 1, these results indicate a distinct aesthetic and added complexity without measured quality degradation, rather than universal aesthetic superiority. Complementary automatic preference results are provided in Sec. B.

[4] We restricted participation to AMT workers with the Master qualification and approval rates above 95%.

Table 3: Two-alternative forced-choice user study. Entries report the percentage of trials preferring our method as the mean ± the 95% confidence-interval half-width. A light-green background marks a preference significantly above the 50% chance level (95% confidence interval entirely above 50%).

Baseline	More detailed / complex (95% CI)	Better prompt match (95% CI)	More artistically appealing (95% CI)
SEGA	$77.3\% \pm 5.1\%$	$69.0\% \pm 5.4\%$	$68.0\% \pm 5.5\%$
DemoFusion	$76.3\% \pm 5.1\%$	$48.5\% \pm 5.6\%$	$46.3\% \pm 5.7\%$
ResDiT	$76.0\% \pm 5.1\%$	$68.3\% \pm 5.5\%$	$71.0\% \pm 5.4\%$

4.5 Ablation Studies

We ablate the contribution of each component. First, we vary the constant HR→LR conditioning strength α between runs. Next, we compare static and jointly denoised LR-derived priors and isolate the effect of MD aggregation. Finally, we vary the detail prompt $\bar{c}$ to demonstrate controls beyond increasing complexity. The α sweep uses a matched subset of 25 prompts, while the component ablation uses all 200 evaluation prompts.

HR→LR Conditioning Strength We evaluate the effect of varying the constant α between runs (Eq. (6)). When $\alpha = 0$, the HR stream still receives information from the LR stream, while the LR stream remains independent of the HR stream. Across the full sweep in Fig. 4, IC9600 increases by 5.5% and DCT by 10.2%. Both metrics remain near their initial levels through $\alpha = 0.5$ before increasing more strongly at $\alpha = 0.75$ and $\alpha = 1$. We omit the text-alignment curve because LMM4LMM correspondence varies by less than one percentage point across the sweep. We retain $\alpha = 0.25$ as a conservative default while leaving headroom for users who prefer more aggressive detail. The matched examples in Fig. 5 likewise show that increasing α produces progressively more complex images, confirming its role as a controllable complexity parameter.

Component Ablation We evaluate four configurations to separate joint LR–HR denoising from the effect of MD aggregation. The *prior-only* variant is the base T2I model, without either component. The *static-prior* variant first generates an LR image and then keeps that image fixed as the in-context reference throughout MD sampling. The *joint-prior* variant jointly denoises a 1024^2 px LR prior and a single 2048^2 px HR window, updating the LR-derived in-context reference at every step but without MD. Finally, the full JoLT combines joint-prior conditioning with MD to produce a 4096^2 px canvas.

These four configurations are summarized in Fig. 6. Between the two configurations using MD at the same resolution, replacing the static prior with joint-prior conditioning increases IC9600 by 5.7%, LMM4LMM correspondence by 2.9%, and QIB Quality by 3.7%. Relative to joint-prior conditioning in the

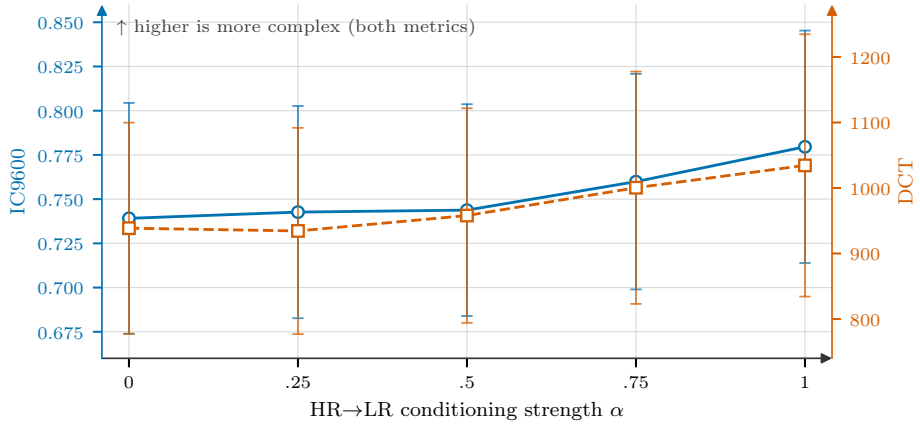


Fig. 4: Complexity response to HR→LR conditioning strength α ($n = 25$ matched prompts, tiled denoising at 4096^2 px final image). Both IC9600 and DCT energy metrics increase most strongly above $\alpha = 0.5$, confirming that HR→LR conditioning provides controllable complexity.

single-window setting, adding MD increases IC9600 by 5.1%, LMM4LMM correspondence by 3.1%, and QIB Quality by 3.2%. Because one window is limited to 2048^2 px, evaluation at 4096^2 px requires MD.

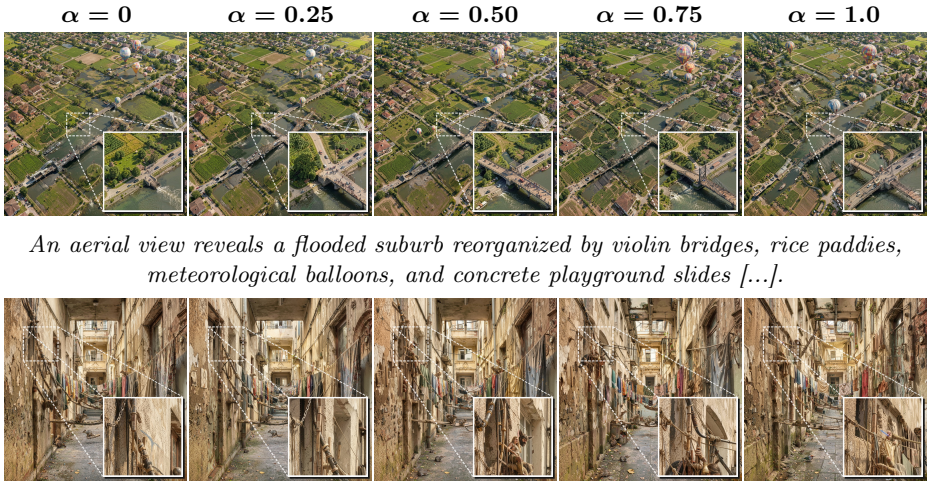


Fig. 5: Qualitative ablation of HR→LR conditioning strength α**.** Across both examples, increasing α produces a clear increase in the number and richness of visible details, especially in the magnified regions shown in the lower-right insets.

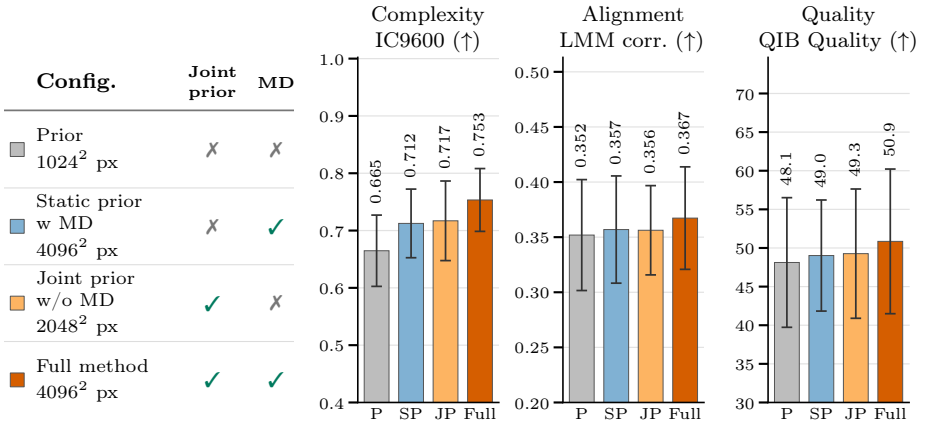
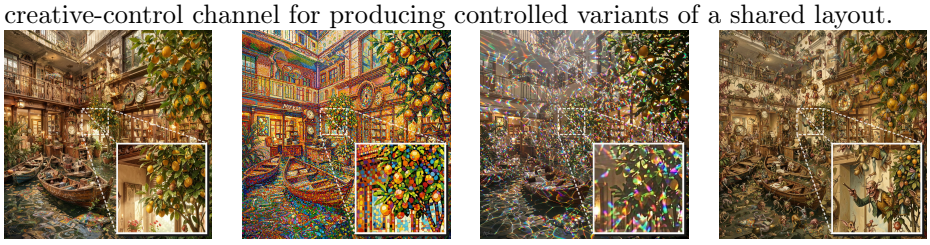


Fig. 6: Component ablation. Left: the four configurations and their native output resolutions; colors match the bars. Right: mean IC9600, LMM4LMM correspondence, and QIB Quality over 200 prompts; error bars show one standard deviation. Since one window is limited to 2048^2 px, 4096^2 px generation requires MD.

Detail-Prompt Ablation We test the detail prompt as a lightweight editing instruction while holding the master prompt and seed fixed. Figure 7 shows that the resulting image follows each requested detail or appearance treatment while preserving the same overall scene. This separation lets artists iterate on rendering style, surface texture, and lighting without rewriting the scene description or disrupting its composition. The detail prompt therefore serves as a compact creative-control channel for producing controlled variants of a shared layout.



"Enhance the image, add lots of details"

"pointillist style"

"Add iridescent caustic light reflections"

"Intricate grotesque imagery"

Fig. 7: Qualitative detail-prompt ablation. We hold the master prompt and seed fixed and vary the detail prompt shown below each image. The master prompt is *"A flooded hotel atrium contains a functioning print shop with reed boats, mechanical clocks, stained-glass fish, and citrus trees [...]."* All results are generated at 4096^2 px.

5 Limitations and Conclusion

Limitations. Our method relies on the potentially noisy one-shot estimate $\hat{x}_1$. We therefore evaluate it on a four-step model, whose estimate is already visually informative after one step. We also tested our method with longer sampling schedules, but these require an alternate clean-estimation approach, such as PiD [30], as well as substantially more compute. Our evaluation is also limited by the weak-to-moderate agreement among text-alignment metrics: across the nine metrics in Sec. C.4, their mean pairwise Spearman correlation is only $\rho \approx 0.25$, so no single score is a definitive proxy for text–image alignment and metric-specific gains should be read cautiously, ideally corroborated across metrics or with human judgment. We further hypothesize that the sheer density of detail JoLT introduces can itself reduce the readability of the full scene: beyond a certain point, additional local content crowds the global composition and makes the whole image harder to parse. We believe this effect partly underlies the modest text-alignment gap we observe alongside our large complexity gains. Future work should also evaluate on larger standardized benchmarks, such as the graph-based DPG-Bench [19] and art-focused GenAI-Bench [26].

Conclusion. We introduced JoLT, a two-stream generator that jointly denoises an LR layout trajectory and an HR detail trajectory, coupling them at every step instead of cascading from LR to HR. A global master prompt governs the overall composition while a detail prompt and a step-wise in-context prior drive local synthesis, and a single feedback strength α turns image complexity into a directly controllable parameter. On a distilled four-step FLUX.2 backbone with MultiDiffusion aggregation, JoLT generates images at 4096^2 px and above with substantially higher measured complexity than strong high-resolution baselines while remaining competitive in text alignment, aesthetics, and quality, and human evaluators consistently prefer it for visual detail. More broadly, our results suggest that treating high-resolution synthesis as two coupled trajectories, rather than a low-to-high cascade, is an effective route to dense, high-fidelity imagery. For artistic practice, this turns choices that vanilla T2I models fix implicitly into explicit controls: α acts as a dial over visual density, and the detail prompt restyles a fixed composition through language alone. At resolutions that withstand the close inspection a print invites, JoLT becomes a practical instrument for large-format artistic creation.

Acknowledgements

This work has been partially supported by the VISA DEEP chair (grant ANR-20-CHIA-0022), the PostGenAI@Paris cluster (grant ANR-23-IACL-0007; France 2030), and funded by the French National Research Agency (ANR) under the Renaissance project (grant ANR-23-CE23-0023; France 2030).

This project was provided with computing HPC & AI and storage resources by GENCI at IDRIS thanks to the grant 2025-AD011017131 on the supercomputer Jean Zay's H100 partition.

References

1. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Transactions on Computers **23**(1), 90–93 (1 1974). https://doi.org/10.1109/T-C.1974.223784, https://doi.org/10.1109/T-C.1974.223784, originally designated volume C-23
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: MultiDiffusion: Fusing diffusion paths for controlled image generation. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 1737–1752. PMLR (7 2023), https://proceedings.mlr.press/v202/bar-tal23a.html
3. Barbero Jiménez, Á.: Mixture of diffusers for scene composition and high resolution image generation (2023), https://arxiv.org/abs/2302.02412
4. Black Forest Labs: FLUX.1 (2024), https://github.com/black-forest-labs/flux, official model release and inference repository
5. Black Forest Labs: FLUX.2 [klein] (1 2026), https://github.com/black-forest-labs/flux2, official model release and inference repository; released 15 January 2026
6. Black Forest Labs, 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 (2025), https://arxiv.org/abs/2506.15742
7. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: HunyuanImage 3.0 technical report. arXiv preprint (2025), https://arxiv.org/abs/2509.23951
8. Chang, P., Tang, J., Gross, M., Azevedo, V.C.: How i warped your noise: A temporally-correlated noise prior for diffusion models. In: ICLR (2024), https://openreview.net/forum?id=pzElnMrgSD
9. Chen, Z., Du, Y., Wen, Z., Zhou, Y., Cui, C., Weng, Z., Tu, H., Wang, C., Tong, Z., Huang, Q., Chen, C., Ye, Q., Zhu, Z., Zhang, Y., Zhou, J., Zhao, Z., Rafailov, R., Finn, C., Yao, H.: MJ-Bench: Is your multimodal reward model really a good judge for text-to-image generation? arXiv preprint arXiv:2407.04842 (2024), https://arxiv.org/abs/2407.04842
10. Cho, J., Hu, Y., Baldridge, J.M., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In: ICLR (2024), https://openreview.net/forum?id=ITq4ZRUT4a
11. Coyote-A: Ultimate SD Upscale. GitHub repository (2023), https://github.com/Coyote-A/ultimate-upscale-for-automatic1111, accessed 12 August 2026
12. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: DemoFusion: Democratising high-resolution image generation with no $$$. In: CVPR. pp. 6159–6168 (2024). https://doi.org/10.1109/CVPR52733.2024.00589, https://doi.org/10.1109/CVPR52733.2024.00589
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: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) ICML. Proceedings of Machine Learning Research, vol. 235, pp. 12606–12633. PMLR (2024)

14. Feng, T., Zhai, Y., Yang, J., Liang, J., Fan, D.P., Zhang, J., Shao, L., Tao, D.: IC9600: A benchmark dataset for automatic image complexity assessment. IEEE TPAMI **45**(7), 8577–8593 (2023). https://doi.org/10.1109/TPAMI.2022.3232328, https://doi.org/10.1109/TPAMI.2022.3232328
15. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: ScaleCrafter: Tuning-free higher-resolution visual generation with diffusion models. In: ICLR (2024), https://openreview.net/forum?id=u48tHG5f66
16. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://doi.org/10.18653/v1/2021.emnlp-main.595
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) NeurIPS. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
19. Hu, X., Wang, R., Fang, Y., Fu, B., Cheng, P., Yu, G.: ELLA: Equip diffusion models with LLM for enhanced semantic alignment (2024), https://arxiv.org/abs/2403.05135
20. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-Context LoRA for diffusion transformers. arXiv preprint (2024), https://arxiv.org/abs/2410.23775
21. Hwang, Y., Lee, D., Min, K., Kang, T., Kim, Y., Jung, K.: Fooling the LVLM judges: Visual biases in LVLM-based evaluation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 23186–23205 (2025). https://doi.org/10.18653/v1/2025.emnlp-main.1182, https://doi.org/10.18653/v1/2025.emnlp-main.1182
22. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: MUSIQ: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (10 2021). https://doi.org/10.1109/ICCV48922.2021.00510, https://doi.org/10.1109/ICCV48922.2021.00510
23. Kim, J., Koo, J., Yeo, K., Sung, M.: SyncTweedies: A general generative framework based on synchronized diffusions. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) NeurIPS. vol. 37, pp. 95198–95237. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-3016, https://doi.org/10.52202/079017-3016
24. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS. vol. 36, pp. 36652–36663. Curran Associates, Inc. (2023). https://doi.org/10.52202/075280-1594, https://doi.org/10.52202/075280-1594
25. Lee, Y., Kim, K., Kim, H., Sung, M.: SyncDiffusion: Coherent montage via synchronized joint diffusions. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS. vol. 36, pp. 50648–50660. Curran Associates, Inc. (2023). https://doi.org/10.52202/075280-2203, https://doi.org/10.52202/075280-2203
26. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., Xia, X., Zhang, P., Neubig, G., Ramanan, D.: Evaluating and improving compositional text-to-visual genera-

tion. In: CVPRW. pp. 5290–5301 (2024). https://doi.org/10.1109/CVPRW63382.2024.00538, https://doi.org/10.1109/CVPRW63382.2024.00538
27. Li, N., Hu, G., Qiao, W., Ba, Y., Hong, Q., Shen, S., Wang, J., Zhou, F., Kang, J., Shang, X., et al.: Qwen-Image-Bench: From generation to creation in text-to-image evaluation. arXiv preprint (2026), https://arxiv.org/abs/2605.28091
28. Lin, Z., Lin, M., Zhao, M., Ji, R.: AccDiffusion: An accurate method for higher-resolution image generation. In: ECCV. pp. 38–53. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72658-3_3, https://doi.org/10.1007/978-3-031-72658-3_3
29. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV. pp. 366–384. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72673-6_20, https://doi.org/10.1007/978-3-031-72673-6_20
30. Lu, Y., Wu, Q., Wu, J.Z., Wang, Z., Ling, H., Fidler, S., Ren, X.: PiD: Fast and high-resolution latent decoding with pixel diffusion. arXiv preprint (2026), https://arxiv.org/abs/2605.23902
31. Ma, Y., Zhou, F., Yin, X., Cao, P., Dang, Y., Yin, J.: ResDiT: Evoking the intrinsic resolution scalability in diffusion transformers. In: CVPR. pp. 40612–40621 (6 2026)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023). https://doi.org/10.1109/ICCV51070.2023.00387, https://doi.org/10.1109/ICCV51070.2023.00387
33. Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.: Merging and splitting diffusion paths for semantically coherent panoramas. In: ECCV. pp. 234–251. Springer Nature Switzerland (2024). https://doi.org/10.1007/978-3-031-72986-7_14, https://doi.org/10.1007/978-3-031-72986-7_14
34. Qwen Team: Qwen3.6-27B: Flagship-level coding in a 27b dense model. Hugging Face model repository (4 2026), https://huggingface.co/Qwen/Qwen3.6-27B
35. 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.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
36. Rajabi, J., Shaban, K., Roohi, K., Lindell, D.B., Taati, B.: SEGA: Spectral-energy guided attention for resolution extrapolation in diffusion transformers. arXiv preprint (2026), https://arxiv.org/abs/2605.22668
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.org/10.1109/CVPR52688.2022.01042
38. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) NeurIPS. vol. 29, pp. 2234–2242. Curran Associates, Inc. (2016)
39. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
40. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)

41. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: OminiControl: Minimal and universal control for diffusion transformer. In: ICCV. pp. 14940–14950 (2025). https://doi.org/10.1109/ICCV51701.2025.01386, https://doi.org/10.1109/ICCV51701.2025.01386
42. Wallace, G.K.: The JPEG still picture compression standard. Communications of the ACM **34**(4), 30–44 (1991). https://doi.org/10.1145/103085.103089, https://doi.org/10.1145/103085.103089
43. Wang, H., Xiang, X., Fan, Y., Xue, J.H.: Customizing 360-degree panoramas through text-to-image diffusion models. In: WACV. pp. 4921–4931 (2024). https://doi.org/10.1109/WACV57701.2024.00486, https://doi.org/10.1109/WACV57701.2024.00486
44. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring CLIP for assessing the look and feel of images. In: AAAI. vol. 37, pp. 2555–2563. AAAI Press (2023). https://doi.org/10.1609/aaai.v37i2.25353, https://doi.org/10.1609/aaai.v37i2.25353
45. Wang, J., Duan, H., Zhao, Y., Wang, J., Zhai, G., Min, X.: LMM4LMM: Benchmarking and evaluating large-multimodal image generation with lmms. In: ICCV. pp. 17312–17323 (2025). https://doi.org/10.1109/ICCV51701.2025.01608, https://doi.org/10.1109/ICCV51701.2025.01608
46. Wang, J., Duan, H., Wang, J., Jia, Z., Zhai, G., Min, X.: TIT-Score: Evaluating long-prompt based text-to-image alignment via text-to-image-to-text consistency. arXiv preprint (2025), https://arxiv.org/abs/2510.02987
47. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint (2023), https://arxiv.org/abs/2306.09341
48. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: ImageReward: Learning and evaluating human preferences for text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) NeurIPS. vol. 36, pp. 15903–15935. Curran Associates, Inc. (2023). https://doi.org/10.52202/075280-0700, https://doi.org/10.52202/075280-0700
49. Ye, W., Ji, C., Chen, Z., Gao, J., Huang, X., Zhang, S.H., Ouyang, W., He, T., Zhao, C., Zhang, G.: DiffPano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In: NeurIPS. vol. 37, pp. 1304–1332. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-0041, https://doi.org/10.52202/079017-0041
50. Zhang, C., Wu, Q., Gambardella, C.C., Huang, X., Phung, D., Ouyang, W., Cai, J.: Taming Stable Diffusion for text to 360° panorama image generation. In: CVPR. pp. 6347–6357 (2024). https://doi.org/10.1109/CVPR52733.2024.00607, https://doi.org/10.1109/CVPR52733.2024.00607
51. Zhang, L.: ControlNet Tile. GitHub repository (2023), https://github.com/lllyasviel/ControlNet-v1-1-nightly, accessed 12 August 2026
52. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: DiffCollage: Parallel generation of large content with diffusion models. In: CVPR. pp. 10188–10198 (2023). https://doi.org/10.1109/CVPR52729.2023.00982, https://doi.org/10.1109/CVPR52729.2023.00982