LIGHTMGT: A COMPACT MASKED GENERATIVE TRANSFORMER FOR IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked generative transformers (MGTs) synthesize images by predicting discrete visual tokens in parallel, but existing models offer limited efficiency advantages over diffusion models. We therefore propose LightMGT, a compact masked generative transformer for image generation and editing. Its 148M backbone reduces model size through a compact double-stream design and factorized token prediction. We introduce MGT-OPD to combine generation and editing capabilities in one model by matching specialist teachers' token predictions on student-generated masked states. Antithetic Posterior Matching (APM) then distills iterative decoding into one step using a pair of masks and an updated auxiliary model, yielding LightMGT-Turbo. LightMGT achieves 0.70 on GenEval, 78.32 on DPG-Bench, and 6.49 on GEditBench-EN. On GenEval, it scores 0.70 versus 0.67 for the 12B-parameter FLUX.1-dev, while delivering $8.8\times$ higher throughput. With a single decoding step, about 0.235 s, LightMGT-Turbo achieves scores of 0.58 on GenEval, 71.30 on DPG-Bench, and 6.21 on GEditBench-EN. MGT-OPD combines text-to-image generation and editing capabilities. In the LightMGT experiments, it improves GenEval and GEditBench-EN by 16.7% and 4.5% relative to weight merging, and by 11.1% and 7.1% relative to joint training, respectively. APM also improves one-step distillation. On native MaskGIT ImageNet-256, it reduces FID by 1.2% and increases IS by 4.8% relative to Di[M]O. Together, this work contributes to the community a compact backbone, capability composition, and one-step distillation for image generation and editing.

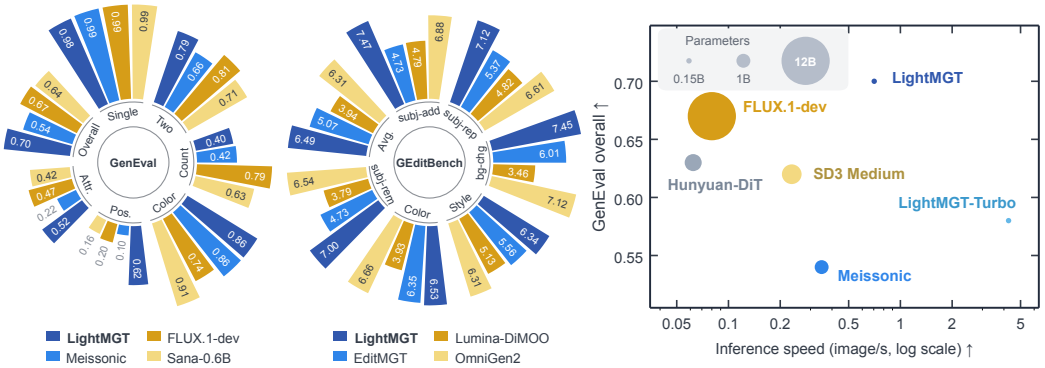


Figure 1 **Overview for LightMGT.** (a) Quality. GenEval and GEditBench metrics versus representative baselines. (b) Efficiency. GenEval overall versus end-to-end inference speed at 1024^2, with bubble area proportional to backbone parameters.

1 INTRODUCTION

Masked generative transformers (MGTs) synthesize images by predicting and committing discrete visual tokens in parallel at each decoding step (Chang et al., 2022; 2023; Li et al., 2023). Their iterative decoding process supports high-resolution image generation and can incorporate reference images as additional conditioning inputs for instruction-based editing (Bai et al., 2024; Chow et al., 2026; Brooks et al., 2023; Zhang et al., 2023). Yet, despite this flexible decoding mechanism, existing MGTs offer limited efficiency advantages over diffusion models in practice (Xie et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2024; Hu et al., 2024a). They still need higher-quality generation, a more compact backbone, a way to combine generation and editing in one model without losing either capability, and a better training method for deployment with fewer decoding steps.

These requirements concern different parts of the system. For model efficiency, a smaller backbone lowers the parameter cost of each pass but still needs a strong conditional image prior (Wang et al., 2026). For capability composition, continuing generation training on editing data changes the conditioning and task distribution, so the resulting editor need not preserve the original generation behavior. For sampling efficiency, fewer decoding steps introduce a further approximation because the model has less opportunity to revise tokens, thus requiring a step distillation method that preserves image quality (Salimans & Ho, 2022; Zhu et al., 2025; Zhou et al., 2026a).

Thus, we propose LightMGT through three strategies: a compact backbone, MGT-OPD for capability composition, and Antithetic Posterior Matching (APM) for step distillation. With a size of only 148M, LightMGT uses a compact double-stream backbone and factorized visual-token prediction to improve model efficiency. Text-to-image training progresses from 256^2 to 512^2 and then 1024^2 pixels using open-source 132.5M text-to-image samples and 14.0M image-editing pairs.

Jointly training text-to-image generation and image editing in a single model can cause the two capabilities to interfere (Zhou et al., 2026b). We therefore propose MGT-OPD to combine them in one student. For each student-generated masked state, the corresponding frozen specialist provides token predictions. Matching these predictions at every masked position supervises the student's decisions during decoding rather than only its final images as shown in Figure 2. To obtain the two specialists, we first train a text-to-image model through the 1024^2-resolution stage, yielding LightMGT-T2I. We then continue training this checkpoint on editing data alone to obtain LightMGT-Edit. Finally, we use MGT-OPD to combine generation and editing capabilities.

After capability composition, APM addresses the remaining cost of iterative sampling. APM uses a pair of masks for each generated sample and updates an auxiliary model before computing the generator update. Querying the updated auxiliary model again on the same masked inputs keeps that update current. The resulting LightMGT variant is named LightMGT-Turbo; neither the teacher nor the auxiliary model is needed at inference.

Extensive experiments demonstrate LightMGT's performance and the effectiveness of MGT-OPD and APM. LightMGT achieves 0.70 on GenEval (Ghosh et al., 2023), 78.32 on DPG-Bench, and 6.49 on GEditBench-EN. On GenEval, it scores 0.70 against 0.67 for the 12B-parameter FLUX.1-dev, with $8.8\times$ higher measured throughput. LightMGT-Turbo achieves 0.58 on GenEval, 71.30 on DPG-Bench, and 6.21 on GEditBench-EN with a single decoding step and 0.235 seconds per image. MGT-OPD combines text-to-image generation and editing. In the LightMGT family, it improves GenEval and GEditBench-EN by 16.7% and 4.5% relative to weight merging and by 11.1% and 7.1% relative to joint training. A separate composition study using Meissonic and EditMGT further tests the method. Using a 512^2-pixel LightMGT-T2I teacher, MGT-OPD also transfers generation capability to a Monetico student with a different visual tokenizer. On GenEval, the student's score rises by 23.3%. APM also improves step distillation and image quality. On MaskGIT ImageNet-256, APM reduces FID by 1.2%, and increases IS by 4.8% relative to Di[M]O (Zhu et al., 2025).

In conclusion, our contributions are:

1. We introduce LightMGT, a 148M compact double-stream backbone to improve model efficiency while supporting generation and editing.
2. We propose MGT-OPD for capability composition, matching specialist teachers' token predictions on student-generated masked states to combine generation and editing.
3. We propose APM for step distillation, using paired masks and an updated auxiliary model to reduce iterative decoding to one step.

2 Related Work

Masked Generative Transformers. MaskGIT (Chang et al., 2022) and related models (Chang et al., 2023; Li et al., 2023; Vaswani et al., 2017) predict discrete visual tokens in parallel and progressively commit confident positions. Meissonic (Bai et al., 2024) applies masked generation to high-resolution text-to-image synthesis. EditMGT (Chow et al., 2026) extends MGTs to

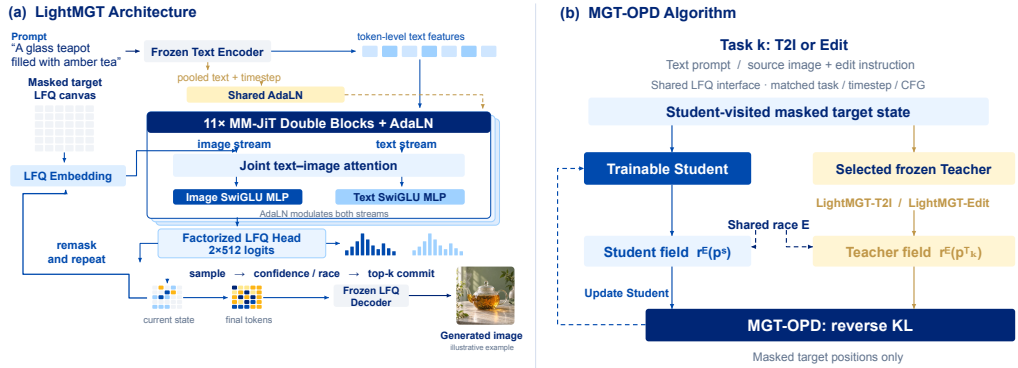


Figure 2 **LightMGT combines a compact backbone with capability composition.** (a) LightMGT uses a double-stream MM-JiT framework to predict factorized visual tokens. (b) MGT-OPD matches the selected frozen teacher on student-generated masked states during training.

instruction-based editing, a task studied in prior work (Brooks et al., 2023; Zhang et al., 2023; Zhao et al., 2024; Liu et al., 2025). LightMGT further improves MGT performance and efficiency, achieving FLUX.1-dev-level performance with a compact backbone and higher measured throughput. To our knowledge, it is the first MGT to compose separately trained text-to-image generation and instruction-based editing capabilities in a single model.

On-Policy Distillation. On-policy distillation trains a student with teacher guidance on states from its own trajectories and is often used to absorb or compose capabilities. In language models the teacher evaluates student-generated sequences so the student can learn from its own mistakes (Agarwal et al., 2024). In diffusion and flow-matching models teacher guidance can likewise supervise student-visited states. DanceOPD uses specialist flow fields on these states to compose text-to-image generation and instruction-based editing (Zhou et al., 2026b). However on-policy distillation remains underexplored for MGTs. MGT-OPD is the first method to use on-policy distillation for capability composition of text-to-image generation and instruction-based editing within an MGT.

Step Distillation. Step distillation reduces iterative sampling to one or a few inference steps (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023; Yin et al., 2024). Di[M]O uses token-level distribution matching and an auxiliary model to distill masked generators into one step, but its generator update uses auxiliary predictions before the auxiliary model learns from the current samples (Zhu et al., 2025). APM instead updates and re-queries the auxiliary model on the same student-generated samples, while paired masks cover complementary masking depths. This improves one-step image quality without increasing inference steps.

3 Approach

To improve the efficiency and performance of MGTs, we develop LightMGT in three parts. Firstly, we introduce a compact backbone to reduce model cost in Sec. 3.1. Secondly, we present MGT-OPD for capability composition to strengthen generation and editing in one model in Sec. 3.2. Finally, we introduce APM for step distillation to get an efficient one-step sampling model in Sec. 3.3.

3.1 Compact LightMGT Backbone

To improve LightMGT's efficiency while supporting both generation and editing, we use a factorized visual-token representation and a compact backbone to reduce output-head and decoding costs.

Visual Tokens. LightMGT uses the Open-MAGVIT2 LFQ tokenizer (Luo et al., 2024) to encode each image into a grid of discrete visual tokens with $16\times$ spatial downsampling. Each token is an 18-bit code. We predict its two 9-bit factors with separate 512-way categorical heads rather than a flat 2^{18}-way classifier. This factorization retains the full code space while reducing output-head cost. Both generation and editing use the same visual-token interface.

Compact Backbone. We build LightMGT as a 148M masked generative transformer with 11 plain MM-JiT double-stream blocks (Wang et al., 2026) and hidden width 768. We use AdaLN (Peebles & Xie, 2023) to combine the timestep and pooled text condition, while joint attention uses token-level features from a frozen text encoder. We assign a learned mask token to padded text positions

and omit the text preamble and GLA blocks (Yang et al., 2023). These choices reduce the cost of each decoding pass while retaining conditioning for generation and editing.

Iterative decoding. LightMGT follow MGT's decoding strategy, where generation starts from an all-mask target canvas. At each step, the model predicts categorical proposals at active positions, ranks their confidence, and commits a subset according to the MaskGIT schedule (Chang et al., 2022; 2023). The frozen LFQ decoder converts the completed canvas into an image. Editing uses the same target-token process conditioned on a clean reference stream and an instruction.

Objective and optimization. For a target sequence of N tokens, sample $u \sim \mathcal{U}(0,1)$ and set the masking ratio to $\rho = \max(10^{-4}, 2\arccos(u)/\pi)$. Let M denote the masked target positions, $\tilde{z}$ the resulting masked target sequence, and $\tau = 1{,}000|M|/N$ the timestep derived from the realized masked fraction. The text condition c is a prompt for text-to-image generation or an editing instruction. Editing additionally conditions on reference tokens r encoded from the original image; text-to-image generation uses no reference stream. Let θ denote LightMGT's parameters, $z_{i,g}$ the ground-truth value of LFQ factor g at target position i, and $p_{\theta,g}$ LightMGT's predicted categorical distribution for that factor. With two 512-way LFQ factors and label smoothing $\epsilon = 0.1$, the shared masked-token objective is:

$$\mathcal{L}_{\text{train}} = \frac{1}{2|M|} \sum_{g=1}^{2} \sum_{i \in M} \text{CE}_{\epsilon}(z_{i,g}, p_{\theta,g}(\cdot \mid \tilde{z}, r, c, \tau)), \tag{1}$$

The loss averages over both LFQ factors and all masked target positions.

3.2 Capability Composition with MGT-OPD

To improve a unified model's text-to-image generation and editing capabilities, we use MGT-OPD to compose two pretrained task specialists. It matches each specialist's token predictions on masked states generated by the student. At each decoding step, a masked decoder predicts tokens at many positions but commits only some. Training on final images or committed tokens alone does not directly supervise every prediction made before commitment. We therefore match the teacher and student distributions at all masked positions.

Student-generated masked states. Let $k \in \{\text{T2I}, \text{Edit}\}$ select the task, let S_θ denote the trainable student with parameters θ, and let T_k denote the corresponding frozen specialist model. The student shares the specialists' backbone and LFQ interface; its composed checkpoint is LightMGT. The condition c_k is a text prompt for T2I and a reference image with an instruction for editing. Let t index decoding steps. We sample a non-terminal masked target canvas s_t^S from the student-state distribution $d_{\theta,k}^S$ induced by S_θ under task k (Hinton et al., 2015; Agarwal et al., 2024). Clean reference tokens remain conditioning rather than prediction targets. The selected teacher is evaluated on the same canvas under the same task condition, timestep, and guidance convention.

Token-level matching with shared randomness. Let $p_M(a \mid s, i)$ denote the categorical token prediction of model M at masked position i, with the shared task condition implicit. Draw $E_{i,a} \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$ once for both models and define

$$\pi_M^E(a \mid s, i) = \text{softmax}_a(\log p_M(a \mid s, i) - \log E_{i,a}), \qquad M \in \{S, T_k\}. \tag{2}$$

Applying the same exponential-race noise to both models couples their token predictions under shared sampling randomness (Kool et al., 2019). The student minimizes

$$\mathcal{L}_{\text{OPD}} = \mathbb{E}_{k,\, s_t^S \sim d_{\theta,k}^S,\, E} \left[\sum_{i \in \text{MASK}(s_t^S)} \text{KL}\big(\pi_S^E(\cdot \mid s_t^S, i) \,\big\|\, \text{sg}\ \pi_{T_k}^E(\cdot \mid s_t^S, i)\big) \right], \tag{3}$$

where $\text{MASK}(s_t^S)$ contains the masked target positions and sg stops gradients through the frozen teacher. The loss covers all active positions, while subsequent states still follow the student's native scheduler. Updating the student-state distribution during training keeps the supervision tied to the decisions made by the current model. The objective matches token predictions at sampled masked states; it does not directly optimize the distribution of completed images.

Extension to different tokenizers. The LightMGT specialists share their state and action spaces. When a teacher and student use different tokenizers, a contextual bridge maps the student canvas

to a teacher canvas $\tilde{s}_t^T$. For a factorized LFQ teacher, let $a_1, a_2 \in \{0, \ldots, 511\}$ denote the values of its two 9-bit categorical factors. Its joint token prediction is transported back into the student's action space: $p_T^{\mathrm{LFQ}}(a_1, a_2 \mid \tilde{s}_t^T) \xrightarrow{\mathcal{T}_{\mathrm{ctx}}} q_T^S(\cdot \mid s_t^S)$. Here, $\mathcal{T}_{\mathrm{ctx}}$ denotes the contextual transport of the teacher distribution, and q_T^S is the resulting categorical distribution over student-token actions. This transport must preserve the joint relation between the two categorical factors. Token-ID matching or rank pairing does not provide that correspondence. The transported teacher distribution replaces the original teacher distribution in the student-space objective.

3.3 One-Step Distillation with APM

To improve inference efficiency after capability composition, APM distills LightMGT's iterative decoding into one step, yielding LightMGT-Turbo. Let T_{APM} denote the frozen iterative LightMGT teacher, G the one-step generator, and F_ϕ the auxiliary model with current parameters ϕ. Di[M]O's auxiliary model tracks the changing generator distribution, so predictions produced before its latest update can become stale (Zhu et al., 2025; Zhou et al., 2026a). APM addresses this by updating the auxiliary model and then querying it again on the same masked inputs before updating the generator. Let K_{cat} denote the number of categories in the categorical interface being distilled. For LightMGT's LFQ interface, $K_{\mathrm{cat}} = 512$ for each factor, and the factor-wise losses and update directions are summed across the two factors.

Generator outputs and paired masks. For a batch of B examples, a single generator forward from a hybrid noisy and masked initial canvas produces logits $\ell_{b,i}^G$ and categorical probabilities $p_{G,b,i} = \mathrm{softmax}(\ell_{b,i}^G)$ for example b and target position i. With task conditioning implicit, APM constructs two masked versions c_b^+ and c_b^- of each generated sample by drawing visible tokens from the stopped generator distribution. For each example b, draw a mask quantile $U_b \sim \mathrm{Uniform}(0, 1)$ and assign each target position i a random permutation rank $R_{b,i} \in \{0, \ldots, N-1\}$, where N is the target-token count. Applying the native masking schedule to $(U_b, R_{b,i})$ and its elementwise complement $(1-U_b, N-1-R_{b,i})$ determines the masked sets $\mathcal{M}_b^+$ and $\mathcal{M}_b^-$. Each context retains the native mask marginal, while the pair covers different masking depths. The frozen teacher and current auxiliary model predictions on context c_b^σ at position i are denoted $p^\sigma_{T_{\mathrm{APM}},b,i}$ and $p^\sigma_{F_\phi,b,i}$, respectively, with $\sigma \in \{+, -\}$.

Update and re-query the auxiliary model. For label-smoothing rate ϵ, define the stopped target $\bar{p}_{G,b,i} = (1-\epsilon)\,\mathrm{sg}\, p_{G,b,i} + \frac{\epsilon}{K_{\mathrm{cat}}} \mathbf{1}_{K_{\mathrm{cat}}}$, where $\mathbf{1}_{K_{\mathrm{cat}}}$ is the all-ones vector of length K_{cat}. The auxiliary loss averages the two separately normalized masked-context losses:

$$\mathcal{L}_F = \frac{1}{2} \sum_{\sigma \in \{+,-\}} \frac{\sum_{b=1}^B \sum_{i \in \mathcal{M}_b^\sigma} \mathrm{CE}\left(\bar{p}_{G,b,i}, p^\sigma_{F_\phi,b,i}\right)}{\sum_{b=1}^B |\mathcal{M}_b^\sigma|}. \tag{4}$$

One auxiliary optimizer update from ϕ gives parameters ϕ^+. We then query F_{ϕ^+} again on the same contexts and labels. Restoring the random-number state from before the first query gives both evaluations the same dropout realization. The second prediction therefore reflects the auxiliary update rather than different dropout noise. The generator update does not backpropagate through the auxiliary update or second query.

Generator update. For each masked position and categorical interface, the teacher prediction and the updated auxiliary model's prediction define the following logit-level update direction:

$$\begin{aligned} \Delta &= \log p_F - \log p_T, \\ v_\beta(p_T, p_F) &= (1-\beta)(p_F - p_T) + \beta\, p_F \odot (\Delta - \langle p_F, \Delta \rangle \mathbf{1}_{K_{\mathrm{cat}}}), \end{aligned} \tag{5}$$

where p_T and p_F denote categorical distributions over the K_{cat} categories, $\odot$ denotes elementwise multiplication, and $\langle \cdot, \cdot \rangle$ is the inner product over categories. The scalar β controls the relative weighting of the two terms. Combining the updates from the two masks gives

$$\bar{v}_{b,i} = \frac{1}{2} \sum_{\sigma \in \{+,-\}} \mathbf{1}_{i \in \mathcal{M}_b^\sigma}\, v_\beta\left(p^\sigma_{T_{\mathrm{APM}},b,i}, p^\sigma_{F_{\phi^+},b,i}\right), \qquad \mathcal{L}_G = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^N \left\langle \ell_{b,i}^G, \mathrm{sg}\, \bar{v}_{b,i} \right\rangle. \tag{6}$$

The generator is updated through $\mathcal{L}_G$ with this direction held fixed. For the LFQ interface, the losses and update directions are computed for each categorical factor and summed, avoiding a flat

Table 1 **Main results for text-to-image generation on GenEval and DPG-Bench**. LightMGT achieves similar performance to 12B FLUX.1-dev with only 148M.

		GenEval							DPG-Bench					
Model	Params.	Single obj.	Two obj.	Counting	Colors	Position	Attr.	Overall	Global	Entity	Attribute	Relation	Other	Overall
• Stable Diffusion v1.5 (Rombach et al., 2022)	860M	0.97	0.38	0.35	0.76	0.04	0.06	0.43	66.26	64.58	51.57	40.46	21.60	52.90
• Stable Diffusion v2.1 (Rombach et al., 2022)	865M	0.98	0.51	0.44	0.85	0.07	0.17	0.50	67.48	74.13	60.16	51.52	26.00	62.23
• PixArt-α (Chen et al., 2024b)	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48	75.68	75.92	64.68	54.92	28.00	65.52
• Stable Diffusion XL (Podell et al., 2024)	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55	70.52	78.90	67.15	59.43	21.60	68.06
• PixArt-Σ (Chen et al., 2024a)	0.6B	0.98	0.59	0.50	0.80	0.10	0.15	0.52	76.60	86.27	79.14	70.92	37.20	77.67
• LUMINA-Next (Zhuo et al., 2024)	2.0B	0.92	0.46	0.48	0.70	0.09	0.13	0.46	75.38	82.10	71.75	63.36	31.60	71.51
• Playground v2.5 (Li et al., 2024a)	2.6B	0.98	0.77	0.52	0.84	0.11	0.17	0.56	73.86	78.84	68.23	59.85	32.00	68.74
• Stable Diffusion 3 Medium (Esser et al., 2024)	2.0B	0.98	0.74	0.63	0.67	0.34	0.36	0.62	76.90	89.89	83.86	75.63	53.60	82.28
• Hunyuan-DiT (Li et al., 2024b)	1.5B	0.97	0.77	0.71	0.88	0.13	0.30	0.63	68.39	84.00	75.40	68.30	44.00	74.66
• Meissonic (Bai et al., 2024; Zhu et al., 2025)	1B	0.99	0.66	0.42	0.86	0.10	0.22	0.54	68.69	75.35	62.81	54.92	31.20	64.44
• IF-XL	5.5B	0.97	0.74	0.66	0.81	0.13	0.35	0.61	66.87	77.05	67.61	57.27	41.60	67.38
• Sana-0.6B (Xie et al., 2024)	0.6B	0.99	0.71	0.63	0.91	0.16	0.42	0.64	76.90	89.50	84.44	77.09	46.80	82.12
• Sana-1.6B (Xie et al., 2024)	1.6B	0.99	0.77	0.62	0.88	0.21	0.47	0.66	61.09	78.35	68.82	63.56	31.60	67.55
• FLUX.1-dev (Labs et al., 2025)	12B	0.99	0.81	0.79	0.74	0.20	0.47	0.67	76.60	89.60	84.64	77.67	66.00	82.56
• **LightMGT**	148M	0.98	0.79	0.40	0.86	**0.62**	**0.52**	**0.70**	74.16	86.66	81.33	72.43	41.20	**78.32**
• **LightMGT-Turbo**	148M	0.83	0.71	0.39	0.66	0.54	0.38	**0.58**	69.91	80.68	73.78	65.14	32.00	**71.30**

Table 2 **Main results for image editing generation on GEditBench-EN**. LightMGT achieves the best overall score and leads on multiple subtasks, even outperforming the 12B model by 4%.

		GEditBench-EN											
Model	Params.	Subj-Add	Subj-Rep	Bg-Chg	Style-Chg	Color-Alt	Subj-Rem	Mat-Alt	Motion-Chg	PS-Human	Text-Chg	Tone-Transfer	Avg
• OmniGen	3.8B	6.33	5.34	5.23	4.88	5.93	6.35	5.44	3.12	3.17	4.31	4.96	5.01
• OmniGen2	4B	6.88	6.61	7.12	6.31	6.66	6.54	5.57	7.42	5.79	4.57	5.94	6.31
• VAREdit-8B	8B	5.86	6.67	6.77	6.46	6.64	7.29	5.40	3.33	4.20	3.87	6.54	5.73
• UltraEdit-SD3	2B	5.06	5.79	5.83	5.73	5.51	3.15	5.86	3.55	5.00	2.24	5.45	4.83
• GoT-6B	6B	5.77	5.65	4.11	4.72	5.75	4.59	3.04	1.71	2.69	1.16	4.24	3.95
• MagicBrush	1B	5.53	5.10	6.17	4.10	5.41	4.13	4.75	1.55	2.90	1.33	5.07	4.19
• AnyEdit	1B	3.72	3.23	4.31	1.95	4.25	3.75	2.64	0.67	1.90	0.77	4.21	2.85
• Instruct-Pix2Pix	1B	3.07	3.48	3.94	4.39	5.40	1.50	3.52	1.27	2.62	1.13	5.10	3.22
• Lumina-DiMOO	8B	4.79	4.82	3.46	5.13	3.93	3.79	3.18	3.52	4.70	2.36	3.67	3.94
• FLUX.1 Kontext-dev	12B	6.95	6.13	7.06	5.55	7.03	6.76	5.52	5.62	4.68	6.10	7.48	6.26
• EditMGT	960M	4.73	5.37	6.01	5.56	6.35	4.73	5.59	4.25	5.26	1.51	6.47	5.07
• **LightMGT**	148M	**7.47**	7.12	**7.45**	6.34	6.53	7.00	**6.40**	5.98	6.19	3.89	6.96	**6.49**
• **LightMGT-Turbo**	148M	7.35	6.55	7.36	6.06	6.15	6.37	6.18	6.17	6.20	3.58	6.29	**6.21**

joint-vocabulary tensor. This alternating semigradient is an approximate update rather than an exact gradient of the full generated-image divergence. The second auxiliary-model query adds training compute; inference discards both the teacher and auxiliary model and uses one generator forward. Experimental settings are given in Appendix D.

4 Experiments

In this section, we evaluate the compact backbone and overall generation and editing performance in Sec. 4.1, capability composition in Sec. 4.2, and step distillation in Sec. 4.3. Experimental settings, evaluation protocols, and throughput measurements are provided in Appendix B.

4.1 Main results

Strong text-to-image ability. With only 148M backbone parameters, LightMGT achieves 0.70 on GenEval and 78.32 on DPG-Bench. Relative to 12B FLUX.1-dev, LightMGT uses 81.1× fewer parameters and improves GenEval from 0.67 to 0.70, a relative gain of 4.5%, while delivering 8.8× higher measured throughput. Relative to Meissonic, LightMGT improves GenEval from 0.54 to 0.70, a relative gain of 29.6%, and DPG-Bench from 64.44 to 78.32, a relative gain of 21.5%. It also leads the listed models on GenEval position and attribute binding, scoring 0.62 and 0.52; these are 82.4% and 10.6% above the next-best scores of 0.34 and 0.47, respectively.

Strong editing performance. LightMGT scores 6.49 overall on GEditBench-EN, a 28.0% relative improvement over EditMGT. It also exceeds 12B FLUX.1 Kontext-dev, whose score is 6.26, by 3.7%. LightMGT leads the listed methods on 5 subtasks. Compared with the second-best methods, LightMGT achieves relative improvements of 7.5%, 6.7%, 4.6%, 14.9%, and 6.9% on subject addition, subject replacement, background change, material alteration, and human, respectively.

Qualitative results. Figure 3 compares LightMGT with the listed baselines on the same generation prompts and editing instructions. For the exact-count prompt, *Four yellow slippers and two burgundy bow ties on a beige carpet*, LightMGT renders four separately recognizable slippers and two distinct burgundy bow ties. Their clear separation makes both the object count and color binding

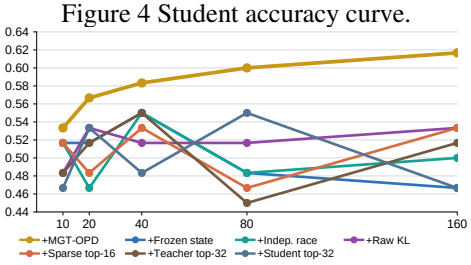
Table 3 MGT-OPD improves performance in two composition settings.

Method	GEditBench-EN												GenEval						
	Subj-Add	Subj-Rep	Bg-Chg	Style-Chg	Color-Alt	Subj-Rem	Mat-Alt	Motion-Chg	PS-Human	Text-Chg	Tone-Transfer	Avg	Sing	Two	Cnt	Col	Pos	C-Attr	Overall
A. Meissonic + EditMGT																			
Meissonic	0.65	1.24	1.89	1.36	0.93	0.10	2.57	1.23	0.86	0.08	0.46	1.03	0.99	0.48	0.38	0.84	0.15	0.20	0.51
EditMGT	4.73	5.37	6.01	5.56	6.35	4.73	5.59	4.25	5.26	1.51	6.47	5.07	0.69	0.26	0.16	0.55	0.16	0.06	0.31
Weight Merge	4.92	4.99	6.26	5.54	6.17	4.18	5.40	4.67	5.07	1.57	6.45	5.02	0.86	0.25	0.18	0.72	0.04	0.17	0.37
MGT-OPD	4.95	5.04	6.50	5.51	6.72	5.69	5.72	4.74	5.38	1.62	6.37	**5.29**	0.76	0.45	0.15	0.73	0.10	0.13	**0.39**
B. LightMGT-T2I + LightMGT-Edit																			
LightMGT-T2I	0.43	1.04	1.31	3.59	1.00	1.97	2.37	3.47	4.15	0.50	3.62	2.13	0.93	0.86	0.52	0.84	0.64	0.55	0.73
LightMGT-Edit	7.47	7.14	7.55	6.45	6.93	7.11	6.36	6.68	6.65	3.67	7.06	6.64	0.85	0.62	0.33	0.69	0.43	0.24	0.53
Joint Training	6.87	6.27	6.81	5.71	5.87	6.32	6.37	6.40	5.96	3.60	6.43	6.06	0.91	0.77	0.35	0.72	0.58	0.43	0.63
Weight Merge	7.35	6.55	7.36	6.06	6.15	6.37	6.18	6.17	6.20	3.58	6.29	6.21	0.85	0.77	0.36	0.72	0.48	0.39	0.60
LightMGT via MGT-OPD	7.47	7.12	7.45	6.34	6.53	7.00	6.40	5.98	6.19	3.89	6.96	**6.49**	0.98	0.79	0.40	0.86	0.62	0.52	**0.70**

Table 4 Ablation results for MGT-OPD.

Method	Single	Two	Count	Color	Position	Attribute	Overall
Teacher: LightMGT-T2I at 512^2	1.00	0.90	0.50	0.90	0.30	0.60	0.7000
Student: Monetico	1.00	0.30	0.50	0.80	0.20	0.20	0.5000
+Frozen state	1.00	0.30	0.50	0.90	0.10	0.00	0.4667
+Indep. race	1.00	0.40	0.40	1.00	0.10	0.10	0.5000
+Raw KL	1.00	0.50	0.60	1.00	0.10	0.00	0.5333
+Sparse top-16	1.00	0.40	0.60	0.90	0.10	0.20	0.5333
+Teacher top-32	1.00	0.40	0.50	1.00	0.00	0.20	0.5167
+Student top-32	1.00	0.20	0.60	0.80	0.10	0.10	0.4667
+ MGT-OPD	1.00	0.40	0.80	0.90	0.40	0.20	**0.6167**

Figure 4 Student accuracy curve.



easy to verify, while several baseline outputs make the count or grouping less distinct. For the spatial prompt, *A green guitar against a black amp, with a microphone to its left*, LightMGT preserves all three objects and places the microphone clearly to the guitar's left. In the editing examples, LightMGT changes *Google* to *Goose* while retaining the search-logo layout, preserves the bench structure while giving it a marble surface, and changes the background to a beach without altering the subject's pose and clothing. These cases show that LightMGT follows the requested object, relation, and edit attributes while keeping the relevant scene structure intact.

4.2 Capability composition with MGT-OPD

MGT-OPD combines text-to-image and editing capabilities in one model. On the full GenEval and GEditBench-EN benchmarks, MGT-OPD improves an EditMGT-initialized student from 0.31 to 0.39 on GenEval, a relative gain of 25.8%, and from 5.07 to 5.29 on GEditBench-EN, a relative gain of 4.3%. It also exceeds weight merging on both benchmarks. On GenEval, it raises the score from 0.37 to 0.39, a relative gain of 5.4%. On GEditBench-EN, it raises the score from 5.02 to 5.29, also a relative gain of 5.4%. In the LightMGT setting, MGT-OPD reaches 0.70 on GenEval and 6.49 on GEditBench-EN, improving over weight merging by 16.7% and 4.5%, and over joint training by 11.1% and 7.1%, respectively. Training details are in Appendix B.5.

MGT-OPD transfers generation capability across different tokenizers. A 512^2-pixel LightMGT-T2I teacher provides token predictions to a Monetico student through a fixed contextual bridge. On the GenEval subst, MGT-OPD raises the student's score from 0.5000 to 0.6167, a relative gain of 23.3%. It also exceeds raw reverse KL and sparse top-16 supervision, which score 0.5333, by 15.6%. Its strongest category gains are counting, which rises from 0.50 to 0.80 for a relative gain of 60%, and position, which rises from 0.20 to 0.40 for a relative gain of 100%.

4.3 One-step distillation with APM

LightMGT-Turbo. LightMGT-Turbo scores 0.58 on GenEval and 71.30 on DPG-Bench. These results are respectively 7.4% and 10.6% above Meissonic, which scores 0.54 and 64.44. LightMGT-Turbo achieves 6.21 on GEditBench-EN, a 22.5% gain over EditMGT, which scores 5.07. Strong subtask scores include a GenEval position score of 0.54, which is 440% above Meissonic's 0.10, and attribute binding at 0.38, which is 72.7% higher. On GEditBench-EN, subject addition scores 7.35, motion change scores 6.17, and text change scores 3.58. These are gains of 55.4%, 45.2%, and 137.1% over EditMGT, respectively, for the same tasks. The one-step setting takes 0.235 seconds per image, corresponding to 6.05× the throughput of the 32-step setting.

MaskGIT in ImageNet-256. Under the common ImageNet-256 sampling convention and 50,000-sample endpoint evaluation, APM achieves FID 6.80 and IS 224.17. Compared with Di[M]O, it reduces FID by 1.2%, and raises IS by 4.8%. It also improves precision from 0.828 to 0.844, a relative gain of 1.9%; density from 1.255 to 1.313, a relative gain of 4.6%; and coverage from 0.967 to 0.978, a relative gain of 1.1%. On the separate fresh-start curves, each checkpoint uses 5,000

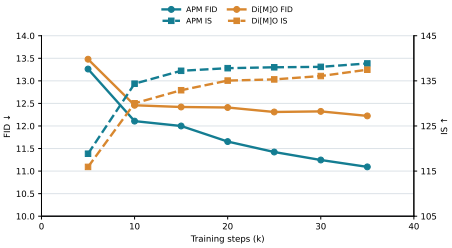


Figure 3 Qualitative comparisons between LightMGT and other image generation models

Table 5 Class-conditional ImageNet-256 results.

Family	Method	Steps ↓	FID ↓	IS ↑	Prec. ↑	Rec. ↑	Den. ↑	Cov. ↑
Teacher	MaskGIT	16	6.60	224.07	0.831	0.402	1.246	0.977
	MaskGIT	8	6.66	221.57	0.827	0.397	1.233	0.974
	MaskGIT	4	10.73	192.29	0.748	0.313	1.011	0.920
	MaskGIT	2	91.35	13.37	0.178	0.164	0.091	0.122
Sampler	θ-trapezoidal	64	6.7	–	–	–	–	–
	θ-trapezoidal	32	7.1	–	–	–	–	–
Distillation	di4c	4	6.79	209.2	–	–	–	–
	di4c-d	4	6.57	213.6	–	–	–	–
	Di[M]O	1	6.91	214.0	0.828	0.377	1.255	0.967
	• **APM**	1	**6.80**	**224.17**	0.844	0.380	1.313	0.978

Figure 5 Training curves on ImageNet-256.



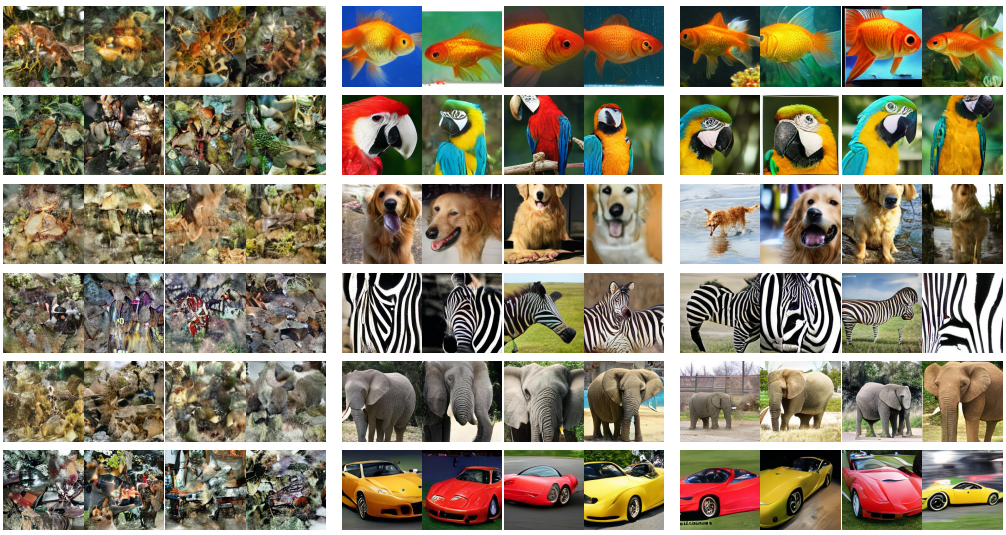


Figure 6 One-step ImageNet-256 samples from MaskGIT initialization, Di[M]O, and APM.

samples. At every matched update from 5,000 to 35,000, APM records lower FID and higher IS than Di[M]O.

Figure 6 gives concrete one-step ImageNet-256 cases. In the zebra row, APM produces consistently recognizable body silhouettes with coherent, repeated stripes; the corresponding Di[M]O examples vary more in crop and body completeness, while the MaskGIT-initialization samples are visibly fragmented. In the parrot row, APM samples retain distinct beaks, eyes, and blue-yellow plumage, with clearer subject structure than the initialization examples. These cases illustrate how APM improves object-level coherence while preserving class-specific detail.

5 Conclusion

We propose LightMGT, a 148M compact masked generative transformer for image generation and editing, with MGT-OPD for capability composition and APM for step distillation. LightMGT achieves 0.70 on GenEval, 78.32 on DPG-Bench, and 6.49 on GEditBench-EN; its GenEval score exceeds that of the 12B-parameter FLUX.1-dev with 8.8× higher throughput. MGT-OPD improves GenEval and GEditBench-EN by 16.7% and 4.5% over weight merging and by 11.1% and 7.1% over joint training. With one decoding step, LightMGT-Turbo achieves 0.58, 71.30, and 6.21 on the same benchmarks. On native MaskGIT ImageNet-256, APM reduces FID by 1.2%, and increases IS by 4.8% relative to Di[M]O. Together, this work contributes a compact backbone, capability composition, and one-step distillation for image generation and editing. We hope our work establish a strong baseline and will inform future research in the image generation community.

AI USE STATEMENT

We used AI tools to assist with language editing and polishing of the manuscript, including improving clarity, grammar, and readability. We did not use generative AI tools for research tasks requiring disclosure under the ICLR AI policy, such as generating research data, developing the proposed method, designing or conducting experiments, or interpreting results. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of this paper.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of new personal data. We train and evaluate LightMGT using existing image-generation and image-editing datasets; the provenance of the training data is documented in Appendix A. Like other image-generation and editing models, LightMGT may produce inaccurate, biased, harmful, or misleading images. Our evaluation focuses on image quality, text and image alignment, and image-editing instruction following, and should not be interpreted as a comprehensive assessment of safety, fairness, or privacy.

REPRODUCIBILITY STATEMENT

We describe LightMGT's architecture, training configurations, data provenance and preprocessing, random seeds, sampling settings, and evaluation protocols in the main paper and Appendices A, and B. Separate protocols for MGT-OPD and APM are reported in the corresponding appendix sections. We will release code and checkpoints.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315. IEEE, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3557–3567. IEEE, 2021.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. Getting it right: Improving spatial consistency in text-to-image models. In *European Conference on Computer Vision*, pp. 204–222. Springer, 2024.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. Scaleedit-12m: Scaling open-source image editing data generation via multi-agent framework. *arXiv preprint arXiv:2603.20644*, 2026.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-Σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024b.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025.

Wei Chow, Linfeng Li, Lingdong Kong, Zefeng Li, Qi Xu, Hang Song, Tian Ye, Xian Wang, Jinbin Bai, Shilin Xu, et al. Editmgt: unleashing potentials of masked generative transformers in image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38038–38048, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 63157–63186, 2026.

Junyao Gao, Sibo Liu, Jiaxing Li, Yanan Sun, Yuanpeng Tu, Fei Shen, Weidong Zhang, Cairong Zhao, and Jun Zhang. Megastyle: Constructing diverse and scalable style dataset via consistent text-to-image style mapping. *arXiv preprint arXiv:2604.08364*, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Aaron Gokaslan, A Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. Common canvas: Open diffusion models trained on creative-commons images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8250–8260. IEEE, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Dongting Hu, Jierun Chen, Xijie Huang, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. *arXiv preprint arXiv:2412.09619*, 2024a.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024b.

Bohan Jia, Wenxuan Huang, Yuntian Tang, Junbo Qiao, Jincheng Liao, Shaosheng Cao, Fei Zhao, Zhaopeng Feng, Zhouhong Gu, Zhenfei Yin, et al. Compbench: Benchmarking complex instruction-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1112–1122, 2026.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International conference on machine learning*, pp. 3499–3508. PMLR, 2019.

Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. Nohumansrequired: Autonomous high-quality image editing triplet mining. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6059–6068. IEEE, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Qingxiao Li, Zikai Wang, Qingli Wang, and Nan Xu. S1-omni-image: A unified model for scientific image understanding, generation, and editing. *arXiv preprint arXiv:2606.24441*, 2026.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2142–2152, 2023.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Delong Liu, Haotian Hou, Zhaohui Hou, Zhiyuan Huang, Shihao Han, Mingjie Zhan, Zhicheng Zhao, and Fei Su. Inter-edit: First benchmark for interactive instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yijun Liu, Jie Huang, Zeyue Xue, Yuming Li, Ruizhe He, Haoran Li, Shijia Ge, and Siming Fu. Hpsv3++: Scaling reward models across the full spectrum of diffusion model capabilities. *arXiv preprint arXiv:2606.14657*, 2026b.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7764–7772, 2026a.

Xu Ma, Yitian Zhang, Qihua Dong, and Yun Fu. Fine-t2i: An open, large-scale, and diverse dataset for high-quality t2i fine-tuning. *arXiv preprint arXiv:2602.09439*, 2026b.

Dongxing Mao, Yilin Wang, Linjie Li, Zhengyuan Yang, and Alex Jinpeng Wang. Textground4m: A prompt-aligned dataset for layout-aware text rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7918–7926, 2026.

OpenAI. Dall-e 3 system card, 2023. URL `https://cdn.openai.com/papers/dall-e-3.pdf`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37226–37235, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*, 2024.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in neural information processing systems*, 36:49659–49678, 2023.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 56783–56799, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025a.

Haoran Wang, Bo Zhao, Jinghui Wang, Hanzhang Wang, Huan Yang, Wei Ji, Hao Liu, and Xinyan Xiao. Sega: A stepwise evolution paradigm for content-aware layout generation with design prior. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19321–19330. IEEE, 2025b.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026. URL `https://peppaking8.github.io/#/post/minit2i`.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025c.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *International Conference on Learning Representations*, volume 2025, pp. 259–271, 2025.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, Ziyi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, et al. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable ai-generated image detection. *arXiv preprint arXiv:2601.19430*, 2026.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. Blip-3: A family of open large multimodal models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6183–6194, 2025.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Junyan Ye, Dongzhi Jiang, Zihao Wang, Leqi Zhu, Zhenghao Hu, Zilong Huang, Jun He, Zhiyuan Yan, Jinghua Yu, Hongsheng Li, et al. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *arXiv preprint arXiv:2508.09987*, 2025.

Keming Ye, Zhipeng Huang, Canmiao Fu, Qingyang Liu, Jiani Cai, Zheqi Lv, Chen Li, Jing Lyu, Zhou Zhao, and Shengyu Zhang. Unicedit-10m: A dataset and benchmark breaking the scale-quality barrier via unified verification for reasoning-enriched edits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37279–37289, 2026a.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Guoqing Zhang, Xingtong Ge, Lu Shi, Xin Zhang, Muqing Xue, Wanru Xu, Yigang Cen, and Yidong Li. Condition weaving meets expert modulation: Towards universal and controllable image generation. *arXiv preprint arXiv:2508.17364*, 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024.

Wei Zhou, Xiongwei Zhu, Lingdong Kong, Bo Chen, Lei Zhang, Yongyuan Liang, Xiaoxia Hou, Ye Tian, Xian Sun, Yingshuo Wang, et al. On-policy self-distillation in diffusion models. *arXiv preprint arXiv:2608.24646*, 2026a.

Wei Zhou, Xiongwei Zhu, Zelin Xu, Bo Dong, Lixue Gong, Yongyuan Liang, Meng Chu, Leigang Qu, Lingdong Kong, Wei Liu, et al. Danceopd: On-policy generative field distillation. *arXiv preprint arXiv:2606.27377*, 2026b.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Di[M]O: Distilling masked diffusion models into one-step generator. *arXiv preprint arXiv:2503.15457*, 2025.

Zihao Zhu, Kuan-Ru Huang, Zhaoming Xu, Renjie Li, Bo Wu, Ruizheng Bai, Mingyang Wu, Sayak Paul, and Zhengzhong Tu. 4klsdb: A large-scale dataset for 4k image restoration and generation. *arXiv preprint arXiv:2605.24762*, 2026.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024.

Kai Zou, Zhiwei Zhao, Bin Liu, and Nenghai Yu. Advancing aesthetic image generation via composition transfer. *International Journal of Computer Vision*, 134(5):252, 2026.

Table 6 LightMGT model and training configuration.

Setting	Value
Initialization	1024^2-pixel T2I step-5,000 EMA weights
Backbone	11 plain MM-JiT double-stream blocks with AdaLN
Hidden width / attention heads	768 / 12 heads of width 64
MLP expansion ratio	2.6875
Text preamble / GLA blocks	0 / 0
RoPE axis dimensions	$(8, 28, 28)$
Text encoder / maximum length	Frozen Qwen3-0.6B / 256 tokens
Text padding	Learned text mask token
Visual tokenizer	Frozen Open-MAGVIT2, 18-bit LFQ, $16\times$ downsampling
Output heads	Two 512-way categorical factors
Resolution	1024^2-pixel tier with aspect-ratio buckets
Data	13,991,529 single-reference Edit pairs
Task mixture	100% Edit examples
Optimizer	AdamW, $\beta = (0.9, 0.96)$
Peak learning rate / floor	10^{-5} / 10^{-6}
Weight decay / clipping norm	0.045 / 1.0
GPUs / batch per GPU	32 / 4
Gradient accumulation / batch	32 / 4,096
Optimizer-update budget	50,000
Target mask schedule	Arccos
Label smoothing	0.1
Text / reference dropout	0.1 / 0.1, sampled independently
Reference timestep	0 in every block
Precision / checkpointing	BF16 / enabled
EMA initialization / decay	Edit step 5,001 / 0.9999
Checkpoint interval	250 optimizer updates
Preview steps / text CFG	20 / 4.0

A Model Training

Model lineage. LightMGT-T2I is the 1024^2-pixel generation specialist. Edit-only continuation yields LightMGT-Edit; MGT-OPD composes these specialists into LightMGT, and APM yields LightMGT-Turbo. The variants share an architecture and LFQ interface. The 512^2-pixel transfer checkpoint and the Meissonic/EditMGT, Monetico, and native MaskGIT experiments remain separate. Table 6 lists optimization settings, including effective batch $4{,}096$ and a $50{,}000$-update budget (204.8M nominal pair exposures). Edit EMA starts from live weights at step $5{,}001$; the T2I EMA accumulator is not carried over.

Edit initialization and conditioning. Edit training starts from the 1024^2-pixel T2I EMA checkpoint at step $5{,}000$ with frozen Qwen3-0.6B and Open-MAGVIT2 LFQ. Masked target tokens precede clean reference tokens; only the target is predicted. The target uses the mask-ratio timestep, while reference tokens use timestep-zero AdaLN in every block. Three-axis RoPE has dimensions $(8, 28, 28)$. Text uses the final Qwen3 layer and last valid token, normalized to $\sqrt{1024}$, with a 256-token limit. Independent text and reference dropout of 0.1 trains the no-reference route.

Edit data and image preprocessing. Table 10 gives the single-reference/single-target manifest. We exclude multi-reference UniPic ($41{,}812$), MICON ($1{,}043$), and OpenGPT4o, whose image payload is unavailable. Missing images are dropped; missing instructions become empty strings. Each processed shard appears once in the manifest, without quotas or further deduplication. EXIF-corrected RGB images use aspect-ratio buckets around 1024^2 pixels, antialiased resizing, and a shared crop anchor. Frozen LFQ encodes pixels in $[-1, 1]$; a square 1024^2-pixel target has 64×64 tokens.

Training Data

All the data we used are shown in the Tables 9 and 10.

B EXPERIMENTAL DETAILS

The following settings and protocols distinguish local measurements from source-reported reference scores.

B.1 EVALUATION SCOPE AND REPORTING

We evaluate the compact backbone, capability composition, and step distillation separately. Tokenizer reconstructions and small-data architecture screens test LightMGT's compact design. GenEval (Ghosh et al., 2023) and DPG-Bench (Hu et al., 2024b) measure text-to-image alignment, while GEditBench-EN (Liu et al., 2025) measures instruction-based editing. We evaluate MGT-OPD on Meissonic/EditMGT (Bai et al., 2024; Chow et al., 2026), within the LightMGT specialist family, and through LightMGT-to-Monetico transfer. We evaluate APM on native MaskGIT (Chang et al., 2022) ImageNet-256. Training-data provenance and benchmark-associated sources are disclosed in Appendices A and A.

B.2 THROUGHPUT MEASUREMENT

Figure 1 reports single-image throughput on one A100-80GB at 1024^2. One warm-up precedes three timed pipeline calls that include prompt encoding, native sampling, and image decoding but exclude model loading and file output. The settings are LightMGT at 32 steps/CFG 16, a one-step LightMGT-base timing proxy at CFG 16, Meissonic at 48/9, FLUX.1-dev at 28/3.5, Hunyuan-DiT at 50/5, and Stable Diffusion 3 Medium at 28/7. The measured speeds are 0.704, 4.258, 0.348, 0.080, 0.062, and 0.233 images/s, respectively. Thus LightMGT has $8.8\times$ the measured throughput of FLUX.1-dev, while the one-step LightMGT-Turbo takes 0.235 seconds per image.

B.3 DPG-BENCH PROTOCOL

For each of 14 runnable public baselines, we generate one image per DPG-Bench prompt (1,065), using native settings and square resolution. GPT-5.5 judges all propositions in one high-detail yes/no request per image. L1 category scores use raw proposition accuracy; overall uses per-image accuracy after the released ELLA dependency zero-out. Each model has 14,392 resolved answers.

B.4 GEDITBENCH-EN PROTOCOL

Local GEditBench-EN results use `GPT-4o-2024-08-06` with fixed VIEScore prompting and aggregation, as most paper choose `GPT-4o-2024-08-06` for testing Liu et al. (2025); Chow et al. (2026).

B.5 MEISSONIC AND EDITMGT

For Table 3, MGT-OPD trains the public EditMGT student for 2,000 updates with frozen Meissonic T2I and reference-conditioned EditMGT teachers. It follows the student's 1024^2-pixel, 32-step, CFG-9 trajectory and matches teacher token distributions at all masked positions on student-generated states, using four reverse-KL terms with shared exponential-race noise. AdamW uses LR 10^{-5}, zero weight decay and clipping 1.0.

C ARCHITECTURE ABLATIONS

Given that training each architecture from scratch is impractical, and to enhance the efficiency of architecture design, we drew upon approaches from related works (Wang et al., 2026) and conducted small-scale experiments to finalize the architecture design, which was then used directly for training. Parameters count the trainable backbone. *Gen@u/Gen(fin.)* are 8-step generation-token accuracy at update u/final update; *acc≥.99/gen≥.99* mark the first update reaching 0.99.

Table 7 Tokenizer reconstruction on 6400 paired images at 256^2 and 1024^2 pixels.

Resolution	Tokenizer	MSE ↓	PSNR ↑	SSIM ↑	LPIPS-Alex ↓
256^2	• Open-MAGVIT2 LFQ	**0.007762**	**21.7885**	**0.7289**	**0.0925**
	Meissonic-VQ	0.008928	21.1444	0.6795	0.1202
1024^2	• Open-MAGVIT2 LFQ	**0.002893**	**26.6317**	**0.7809**	**0.0770**
	Meissonic-VQ	0.003537	25.5998	0.7497	0.0993

Table 8 Selected backbone and component ablations on 6400 images at 256^2 pixels after 2,000 steps.

Configuration	Params (M)	Gen(fin.)	gen≥.99 (step)	Train mem (GB)
Framework screen				
Hybrid + AdaLN	155.4	0.997	600	11.3
MM-DiT + AdaLN	171.9	0.995	600	11.0
MM-JiT without AdaLN	167.4	0.844	–	7.9
MM-JiT + AdaLN + two preamble blocks	164.8	0.997	600	8.3
MM-DiT without AdaLN	161.5	0.996	700	9.8
Component isolation relative to MM-JiT + AdaLN + two preamble blocks				
• 0 pre / AdaLN / MM-JiT	147.7	0.998	500	7.56
4 pre / AdaLN / MM-JiT	180.2	0.993	600	8.98
2 pre / AdaLN / FLUX	164.9	0.998	600	9.29
2 pre / no AdaLN / FLUX	154.4	0.990	1900	8.10
2 pre / no AdaLN / MM-JiT	154.4	0.880	–	7.33

C.1 Tokenizer Reconstruction Ablation

We compare Open-MAGVIT2 18-bit LFQ with Meissonic-VQ at matched 16× downsampling on 6400 images at 256^2 and 1024^2 pixels.

Deterministic center crops and bicubic resizing precede each tokenizer's native encode/decode path. We compare decoded serialized IDs against the same RGB target; LPIPS uses AlexNet, with 4,000 paired bootstrap resamples.

Findings. LFQ improves all four metrics at both resolutions in Table 7; paired 95% bootstrap intervals exclude zero. At 1024^2 pixels, PSNR rises by 1.032 dB, SSIM by 0.0313, and LPIPS falls by 0.0223.

Model cost. Two 512-way LFQ heads replace Meissonic-VQ's 8,192-way head, saving 11.8M backbone parameters. This reconstruction test does not measure downstream generation.

C.2 Text encoder ablation

On the intermediate pilot backbone with four early softmax-attention blocks, eight GLA blocks, and four standard softmax-attention blocks (88.4M parameters), frozen decoder LLMs, T5, and BERT use suitable pooling for 2,000 steps on 10 images at 256^2 pixels. GPU contention precludes throughput comparison. All encoders overfit; T5-base reaches gen≥.99 at step 200 and BERT at 500.

C.3 Backbone design ablation

Eight approximately 160M configurations compare AdaLN, FLUX sandwich norm versus MM-JiT pre-norm, and a two-block text preamble on 6400 images at 256^2 pixels for 2,000 steps. Width, heads, and frozen Qwen3-0.6B are shared. Figure 7(a) compares the representative backbone variants using 8-step generation-token accuracy; Table 8 retains the endpoint and resource comparisons needed for the final design choice.

AdaLN stabilizes plain MM-JiT. In the backbone screen, MM-JiT with AdaLN and two preamble blocks reaches 0.997 final generation-token accuracy, whereas MM-JiT without AdaLN reaches

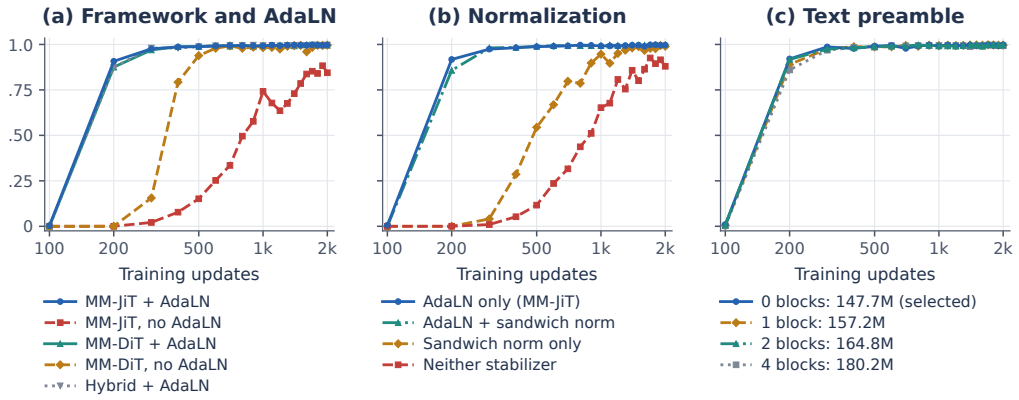


Figure 7 **Architecture and component ablations.** All curves report 8-step generation-token accuracy on the 6400-image, 256^2-pixel probe over 2,000 training updates. (a) Representative backbone and AdaLN configurations. (b) Normalization choices with two text-preamble blocks. (c) Text-preamble depth with MM-JiT and AdaLN.



Figure 8 **MM-JiT stabilizer examples.** The 12 columns show the same cases across three rows: (a) tokenizer reconstructions, (b) 8-step generation with AdaLN, and (c) 8-step generation without AdaLN. Both generation configurations use two text-preamble blocks.

0.844. The MM-DiT configuration without AdaLN remains stable (0.996). The component isolation in Table 8 confirms that removing both stabilizers degrades MM-JiT to 0.880.

Findings. MM-JiT with AdaLN and two preamble blocks first exceeds 0.99 generation-token accuracy at step 600 and finishes at 0.997 using 8.3 GB. It serves as the two-preamble baseline for component isolation.

C.4 Component isolation

The 6400-image screens isolate the effects of preamble depth and normalization across configurations of roughly 150M to 180M parameters. Figure 7(b,c) reports generation-token accuracy for these controls; Table 8 gives their final values and resource costs.

Preamble. Zero preamble first reaches 0.99 generation-token accuracy at step 500 and finishes at 0.998 with 147.7M parameters and 7.56 GB. Figure 7(c) shows that added blocks cost about 8M each without improving this probe.

Normalization. In Figure 7(b), AdaLN-only reaches 0.99 at step 600; sandwich-only first reaches this threshold at step 1,900, and removing both finishes at 0.88. Combining AdaLN with sandwich norm gives no gain. Figure 8 compares reconstructions and generated outputs on the same 12 examples.

Text mask token. Four-seed final accuracy is 0.9971 ± 0.0009 with the token and 0.9965 ± 0.0018 without it. This probe shows no clear gain; the formal model retains the token for varied padding.

The selected LightMGT uses plain MM-JiT, AdaLN, no text preamble, and the learned text mask token at 147.7M parameters.

Table 9 Text-to-image training data and public provenance.

Data source / category	Public source and paper	Samples
Web-scale image–text collections	Conceptual Captions 12M (Changpinyo et al., 2021); CommonCatalog (Gokaslan et al., 2024); DALL-E 3 synthetic (OpenAI, 2023); FLUX Journey; JourneyDB part 1 (Sun et al., 2023) and part 2 (Sun et al., 2023); LAION-Aesthetics 512 (Schuhmann et al., 2022) and LAION-Aesthetics 1024 (Schuhmann et al., 2022); MultiGen-20M (Zhang et al., 2025); Text-to-Image 2M (Zou et al., 2026)	24,979,177
High-resolution and synthetic image collections	MJHQ-30K (Li et al., 2024a); Pexels Aesthetic 1024; LAION-Aesthetics 1024 bucketed (Schuhmann et al., 2022); CC12M realistic 1024 (Changpinyo et al., 2021); 4KLSDB (Zhu et al., 2026); FLUX.2 synthetic 2M;	24,255,192
Curated high-quality images	CC12M–ImageNet21K recaptioned (Changpinyo et al., 2021); RenderedText Fine-T2I (Ma et al., 2026b); TextAtlas5M subsets (Wang et al., 2025a).	14,225,293
Dense captioning and spatial layout	PixelProse shards (Singla et al., 2024); TextGround4M (Mao et al., 2026); SPRIGHT (Chatterjee et al., 2024) and SPRIGHT–COCO (Chatterjee et al., 2024)	12,807,213
Dense captioning, reasoning, and style datasets	BLIP3o long captions (Xue et al., 2025) and short captions (Xue et al., 2025); Stable Diffusion 3 Recaption (Esser et al., 2024); FLUX-Reason-6M (Fang et al., 2026); AnyWord-3M (Tuo et al., 2024); MegaStyle-1.4M (Gao et al., 2026); Midjourney Niji; Takara image captions; LLaVA-ReCap-558K	41,267,042
Preference, product, and human-captioned images	FACAD fashion captions; Pick-a-Pic v2 (Kirstain et al., 2023); X-AIGD (Xiao et al., 2026); Shopify product catalogue; COCO captions (Lin et al., 2014); Flickr30K (Young et al., 2014); Shopify stock images; Flickr8K (Hodosh et al., 2013); DALL-E 3 contrastive captions (OpenAI, 2023); DALL-E 3 palette (OpenAI, 2023)	376,636
Text rendering and poster data	MARIO-10M (Chen et al., 2023); GenPoster-100K (Wang et al., 2025b)	10,515,395
Preference-selected synthetic images	HPDv3-PlusPlus (Liu et al., 2026b)	345,734
Supplemental capability datasets	300K image–text captions; Qwen DeepFashion; Synthetic Characters; Apple synthetic scenes; LAION 110K GPT-4V captions; Movie scenes; Synthetic object relations; Lemon; ArchDaily 30K; Pexels advertisement photography; Photo typography; Aesthetic image captions 10K	1,242,608
Additional aesthetic images	LAION 600K aesthetic (Schuhmann et al., 2022); Midjourney v6	546,904
Large synthetic scene collection	MeissonFlow Park	1,919,358
Total		**132,480,552**

D Details for step distillation

Adam uses $(0.9, 0.999)$, 100 warmup updates to 2×10^{-5}, then a constant rate; teacher CFG is 2.0, mask ratio 0.6, embedding noise 0.1, and EMA cap 0.9999. APM uses $N = 256$, auxiliary smoothing 0.1, $\beta = -0.2$.



Figure 9 One-step APM samples from ImageNet-256 classes.

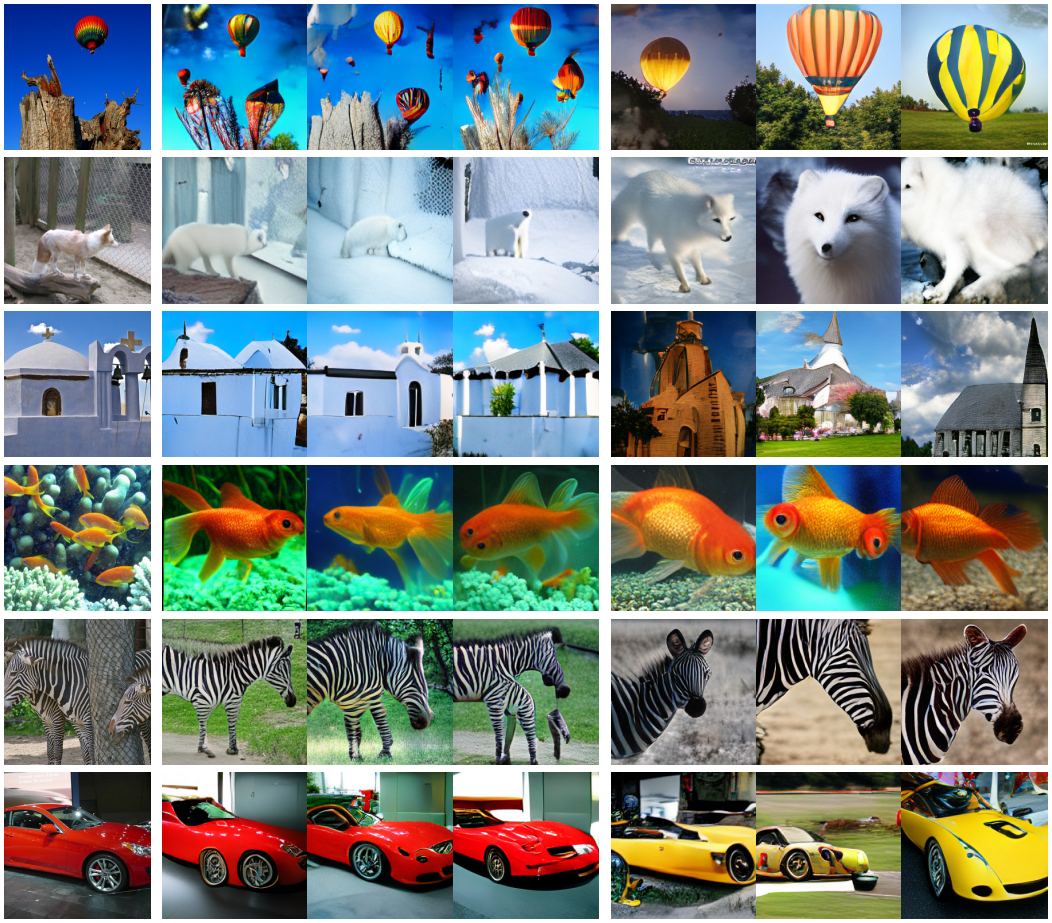


Figure 10 One-step reconstruction and random generation with APM on ImageNet-256. Each row shows a real image on the left, three reconstructions initialized from encoded image tokens in the middle, and three samples initialized from random tokens on the right.



Figure 11 Qualitative comparison across sampling budgets.

Table 10 Image editing training data and public provenance.

Public source and paper	Samples
GPT-Image-Edit-1.5M (Wang et al., 2025c)	678,389
ImgEdit (Ye et al., 2026b)	707,788
Nano-consistent-150k (Ye et al., 2025)	123,268
OmniEdit-Filtered-1.2M (Wei et al., 2025)	367,884
Pico-Banana-400K (Qian et al., 2026)	10,765
instruction-tuning-sd/cartoonization	5,000
CompBench (Jia et al., 2026)	3,403
CrispEdit-2M (Chow et al., 2026)	2,283,164
ImageEditingRequestV1	3,934
InstructPix2Pix CLIP-filtered (Brooks et al., 2023)	313,010
Inter-Edit-Train (Liu et al., 2026a)	1,096,032
low-level-image-proc	1,917
MagicBrush (Zhang et al., 2023)	9,335
multi-edit-clipd-improved	4,940
NHR-Edit (Kuprashevich et al., 2026)	640,088
ScaleEdit-12M (Chen et al., 2026)	606,378
SciGenEdit-10K (Li et al., 2026)	2,355
ShareGPT-4o-Image (Chen et al., 2025)	46,494
TextEdit (Tian et al., 2026)	2,148
UltraEdit 500K (Zhao et al., 2024)	499,978
UnicEdit-10M (Ye et al., 2026a)	484,601
X2Edit (Ma et al., 2026a)	5,981,687
xLingual-picobanana-12k (Qian et al., 2026)	9,763
Total	**13,991,529**