STATEWISE: FROM NOISE AWARENESS TO STATE-CALIBRATED MLLM CONDITIONING FOR UNIFIED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified multimodal models aim to improve image generation and editing through visual understanding. Connecting a pretrained multimodal large language model (MLLM) to a diffusion generator allows its understanding capabilities to guide image synthesis. However, reusing a fixed condition throughout denoising prevents the MLLM from adapting its guidance to the evolving latent. Adapting this condition requires identifying the current latent state, which may depart from the state expected at the input timestep during finite-step sampling. We introduce StateWise, a state-calibrated MLLM conditioning framework that reads noisy latents directly in the generator's native VAE space without an auxiliary visual encoder on this pathway. During training, we vary each latent's noise level while keeping its input timestep fixed, teaching the MLLM not only to perceive the latent's noise state but also to estimate how much noisier or cleaner it is than expected at that timestep. These estimates guide adjustments to the condition supplied to the diffusion generator at each denoising step, while an additional constraint discourages changes when the observed and expected noise levels agree. Using only publicly available training data, StateWise achieves 0.90 on GenEval and 77.4 on GenEval2, alongside competitive image editing. Code will be publicly released.

1 INTRODUCTION

Unified multimodal models (UMMs) aim to improve image generation and editing through visual understanding (Xie et al., 2025b; Zhou et al., 2025; Wu et al., 2025b; Ma et al., 2025; Xie et al., 2025c). Pairing a pretrained multimodal large language model (MLLM) with a diffusion transformer (DiT) offers a practical way to retain pretrained understanding while benefiting from a specialized image generator (Pan et al., 2025; Chen et al., 2025a; Wu et al., 2025d). The MLLM contributes knowledge, semantic interpretation, and instruction following, while the DiT synthesizes the image through iterative denoising. This division of labor makes the conditioning interface central to how understanding can assist generation.

However, a stronger encoder does not by itself make conditioning responsive to the generation process. Many diffusion models reuse the same prompt representation throughout denoising, whether it is produced by a CLIP or T5 text encoder (Esser et al., 2024) or an MLLM (Pan et al., 2025; Wu et al., 2025d). The generator processes this representation together with the evolving latent, but the encoder does not update its output based on the current generation state. In this setting, the MLLM specifies what to generate but does not observe how the sample develops. Because the effect of semantic guidance varies across denoising stages (Wu et al., 2023; Lv et al., 2025; Guo et al., 2026), guidance should account for the sample's current state as well as the prompt. This motivates allowing the MLLM to read the current latent and adapt its condition throughout generation.

Recent work has begun to address this limitation by conditioning the MLLM on the evolving noisy latent. RepFusion feeds evolving RAE features to the MLLM and trains a separate decoder for image reconstruction (Pan et al., 2026), while TUNA adapts a SigLIP2 encoder to process noisy VAE latents (Liu et al., 2026b). These representations make the MLLM noise-aware, but do not by themselves teach it to recognize when the current latent departs from the state expected at its input timestep (Fig. 1). During training, standard flow matching constructs noisy latents at noise levels

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.

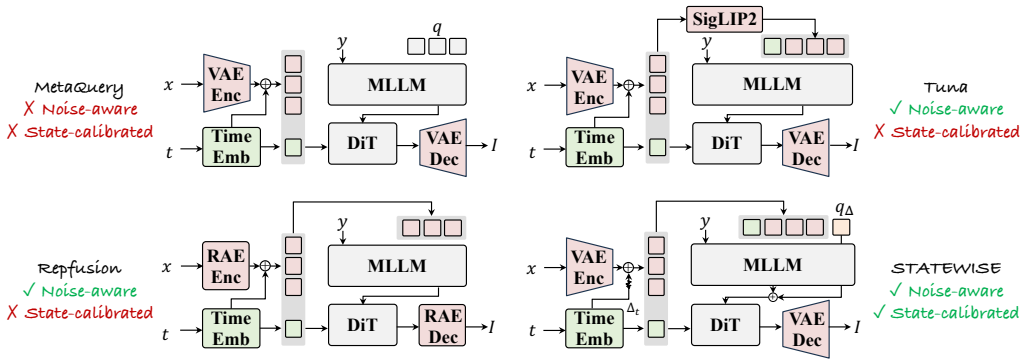


Figure 1: **From static to state-calibrated conditioning.** Static methods reuse a fixed condition, while noise-aware methods incorporate the evolving latent into MLLM conditioning. StateWise reads native VAE latents and uses a learned state estimate to calibrate the DiT condition.

matched to their input timesteps. During sampling, however, latents are recursively updated using model predictions, and accumulated prediction and numerical integration errors can cause them to depart from the states expected at those timesteps (Fig. 2) (Ning et al., 2023; Zhang et al., 2025; Li et al., 2024a; Lee et al., 2024; Zhong et al., 2026; Yu et al., 2026). A diagnostic with a frozen Sana model shows a late-stage difference in noise-prediction norms between sampled latents and forward-noised references at matched timesteps (App. D.1). Under schedule-matched training, the timestep already specifies the construction noise level. Learning from these matched pairs alone does not ensure that the MLLM can recognize when a latent is noisier or cleaner than expected. Consequently, the condition supplied to the DiT may not suit the latent's current state. We retain the generator's native VAE space and focus on teaching the MLLM to assess the latent relative to its timestep and use the resulting estimate to adjust the DiT condition.

We introduce StateWise, a framework for *state-calibrated MLLM conditioning*. A lightweight tokenizer maps the current noisy latent from the generator's native VAE space into MLLM input tokens. Both the VAE and MLLM remain frozen, and no additional vision encoder is introduced. During training, we sample an input timestep and construct latents at different noise levels. Supervising the difference from the level specified by the timestep teaches the MLLM not only to perceive the latent's noise state but also to estimate how much noisier or cleaner it is than expected. Crucially, the same timestep is paired with different noise states, making time alone insufficient for predicting this difference. The estimated difference modulates a learned update to the DiT condition. We train this update through the generation objective and discourage changes to the base condition on matched examples. At inference, StateWise recomputes the condition from the current latent and timestep at every denoising step, using its learned estimate to adapt the guidance throughout sampling.

Using only publicly available pretrained models and training data, we evaluate latent-state sensitivity and downstream generation. Fixed-timestep diagnostics show that state predictions follow the relative denoising progress of generated latents. Ablations show lower flow-matching training losses during pretraining and fine-tuning, alongside improved image generation. The final model achieves 0.90 on GenEval and 77.4 on GenEval2, alongside competitive image editing.

Our contributions are threefold:

- We introduce StateWise, a state-calibrated MLLM conditioning framework that reads noisy latents directly in the generator's native VAE space and adapts the DiT condition at each denoising step, without an auxiliary visual encoder on this pathway.
- We decouple latent noise levels from input timesteps to supervise the MLLM's prediction of noise-level deviations. These predictions regulate learned condition updates, while an additional constraint discourages unnecessary changes when the noise level matches the timestep.
- Experiments using only publicly available training data show improved image generation and competitive image editing.

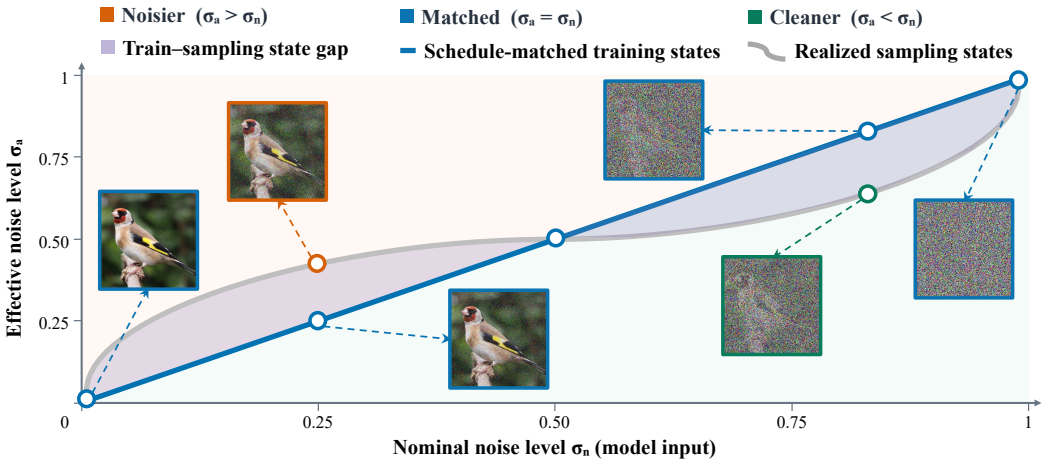


Figure 2: **Training–sampling mismatch: a schematic illustration.** Standard training pairs constructed latents with matching timesteps, whereas recursively sampled latents may deviate from this correspondence.

2 Related Work

2.1 Unified Multimodal Generation and Editing

Unified multimodal models combine understanding and synthesis either within a shared transformer, using autoregressive, discrete-diffusion, or continuous objectives (Xie et al., 2025b; Zhou et al., 2025; Wu et al., 2025b; Ma et al., 2025; Xie et al., 2025c; Wu et al., 2025e; Qu et al., 2025), or through modular systems that connect pretrained language models to diffusion decoders (Wu et al., 2024; Li et al., 2024b; Lin et al., 2025b). Query-based variants such as MetaQueries, OpenUni, and BLIP3-o provide a compact MLLM–generator interface (Pan et al., 2025; Wu et al., 2025d; Chen et al., 2025a).

This scope increasingly includes editing. Instruction-based editors use source images and natural-language commands (Brooks et al., 2023; Huang et al., 2024), while generalist systems unify generation and editing through shared sequences, multi-frame formulations, joint training, or reference conditioning (Xiao et al., 2025; Le et al., 2025; Xia et al., 2025; Chen et al., 2025d; Fu et al., 2025; Tang et al., 2026). WeMMU combines randomly sampled Gaussian query tokens with source-image VAE features (Yang et al., 2026); these queries are distinct from the evolving target latent. RepFusion, UniDDT, and TUNA instead encode the current noisy latent through representation-specific visual interfaces (Pan et al., 2026; Wang et al., 2026a; Liu et al., 2026b). These interfaces can already incorporate timestep information, but do not explicitly supervise the discrepancy between latent corruption and nominal time. StateWise learns this state–time relation directly in the generator's native VAE space and uses the predicted discrepancy to calibrate the condition, without an added vision foundation model on the noisy-latent path.

2.2 Time-Varying Conditioning and Trajectory Mismatch

Semantic requirements vary across denoising stages. Time-dependent attention methods adapt cross-modal interaction (Wu et al., 2023; Lv et al., 2025; Guo et al., 2026), while CADS and limited-interval guidance schedule condition noise or guidance strength (Sadat et al., 2024; Kynkäänniemi et al., 2024). These approaches motivate dynamic conditioning, but index it by a nominal timestep or predefined interval rather than infer the corruption realized by the evolving latent.

Complementary work studies the training–sampling trajectory gap. Input Perturbation simulates exposure bias, Time-Shift Sampler remaps generated states, Anti-Exposure Bias compensates drift through sampling-time prompts, and rectified-flow analyses examine finite-step integration (Ning et al., 2023; Li et al., 2024a; Zhang et al., 2025; Lee et al., 2024).Noise Awareness Guidance (NAG)

diagnoses noise-level shifts and guides sampling toward the prescribed noise schedule (Zhong et al., 2026).SNR–t Bias directly diagnoses disagreement between sample SNR and assigned time, but corrects the sampling state (Yu et al., 2026). These methods show that recursively generated states can depart from nominal coordinates, but address the mismatch through perturbed inputs, altered sampling coordinates or guidance, or latent correction. StateWise instead turns state–time deviations into supervision for the MLLM conditioner: it estimates whether a native VAE latent is cleaner or noisier than its nominal timestep and calibrates the semantic condition, while retaining the sampling schedule and velocity labels, without adding a separate latent-correction step during sampling.

3 METHOD

3.1 PRELIMINARIES: CONDITIONING PARADIGMS

Let $\mathbf{y}$ denote prompt embeddings, $\mathbf{z}_\sigma$ a latent at noise level σ, and $\mathbf{C}$ the condition supplied to the generator. Text-only conditioning uses a prompt encoder $\mathcal{E}_{\text{text}}$: CLIP-based and T5 encoders in Stable Diffusion 3 (Esser et al., 2024), or MLLMs with learnable queries in MetaQueries and OpenUni (Pan et al., 2025; Wu et al., 2025d). For text-to-image generation, its output stays fixed throughout denoising.

Noise-aware conditioning also reads the evolving latent and timestep: RepFusion uses projected RAE features (Pan et al., 2026), whereas TUNA uses SigLIP2-encoded VAE latents (Liu et al., 2026b). With $\mathcal{V}$ mapping latents to MLLM tokens and $\mathcal{F}$ producing the condition, we write

$$\underbrace{\mathbf{C}_{\text{text}} = \mathcal{E}_{\text{text}}(\mathbf{y})}_{\text{text-only conditioning}}, \qquad \underbrace{\mathbf{C}_\sigma^{\text{NA}} = \mathcal{F}\big(\mathbf{y}, \mathcal{V}(\mathbf{z}_\sigma; \sigma), \sigma\big)}_{\text{noise-aware conditioning}}. \tag{1}$$

StateWise instead maps native VAE latents to MLLM tokens through a lightweight tokenizer $\mathcal{P}$. It distinguishes the construction noise level σ_{a} from the supplied noise level σ_{n}:

$$\underbrace{\mathbf{C}^{\text{SC}}_{\sigma_{\text{a}},\sigma_{\text{n}}} = \mathcal{F}\big(\mathbf{y}, \mathcal{P}(\mathbf{z}_{\sigma_{\text{a}}}, \sigma_{\text{n}}), \sigma_{\text{n}}\big)}_{\text{state-calibrated conditioning}}. \tag{2}$$

We explicitly supervise the discrepancy between these levels and use the MLLM's prediction to calibrate the DiT condition, as illustrated in Fig. 3.

3.2 NATIVE-VAE CONDITIONING

As illustrated in Fig. 3, we first construct a dynamic condition by allowing the MLLM to read the current noisy latent directly in the generator's native VAE space. We keep the VAE and MLLM frozen and map noisy latents into MLLM tokens through a lightweight tokenizer, without an auxiliary visual encoder on this pathway.

The tokenizer $\mathcal{P}$ applies timestep-conditioned adaptive LayerNorm and projects latent features into the MLLM token space. We also project the generator's timestep embedding into an explicit time token $\mathbf{q}_{\sigma_{\text{n}}}$, providing a reference for the intended denoising stage. We arrange the prompt, time, and latent tokens in this order:

$$\mathbf{Q}_z = \mathcal{P}(\mathbf{z}_{\sigma_{\text{a}}}; \sigma_{\text{n}}), \qquad \mathbf{H} = \mathcal{M}([\mathbf{y}; \mathbf{q}_{\sigma_{\text{n}}}; \mathbf{Q}_z]), \qquad \mathbf{C}^0 = \mathcal{G}(\mathbf{H}_z). \tag{3}$$

Here, $\mathcal{M}$ is the frozen MLLM, and $\mathbf{H}_z$ denotes its latent-token hidden states, which attend to the preceding prompt and time tokens under causal attention. The connector $\mathcal{G}$ maps these states to the DiT condition space. We recompute the base condition $\mathbf{C}^0$ at every sampling step, leaving the DiT's original latent and timestep inputs unchanged. Tokenization details are provided in App. C. The next section introduces explicit discrepancy supervision to calibrate this dynamic condition.

3.3 STATE CALIBRATION

Building on the dynamic condition $\mathbf{C}^0$ in Sec. 3.2, we train the MLLM to estimate how much noisier or cleaner a latent is than expected at the supplied timestep (Fig. 2). This estimate regulates a learned update to the condition supplied to the DiT.

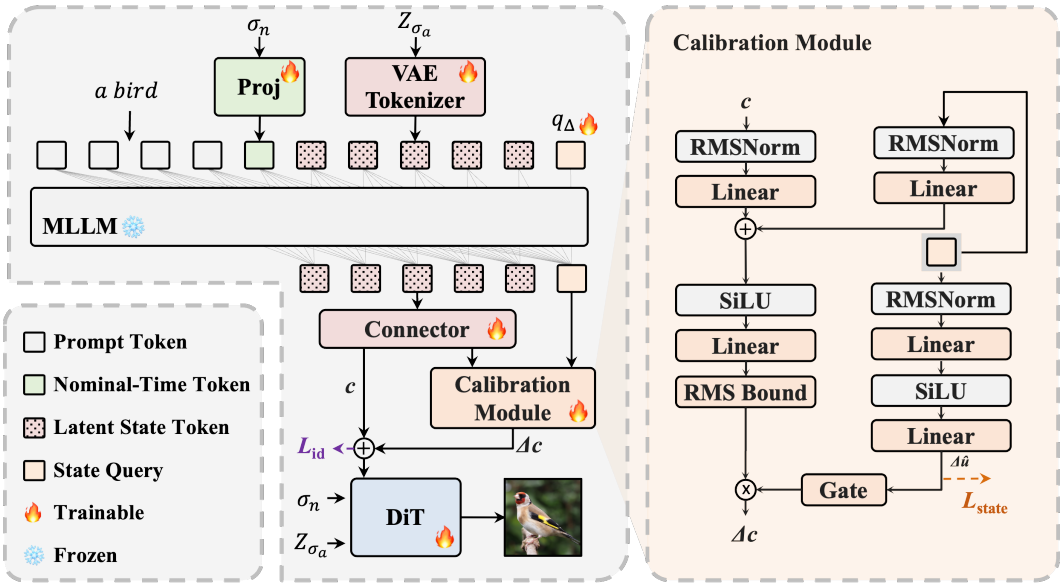


Figure 3: **Overview of StateWise.** The frozen MLLM reads text, native-VAE latent tokens, and the input timestep to produce a dynamic condition. A state query predicts the noise-level discrepancy, which modulates a learned update to the DiT condition.

Noise-level augmentation. During training, we sample a nominal noise level σ_{n} and vary the latent construction level using a bounded signed offset δ, while retaining σ_{n} as the input to both the MLLM and DiT. Given a clean VAE latent $\mathbf{x}_0$ and Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we construct

$$\sigma_{\mathrm{a}} = \mathrm{sigmoid}(\mathrm{logit}(\sigma_{\mathrm{n}}) + \delta)\,, \qquad \mathbf{z}_{\sigma_{\mathrm{a}}} = (1-\sigma_{\mathrm{a}})\mathbf{x}_0 + \sigma_{\mathrm{a}}\boldsymbol{\epsilon}, \qquad \Delta u^\star = \mathrm{logit}(\sigma_{\mathrm{a}}) - \mathrm{logit}(\sigma_{\mathrm{n}}). \tag{4}$$

Here, $\Delta u^\star$ measures the noise-level difference in logit space. Positive and negative offsets produce noisier and cleaner states, respectively; we retain zero-offset examples as matched pairs.

State calibration. We append a learned state query to the sequence in Sec. 3.2, yielding $[\mathbf{y}; \mathbf{q}_{\sigma_{\mathrm{n}}}; \mathbf{Q}_z; \mathbf{q}_\Delta]$. Under causal attention, this query reads the preceding text, time, and latent tokens without altering the hidden states that produce $\mathbf{C}^0$. A lightweight head predicts the discrepancy from its output $\mathbf{H}_\Delta$:

$$\widehat{\Delta u} = h_\Delta(\mathbf{H}_\Delta), \qquad \mathcal{L}_{\mathrm{state}} = \ell_{\mathrm{SL1}}\big(\widehat{\Delta u}, \Delta u^\star\big). \tag{5}$$

Here, ℓ_{SL1} denotes the Smooth L1 loss.

To calibrate the condition, an adapter uses $\mathbf{C}^0$ and $\mathbf{H}_\Delta$ to produce a bounded residual $\overline{\mathbf{R}}$, while a gate $g(\widehat{\Delta u})$ controls its sign and magnitude. For matched examples, we also penalize changes to the base condition:

$$\mathbf{C} = \mathbf{C}^0 + g(\widehat{\Delta u})\overline{\mathbf{R}}, \qquad \mathcal{L}_{\mathrm{id}} = \mathrm{MSE}(\mathbf{C}, \mathbf{C}^0), \quad \text{for } \Delta u^\star = 0. \tag{6}$$

The state loss supervises discrepancy estimation, while the identity loss anchors the calibrated condition to $\mathbf{C}^0$ on matched examples. The residual update is learned through the generation objective.

For the augmented latent–timestep pairs, we retain the linear-path velocity target $\mathbf{v}^\star = \boldsymbol{\epsilon} - \mathbf{x}_0$, which is independent of the interpolation level. The joint training objective is

$$\mathcal{L}_{\mathrm{FM}} = \mathrm{MSE}(\mathbf{v}_\theta(\mathbf{z}_{\sigma_{\mathrm{a}}}, \sigma_{\mathrm{n}}, \mathbf{C}), \mathbf{v}^\star)\,, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{state}}\mathcal{L}_{\mathrm{state}} + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}. \tag{7}$$

At inference, the current latent and supplied timestep yield a learned discrepancy estimate that modulates $\mathbf{C}^0$ at each denoising step; neither the synthetic offset δ nor its target $\Delta u^\star$ is needed. Implementation and sampling settings are provided in App. C.

3.4 Training Strategy

We train StateWise in three stages, using the state-calibrated conditioning and shared objective in Sec. 3.3 throughout. The VAE and MLLM remain frozen; we optimize the DiT and conditioning interface, together with the source adapters introduced in Stage 3.

Stage 1: Large-Scale Generation Pretraining. We use BLIP3-o-Pretrain image–text pairs to learn the native-VAE conditioning interface and establish text-to-image generation capability. **Stage 2: Generation Refinement.** High-quality text-to-image pairs are used to improve prompt following and visual quality. **Stage 3: Joint Generation and Editing.** We use a 65% Edit / 35% T2I mixture to learn source-conditioned editing and retain generation capability. Tab. 1 summarizes the data composition and optimization settings.

Table 1: **Three-stage training setup.** Batch denotes the effective global batch size; Seen counts sample presentations, including repeats.

Stage	Res.	Batch	Opt. steps	Seen	LR	Warmup	Data (unique)	Task ratio
1	1024^2	1920	200K	384M	3×10^{-4}	1K	BLIP3-o-Pretrain (38.3M)	T2I 100%
2	1024^2	576	45K	25.9M	1×10^{-4}	450	T2I SFT mix (172.5K)	T2I 100%
3	1024^2	576	100K	57.6M	1×10^{-4}	1K	Joint mix (2.50M)	T2I 35% / Edit 65%

For editing, we additionally condition the target-generation pathway on the source image (Wu et al., 2025a;c). The MLLM's existing visual encoder provides source semantics, while clean source VAE latents supply spatial and appearance information. These latents are incorporated into the DiT condition and initial target patch embeddings; T2I samples bypass the source-specific paths. Both tasks retain the state-calibrated conditioning and target flow objective of Sec. 3.3. Source-conditioning details and additional editing supervision are provided in App. B.

4 Experiments

4.1 Implementation Details

Architectures. We use frozen Qwen3-VL-4B-Instruct and DC-AE backbones, and train the SANA-1.5 1.6B diffusion transformer together with the conditioning interface at 1024×1024 resolution (Bai et al., 2025; Chen et al., 2025c; Xie et al., 2025a). DC-AE produces a $32\times32\times32$ latent, which a fixed 2×2 space-to-channel rearrangement converts into 256 latent tokens. Editing uses the source-conditioning pathways described in Sec. 3.4.

Training data. We use only publicly available datasets for the three training stages (Tab. 1). Stage 1 uses filtered BLIP3-o-Pretrain (Chen et al., 2025a). Stage 2 combines BLIP3-o-60K, Echo-4o, and ShareGPT-4o-Image (Chen et al., 2025a; Ye et al., 2025a; Chen et al., 2025b). Stage 3 augments this T2I pool with OSP1024, the generation split of OpenGPT-4o-Image, and Fine-T2I (Lin et al., 2025a; Chen et al., 2025f; Ma et al., 2026), and introduces editing examples from GPT-Image-Edit-1.5M, Uni-Edit-148K, and the editing splits of ShareGPT-4o-Image and OpenGPT-4o-Image (Wang et al., 2025; Zheng et al., 2026; Chen et al., 2025b;f). Editing examples are normalized and filtered for validity and geometric compatibility (App. A).

Inference. For both generation and editing, we use second-order multistep DPM-Solver++ (Lu et al., 2025) with SANA's flow-matching schedule. We use 20 sampling steps and a classifier-free guidance scale of 4.5.

Evaluation. GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025) assess compositional text-to-image generation, while DPGBench (Hu et al., 2024) evaluates alignment with detailed prompts. ImageEdit (Ye et al., 2025b) and GEdit-Bench-EN (Liu et al., 2025) assess instruction-following editing and visual quality.

4.2 Main Results

Image generation. StateWise achieves the highest overall scores on GenEval (0.90) and GenEval2 (77.4) among the methods in Tab. 2, alongside a competitive DPGBench score of 86.22. Relative

Table 2: **Text-to-image results on GenEval, DPGBench, and GenEval2.** † indicates prompt rewriting; ‡ indicates self-CoT.

METHOD	GenEval							DPGBench (↑)	GenEval2 (↑)
	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall (↑)		
Generation. Only									
SDv1.5 *CVPR 22*	0.97	0.38	0.35	0.76	0.04	0.06	0.43	63.18	-
SDv2.1 *CVPR 22*	0.98	0.51	0.44	0.85	0.07	0.17	0.50	-	-
SD3-Medium *ICML 24*	0.99	0.94	0.72	0.89	0.33	0.60	0.74	84.08	-
SDXL *ICLR 24*	0.98	0.74	0.39	0.85	0.15	0.23	0.55	74.65	45.3
PixArt-α *ICLR 24*	0.98	0.50	0.44	0.80	0.08	0.07	0.48	71.11	-
DALL-E 2 *arXiv 22*	0.94	0.66	0.49	0.77	0.10	0.19	0.52	-	-
DALL-E 3 *Tech. Rep. 23*	0.96	0.87	0.47	0.83	0.43	0.45	0.67	83.50	-
Unified Models									
Chameleon *arXiv 24*	-	-	-	-	-	-	0.39	-	-
Show-o *ICLR 25*	0.98	0.80	0.66	0.84	0.31	0.50	0.68	-	-
Show-o2-7B *NeurIPS 25*	1.00	0.87	0.58	0.92	0.52	0.62	0.76†	86.14	-
Janus *CVPR 25*	0.97	0.68	0.30	0.84	0.46	0.42	0.61	79.68	53.7
JanusFlow *CVPR 25*	0.97	0.59	0.45	0.83	0.53	0.42	0.63	80.09	42.7
Janus-Pro-1B *arXiv 25*	0.98	0.82	0.51	0.89	0.65	0.56	0.73	82.63	-
Janus-Pro-7B *arXiv 25*	0.99	0.89	0.59	0.90	0.79	0.66	0.80	84.19	54.5
MetaQuery-B† *arXiv 25*	-	-	-	-	-	-	0.74	80.04	-
MetaQuery-L† *arXiv 25*	-	-	-	-	-	-	0.78	81.10	-
MetaQuery-XL† *arXiv 25*	-	-	-	-	-	-	0.80	82.05	-
Openuni *arXiv 25*	0.99	0.92	0.76	0.91	0.82	0.77	0.86	83.08	-
UniWorld *arXiv 25*	-	-	-	-	-	-	0.84†	-	70.0
BLIP3-o-4B *arXiv 25*	-	-	-	-	-	-	0.81	79.36	-
BLIP3-o-8B *arXiv 25*	-	-	-	-	-	-	0.84	81.60	59.8
BAGEL-7B *arXiv 25*	0.99	0.94	0.81	0.88	0.64	0.63	0.82	84.03	-
BAGEL-7B† *arXiv 25*	0.98	0.95	0.84	**0.95**	0.78	0.77	0.88	-	71.7‡
Scale-RAE *arXiv 26*	-	-	-	-	-	-	0.83	79.70	-
WeMMU *CVPR 26*	-	-	-	-	**0.86**	0.77	0.88	83.69	-
HBridge† *CVPR 26*	1.00	**0.96**	0.80	0.94	0.77	0.78	0.87	85.23	-
Query-Kontext† *ECCV 26*	0.98	0.94	0.81	0.91	0.85	0.79	0.88	-	-
UniDDT *ECCV 26*	0.99	0.93	0.71	0.92	0.85	0.80	0.87	**86.90**	-
UNO *arXiv 26*	-	-	-	-	-	-	-	86.12	75.1‡
Repfusion *arXiv 26*	-	-	-	-	-	-	0.87	85.11	-
StateWise	**1.00**	0.95	**0.85**	0.93	0.85	**0.82**	**0.90**	86.22	**77.4**

Table 3: **Image-editing results on ImageEdit and GEdit-Bench-EN.**

Type	Method	Size	ImageEdit			GEdit-Bench-EN		
			Hybrid↑	Action↑	Overall↑	G_SC↑	G_PQ↑	G_O↑
Gen. Only	Gemini 2.5 Flash Image *Model Card 25*	–	3.66	4.59	4.28	7.41	**7.96**	7.10
Unified	EMU3.5 *arXiv 25*	34B	3.69	4.57	**4.41**	8.11	7.70	7.59
	GPT-4o *Online 25*	–	**3.96**	**4.89**	4.2	7.85	7.62	7.53
	QWen-Image *arXiv 25*	27B	3.82	4.69	4.27	8.00	7.86	7.56
	Bagel *arXiv 25*	7B	2.38	4.17	3.2	7.36	6.83	6.52
	OmniGen2 *arXiv 25*	7B	2.52	4.68	3.44	7.16	6.77	6.41
	UniWorld-V1 *arXiv 25*	8B	2.96	2.74	3.26	4.93	7.43	4.85
	WeMMU *CVPR 26*	8B	2.82	3.15	3.31	5.86	6.80	5.75
	HBridge *CVPR 26*	-	3.44	4.32	3.54	-	-	-
	Query-Kontext *ECCV 26*	17B	–	–	–	**8.36**	7.37	**7.66**
	StateWise	5.6B	3.93	4.14	4.28	7.84	7.48	7.38

to OpenUni, the clearest gains occur in counting ($0.76 \rightarrow 0.85$) and color attribution ($0.77 \rightarrow 0.82$), highlighting the full model's ability to satisfy object-count and attribute-binding constraints.

Image editing. In Tab. 3, StateWise achieves overall scores of 4.28 on ImageEdit and 7.38 on GEdit-Bench-EN, improving all six metrics over WeMMU. On GEdit-Bench-EN, semantic consistency improves from 5.86 to 7.84, reflecting stronger adherence to the requested edits.

Both evaluations use the final Stage 3 model, demonstrating strong generation and competitive editing within one system. Fig. 4 provides qualitative examples of both capabilities.

4.3 Method Ablation

Ablation setting. We conduct T2I-only component ablations with InternVL3-2B and SANA-1.5 (1.6B), using Stage 1 and Stage 2 data (Sec. 4.1). The noise-aware variants use 40K pre-training and 45K fine-tuning updates. Starting from an OpenUni-style reference (Wu et al., 2025d), we first introduce native-VAE noise-aware conditioning. We then add noise-level augmentation in both

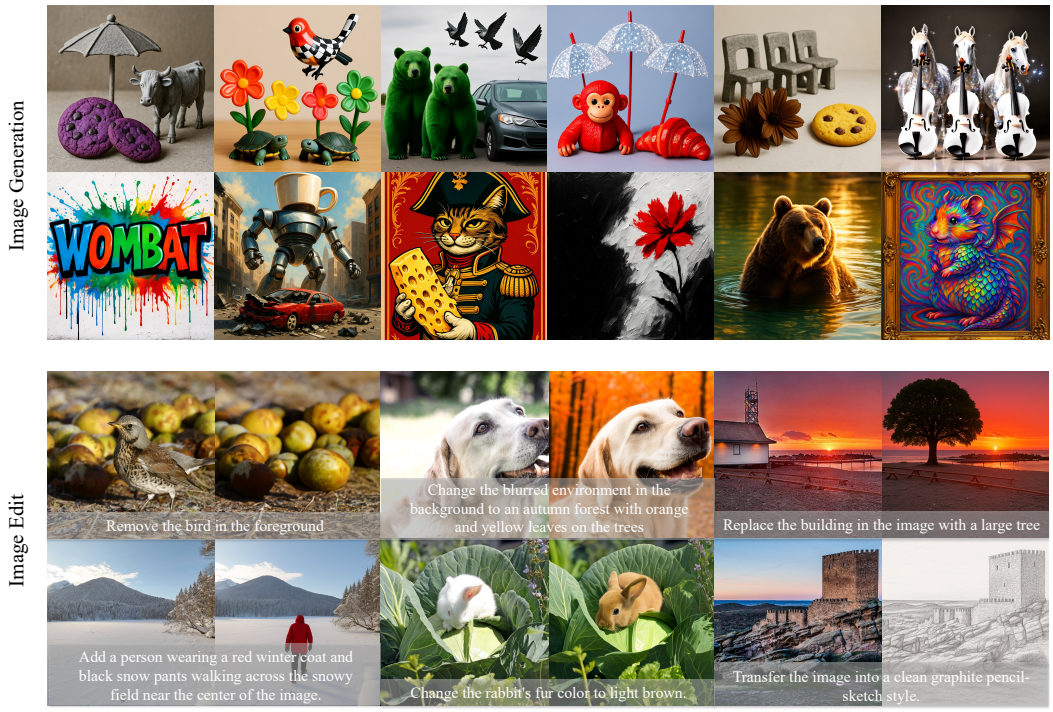


Figure 4: **Generation and editing examples of StateWise.** Top: text-to-image results spanning object compositions and rendering styles. Bottom: source images (left) and edited outputs (right), with instructions.

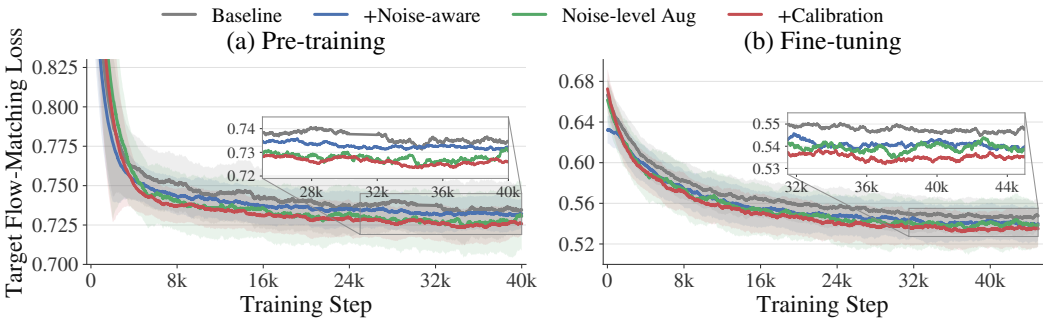


Figure 5: **Training loss across ablation configurations.** EMA-smoothed (0.99) flow-matching losses during (a) pre-training and (b) instruction fine-tuning. Bands show ± 1 rolling standard deviation of the loss residuals around the EMA; insets enlarge late training.

stages, retaining only the flow-matching objective. Finally, we add state prediction and condition calibration (Sec. 3.3) to obtain full StateWise.

Training dynamics. Fig. 5 reports flow-matching training losses during pre-training and fine-tuning (Appendix C). Under the same noise-level augmentation scheme, StateWise exhibits lower late-training losses than the augmentation-only variant in both stages. Together with the GenEval results in Tab. 4, these observations support the effectiveness of the complete calibration module.

Compositional generation. Tab. 4 evaluates the same progression on GenEval. Noise-aware conditioning raises the overall score from 0.86 to 0.87, and noise-level augmentation increases it to 0.88. Adding state calibration further improves counting ($82.81 \rightarrow 84.38$) and position ($82.25 \rightarrow 84.00$), alongside gains in two-object and color scores.

Table 4: **Component ablation on GenEval.** Subscores are percentages; Overall uses a 0–1 scale.

Configuration	Noise-aware cond.	Noise-level aug.	State calib.	GenEval (↑)						
				Single object	Two object	Counting	Colors	Position	Color attr.	Overall
Baseline	✗	✗	✗	**99.69**	91.67	67.50	**94.41**	81.25	**81.75**	0.86
+ Noise-aware	✓	✗	✗	98.44	93.18	79.69	90.96	83.25	77.75	0.87
+ Noise-level Aug	✓	✓	✗	99.38	93.18	82.81	92.29	82.25	80.50	0.88
StateWise	✓	✓	✓	99.38	**94.19**	**84.38**	93.35	**84.00**	81.50	**0.89**

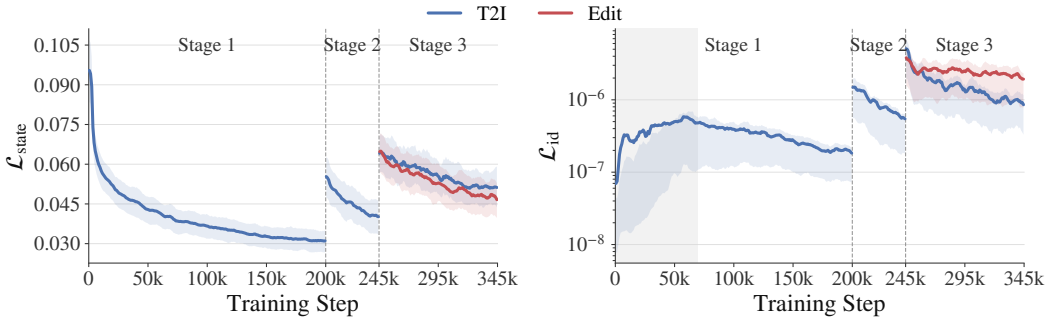


Figure 6: **Dynamics of selective state calibration.** Unweighted losses are summarized by 1K-step means and interquartile ranges; Stage 3 separates T2I and Edit. Gray shading denotes the initial activation period of the zero-initialized adapter.

4.4 State Calibration Analysis

We assess latent-state sensitivity and the restraint of condition updates on matched examples.

State sensitivity. We fix text and timestep and compare earlier, current, and later latents from 48 gate-disabled trajectories at three timestep settings. All 144 triplets yield higher predictions for earlier latents and lower predictions for later latents (App. D.2). With time unchanged, this ordering indicates sensitivity to relative latent progress beyond the timestep alone.

Training dynamics. Fig. 6 shows an overall decrease in $\mathcal{L}_{\text{state}}$ within each stage, including both Stage 3 tasks, with a transient increase after the Stage 1 restart. This tracks fitting of constructed noise-level offsets. Zero initialization sets $\mathbf{C} = \mathbf{C}^0$, so $\mathcal{L}_{\text{id}}$ starts near zero. Its initial rise accompanies activation of the correction branch; its subsequent decline indicates smaller condition updates on matched examples.

5 Conclusion

We presented StateWise, a state-calibrated MLLM conditioning framework for modular unified visual generation. The key observation is that exposing an MLLM to noisy latents does not ensure state identification: schedule-matched training leaves a timestep shortcut precisely where finite-step sampling may depart from the nominal schedule. StateWise turns this discrepancy into supervision in the generator's native VAE space and uses the inferred state to selectively calibrate the DiT condition. Experiments show lower flow-matching training losses and improved compositional generation, together with competitive source-conditioned editing and consistent calibration dynamics across the three training stages. More broadly, the results identify state reasoning as a complementary axis alongside stronger visual representations and improved samplers: a unified conditioner should reason about both what to generate and where the evolving latent lies relative to its denoising schedule.

AI USE STATEMENT

No generative AI tools were used in the preparation of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

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, 2023.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

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

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*, 2025b.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 96539–96560, 2025c.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12501–12511, 2025d.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025e.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025f.

Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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.

Tsu-Jui Fu, Yusu Qian, Chen Chen, Wenze Hu, Zhe Gan, and Yinfei Yang. Univg: A generalist diffusion model for unified image generation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17160–17170, 2025.

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

Xiefan Guo, Xinzhu Ma, Haiyu Zhang, and Di Huang. Ctcal: Rethinking text-to-image diffusion models via cross-timestep self-calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43558–43567, June 2026.

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*, 2024.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2671–2682, 2025.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, 2024.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *International Conference on Learning Representations*, 2024a.

Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6173–6188, 2024b.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, 2023.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025a.

Han Lin, Jaemin Cho, Amir Zadeh, Chuan Li, and Mohit Bansal. Bifrost-1: Bridging multimodal llms and diffusion models with patch-level clip latents. *Advances in Neural Information Processing Systems*, 38:69700–69725, 2025b.

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.

Zeyu Liu, Zanlin Ni, Yang Yue, Cheng Da, Huan Yang, Di Zhang, Kun Gai, and Gao Huang. Steering visual generation in unified multimodal models with understanding supervision. *arXiv preprint arXiv:2605.05781*, 2026a.

Zhiheng Liu, Weiming Ren, Haozhe Liu, Zijian Zhou, Shoufa Chen, Haonan Qiu, Xiaoke Huang, Zhaochong An, Fanny Yang, Aditya Patel, Viktar Atliha, Tony Ng, Xiao Han, Chuyan Zhu, Chenyang Zhang, Ding Liu, Juan-Manuel Perez-Rua, Sen He, Jürgen Schmidhuber, Wenhu Chen, Ping Luo, Wei Liu, Tao Xiang, Jonas Schult, and Yuren Cong. Tuna: Taming unified visual representations for native unified multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15740–15751, June 2026b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Zhengyao Lv, Tianlin Pan, Chenyang Si, Zhaoxi Chen, Wangmeng Zuo, Ziwei Liu, and Kwan-Yee K. Wong. Rethinking cross-modal interaction in multimodal diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5934–5943, October 2025.

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*, 2026.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7739–7751, 2025.

Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21492–21501, 2025.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

OpenAI. Gpt-4o, 2025. URL https://openai.com/index/introducing-4o-image-generation.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Xichen Pan, Aashu Singh, Satya Narayan Shukla, Xiangjun Fan, Shlok Kumar Mishra, and Saining Xie. Repfusion: Leveraging multimodal priors for denoising in representation space. *arXiv preprint arXiv:2606.14700*, 2026.

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*, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2545–2555. IEEE, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, volume 2024, pp. 23723–23755, 2024.

Yuxin Song, Wenkai Dong, Shizun Wang, Qi Zhang, Song Xue, Tao Yuan, Hu Yang, Haocheng Feng, Hang Zhou, Xinyan Xiao, et al. Query-kontext: An unified multimodal model for image generation and editing. *arXiv preprint arXiv:2509.26641*, 2025.

Hao Tang, Chen-Wei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing. In *International Conference on Learning Representations*, volume 2026, pp. 60440–60465, 2026.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemini Team. Gemini 2.5 flash & gemini 2.5 flash image model card, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

Shuai Wang, Liang Li, Yang Chen, Ruopeng Gao, Yao Teng, and Limin Wang. Uniddt: Unifying multimodal understanding and generation with decoupled diffusion transformer. In *European Conference on Computer Vision*, 2026a.

Xiang Wang, Zhifei Zhang, He Zhang, Zhe Lin, Yuqian Zhou, Qing Liu, Shiwei Zhang, Yijun Li, Shaoteng Liu, Haitian Zheng, et al. Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14767–14778, 2026b.

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*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12966–12977, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025c.

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7766–7776, October 2023.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, pp. 53366–53397, 2024.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.23661*, 2025d.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 93620–93638, 2025e.

Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28533–28543, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13294–13304, 2025.

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.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025a.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems*, pp. 47490–47518, 2025c.

Jian Yang, Dacheng Yin, Xiaoxuan He, Yong Li, Fengyun Rao, Jing Lyu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Wemmu: Enhanced bridging of vision-language models and diffusion models via noisy query tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17609–17618, June 2026.

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*, 2025a.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Meng Yu, Lei Sun, Jianhao Zeng, Xiangxiang Chu, and Kun Zhan. Elucidating the snr-t bias of diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43461–43470, June 2026.

Junyu Zhang, Daochang Liu, Eunbyung Park, Shichao Zhang, and Chang Xu. Anti-exposure bias in diffusion models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023.

Dian Zheng, Manyuan Zhang, Hongyu Li, Hongbo Liu, Kai Zou, Kaituo Feng, and Hongsheng Li. Uni-edit: Intelligent editing is a general task for unified model tuning. *arXiv preprint arXiv:2605.21487*, 2026.

Jincheng Zhong, Boyuan Jiang, Xin Tao, Pengfei Wan, Kun Gai, and Mingsheng Long. Mitigating noise shift in denoising generative models with noise awareness guidance. In *International Conference on Learning Representations*, volume 2026, pp. 34193–34211, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

A Edit-Data Preprocessing

We standardize instructions and remove invalid image–instruction pairs. The source and target are fitted to a shared content box, and a fixed geometric filter rejects poorly compatible pairs based on aspect-ratio agreement and target crop retention. This leaves 1.67M real Edit pairs. This preprocessing is shared by all Edit samples in Stage 3.

B Image-Editing Details

Source conditioning. Editing adds the clean source VAE latent as a diffusion input (Brooks et al., 2023). A token-gated adapter augments the DiT condition, and a parallel projection augments the target patch embeddings. Both source output projections are zero initialized, following the conditional-adaptation design of ControlNet and GLIGEN (Zhang et al., 2023; Li et al., 2023). The original target projection is unchanged; T2I samples bypass both source paths and the MLLM's source-image input.

Auxiliary supervision. Smoothed source–target differences provide soft preservation weights p, with larger values in similar regions and zero weight on padding. Approximately 15% of Edit examples are identity pairs that ask the model to keep the source unchanged. Editing uses

$$\mathcal{L}_e = \mathcal{L} + 0.10\ell_{\text{keep}} + 0.10\ell_{\text{copy}} + 0.05\ell_{\text{src}}, \tag{8}$$

where $\mathcal{L}$ is the shared objective in Eq. (7). Using Smooth L1, ℓ_{keep} compares the predicted clean latent with the source, weighted by p; ℓ_{copy} applies the same constraint to identity pairs, which also contribute to the first term. ℓ_{src} supervises the spatial source gate toward $0.1 + 0.8p$. This gate controls source injection, separately from the state-calibration gate g in Sec. 3.3. Errors are weighted by valid content and geometric compatibility. These Edit-only constraints preserve source content, whereas $\mathcal{L}_{\text{id}}$ regularizes the DiT condition on matched states. Their motivation follows mask-guided preservation (Nguyen et al., 2025).

C State-Calibration Details

Latent tokenizer. The tokenizer in Sec. 3.2 groups each 2×2 VAE latent neighborhood along the channel dimension, applies timestep-conditioned adaptive LayerNorm, and projects it into the MLLM hidden space with learned spatial embeddings. A timestep embedder and a zero-initialized modulation head produce the scale and shift. The explicit time token separately projects the generator's timestep embedding.

Condition calibration. The adapter predicts $\mathbf{R} = \mathcal{A}(\mathbf{C}^0, \mathbf{H}_\Delta)$ from the base condition and state-query representation. Its output projection is zero initialized, so calibration initially leaves $\mathbf{C}^0$ unchanged and $\mathcal{L}_{\text{id}}$ starts at zero. For condition token j, we use

$$\overline{\mathbf{R}}_j = \frac{\mathbf{R}_j}{1 + \text{RMS}(\mathbf{R}_j)}, \qquad g(\widehat{\Delta u}) = \eta \tanh\left(\frac{\widehat{\Delta u}}{\tau}\right). \tag{9}$$

RMS is computed over channels, adding 10^{-6} to the mean square for numerical stability. We use $\eta = 0.05$ and $\tau = 0.5$, bounding the per-token correction RMS below 0.05.

Training settings. Across all three stages, we sample σ_{n} uniformly from the 1,000-point grid between 0.001 and 1. We propose matched, noisier, and cleaner states with probabilities 0.5, 0.4, and 0.1, respectively. Matched proposals have zero offset; nonzero logit offsets have magnitudes sampled uniformly between 0.1 and 0.8. Targets use the realized logit difference after numerical clipping. We set $\lambda_{\text{state}} = 0.1$ and $\lambda_{\text{id}} = 1$ in Eq. (7). The state loss averages Smooth L1 errors ($\beta = 0.1$) across the batch; the identity loss averages squared corrections over tokens and channels within matched examples, then over those examples. The flow loss averages over latent elements and the batch, excluding padding for Edit samples.

Inference. We average the conditional and unconditional discrepancy predictions before applying g for classifier-free guidance, sharing the resulting gate while retaining each branch's own residual. For editing, both branches retain the source image; only the unconditional instruction is removed. We follow the sampling protocol in Sec. 4.1, with a flow shift of 3.0.

Ablation loss measurement. Fig. 5 reports the flow-matching component of the training objective under each variant's training-time noise sampling, with EMA smoothing of 0.99. The noise-aware model uses matched latent–timestep pairs; the augmentation-only and full variants share the same noise-level augmentation scheme. The baseline uses a different pre-training schedule. We compare late-training losses between the two augmented variants.

Calibration loss visualization. Fig. 6 reports unweighted losses against cumulative optimizer updates. We merge resumed logs by iteration, keeping the later record where sessions overlap, and retain the original loss scales. Curves and bands show within-run means and interquartile ranges, computed in 1K-update bins and smoothed with a five-bin moving average within each stage and task.

D Sampling-State Diagnostics

We examine two complementary questions: whether sampled latents differ from forward-noised references at the same timestep, and whether StateWise reads the relative denoising progress of generated latents.

D.1 Forward-Noised and Sampled Latents

Setup. Following the prediction-norm diagnostic of Yu et al. (2026), we use a frozen Sana-600M model (Xie et al., 2024) with its native DC-AE and 20-step second-order Flow-DPM sampler at 512×512 resolution and CFG scale 1. This is a standalone diagnostic without StateWise calibration. We define the public reference distribution using 2,000 JourneyDB image–caption pairs from BLIP3-o-Pretrain, fixed before evaluation and shared across three noise seeds (16, 42, and 99). Forward references are clean DC-AE encodings mixed with independently sampled Gaussian noise; sampled latents follow the model's recursive updates from an independent Gaussian initialization. At each step, both groups receive identical text conditions and scheduler timesteps, with forward noise levels matched to the scheduler. All measurements precede the solver update.

Measurement and results. For Sana's velocity prediction $\mathbf{v}_\theta$, the corresponding noise prediction is

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{z} + (1 - \sigma)\mathbf{v}_\theta(\mathbf{z}, t, \mathbf{y}), \tag{10}$$

where σ and t are the scheduler's noise level and model time label. We average the per-sample ℓ_2 norms, without dimension normalization. Fig. 7 shows the comparison for seed 42: prediction norms are close at high noise and separate later in sampling. This late-stage difference is consistent across three noise seeds, with sampled latents exhibiting 10.4–10.8% lower mean prediction norms at the last measured pre-update state ($\sigma \approx 0.136$).

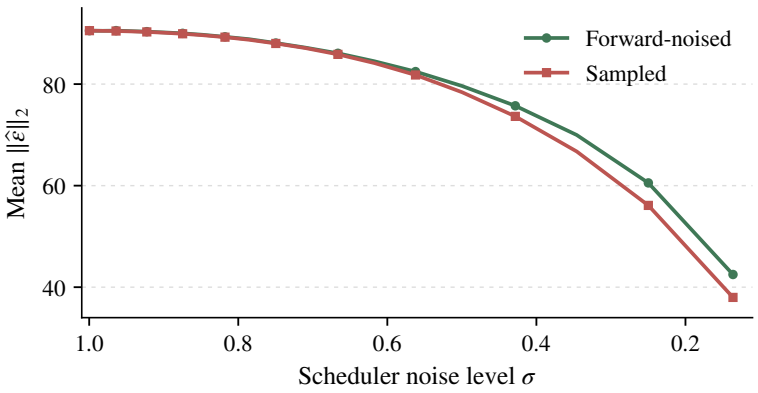


Figure 7: **Forward-noised and sampled latents at matched timesteps.** Mean noise-prediction norms of the frozen Sana-600M model over all 2,000 JourneyDB image–caption pairs (seed 42). Sampling proceeds from left to right.

The late-stage separation reveals different prediction statistics for sampled latents and the JourneyDB reference despite identical time labels.

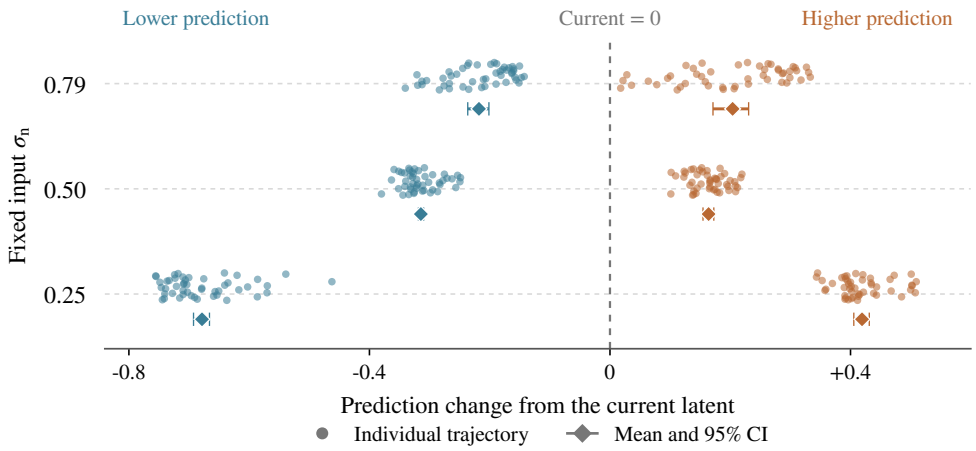


Figure 8: **State predictions follow relative latent progress.** With text and input timestep fixed, orange/blue dots show prediction changes from the current latent for earlier/later latents across 48 gate-disabled trajectories. Diamonds and bars show means and pointwise 95% block-bootstrap confidence intervals.

Table 5: **State-head responses at fixed timesteps.** Earlier/Later report mean changes from the current latent. Ordered counts triplets satisfying earlier $>$ current $>$ later.

σ_n	Earlier	Later	Ordered
≈ 0.79	$+0.204$	-0.218	48/48
≈ 0.50	$+0.164$	-0.314	48/48
≈ 0.25	$+0.419$	-0.678	48/48

D.2 State Sensitivity on Generated Trajectories

Setup. We next test whether StateWise responds to latent progress while the input timestep is held fixed. Using the final Stage 3 model, we collect 48 trajectories from 24 GenEval prompts with two seeds each, using 320–1,280 solver steps with the calibration gate disabled. At each fixed timestep ($\sigma_n \approx 0.79, 0.50, 0.25$), we read earlier, current, and later latents from neighboring nodes of a shared 20-step grid. Text and input timestep remain fixed within each triplet; only the latent changes.

Measurement and results. As at inference, we average the conditional and unconditional state-head predictions. Fig. 8 reports changes relative to the current latent, whose response is zero by construction. Pointwise 95% confidence intervals use 10,000 bootstrap resamples of eight six-sample blocks. Earlier latents yield higher predictions and later latents lower predictions in all 144 triplets (Tab. 5). This ordering shows that the state head reads relative latent progress beyond the unchanged timestep.