MT Forcing: Towards Stable and Consistent Multi-Turn Image Editing with Autoregressive Diffusion

Houyuan Chen[1, 2, *], Fanzhou Wang[1, *], Xiangyu Fan[1, †], Zesong Qiu[1], Yukun Wei[1], Songlin Yang[2]
Zhitao Yang[1], Lvmin Zhang[3], Pei Yang[1, 2], Hao Zhao[4, ‡], Lei Yang[1, ‡], Anyi Rao[2, ‡]

*Equal contribution, †Project lead, ‡Co-corresponding authors

[1]SenseTime Research [2]HKUST MMLab [3]Stanford University [4]AIR, Tsinghua University

Abstract

Multi-turn image editing aims to modify an image according to a sequence of instructions, enabling more effective interactive visual content creation for users. However, existing methods often exhibit noticeable drift after only a few editing turns, manifesting in two aspects: 1) ***unstable*** visual quality, with accumulating artifacts and degraded details; and 2) ***inconsistent*** image content, with unintended changes to preserved regions and prior edits. To address these challenges, we propose **MT Forcing**, a novel training paradigm for multi-turn image editing with unified multimodal models (UMMs) based on autoregressive diffusion. Specifically, we propose **Single-Turn D**istribution **M**atching **D**istillation (**Single-Turn DMD**), which supervises student self-rollouts with a teacher that applies cumulative edits to the source image in a single turn, providing visual-quality guidance without cross-turn error accumulation. Additionally, we introduce **Teacher-Forcing Guidance**, which supervises history-dependent transitions on reference trajectories during warmup and as a regularizer during distillation, encouraging cross-turn content consistency beyond what per-turn distribution matching alone can ensure. Extensive experiments demonstrate that our approach significantly improves the multi-turn image editing capabilities of UMMs, yielding more stable visual quality and stronger cross-turn consistency.

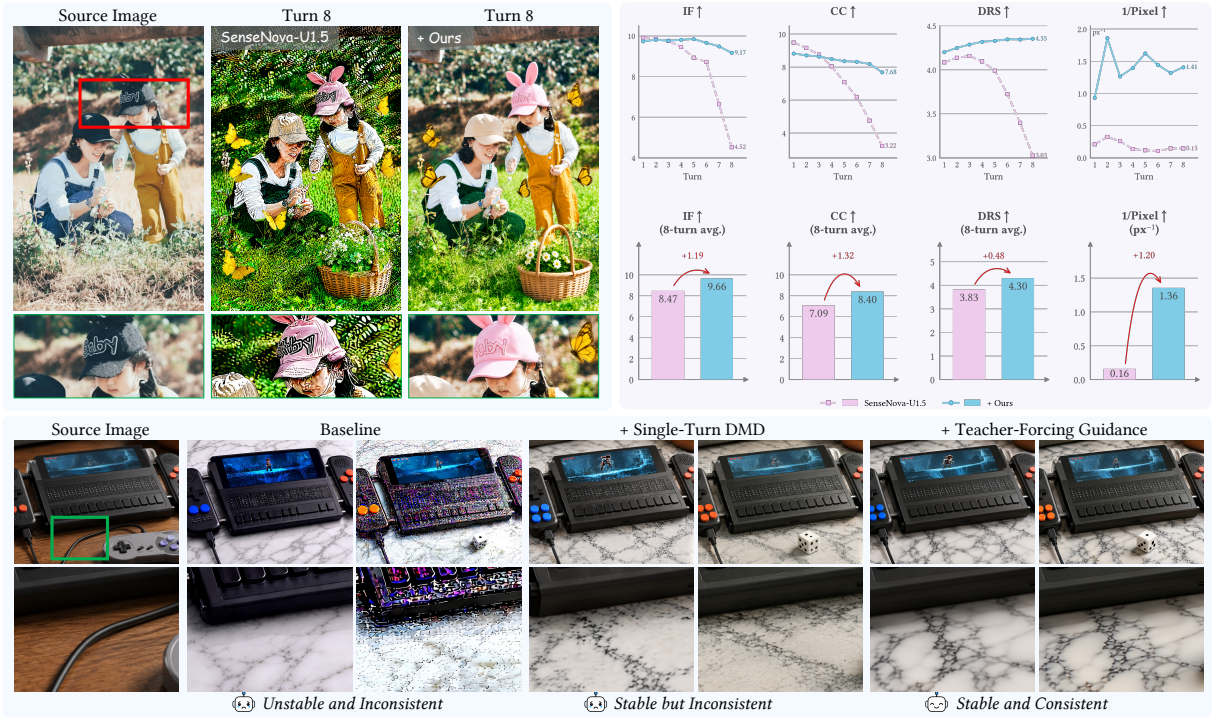


Figure 1. Overview of our approach. Top left: MT Forcing reduces artifacts after eight editing turns. Top right: turn-wise and eight-turn mean scores for SenseNova-U1.5 with and without MT Forcing. Bottom: qualitative comparisons illustrate the roles of Single-Turn DMD and Teacher-Forcing Guidance in improving visual quality and cross-turn consistency.

1 Introduction

Multi-turn image editing [27, 39, 60, 92, 112] aims to progressively modify an image through a sequence of user instructions. With rapid advances in unified multimodal models (UMMs) [11, 14, 66], this task has attracted growing attention from the research community. Compared with conventional single-turn editing [3, 9, 29, 35, 56, 67, 100], it enables users to continuously refine visual content as their creative intentions evolve, making it well suited to creative workflows requiring precise control over successive edits.

Since each edit builds on the results of preceding turns, multi-turn image editing is inherently causal [61, 86], which naturally aligns with autoregressive diffusion [12, 18, 22, 36] in UMMs. Such a design enables subsequent edits to draw on the accumulated interaction history while responding to new user instructions. However, existing methods often exhibit noticeable drift after only a few editing turns, manifesting in two aspects: 1) ***unstable*** visual quality, characterized by accumulating visual artifacts and progressive degradation of fine-grained textures and local structures; and 2) ***inconsistent*** image content, characterized by unintended alterations to regions outside the requested edit and failures to retain prior edits that should remain unchanged (See Fig. 1 for examples). We attribute these failures to exposure bias [59, 65], arising from the mismatch between training and inference conditions. During supervised training, the model typically conditions on ground-truth histories, whereas inference requires recursively conditioning on its own imperfect outputs.

To overcome these challenges, we propose **MT Forcing**, a training paradigm for UMMs based on AR diffusion that enables stable and consistent multi-turn image editing.

First, we introduce Single-Turn **D**istribution **M**atching **D**istillation (**Single-Turn DMD**) to improve visual quality stability. Our design is motivated by two observations: 1) editing models can execute multiple actions within a single turn [72], and 2) a single editing turn typically introduces limited visual degradation [71]. These observations suggest that we can use a model as a teacher for multi-turn editing by having it perform single-turn edits. Specifically, at each turn, we reorganize all editing instructions up to that turn into a single-turn editing prompt, termed a *bundled prompt*, and condition the teacher on this prompt and the source image. We use a DMD loss [15, 43, 94, 95] to distill the teacher into the student by matching their per-turn output distributions. By applying the bundled edits to the source image within a single turn, the teacher provides guidance free from cross-turn error accumulation. The student performs autoregressive self-rollouts during training [1, 28], conditioning each edit on its own imperfect outputs. This training teaches the student to maintain visual quality under imperfect self-generated histories using high-quality teacher guidance.

Additionally, to improve cross-turn consistency, we introduce **Teacher-Forcing Guidance**, which applies teacher-forced supervision [26, 30, 99] on ground-truth multi-turn editing trajectories in two stages. 1) *Teacher-Forcing Warmup*: before Single-Turn DMD training, we perform supervised fine-tuning to teach the model to produce consistent edits conditioned on coherent reference histories. 2) *Teacher-Forcing Regularization*: during Single-Turn DMD training, we retain this supervision with a lower weight as a regularizer to preserve cross-turn consistency while mitigating oversaturation and visual artifacts introduced during distillation [6, 78, 97].

Extensive quantitative and qualitative experiments demonstrate that MT Forcing achieves more stable visual quality, greater content consistency, and better instruction following across successive editing turns. We further introduce a metric based on image-matching [23] that quantifies pixel-level alignment. MT Forcing also achieves better alignment under this metric, providing additional evidence of improved content preservation over successive edits.

In summary, we 1) discuss visual quality degradation and content inconsistency in multi-turn image editing; 2) propose MT Forcing, in which Single-Turn DMD stabilizes visual quality and Teacher-Forcing Guidance preserves cross-turn content consistency; and 3) conduct extensive quantitative and qualitative experiments demonstrating improved visual quality stability, content consistency, and instruction following, and introduce a metric based on image matching to quantify pixel-level alignment across successive edits.

2 Related Work

Multi-Turn Image Editing extends instruction-based image editing [3, 25, 34, 45, 69, 72, 74, 79] to an interactive setting, where users progressively refine an image through a sequence of instructions. A straightforward approach is to recursively apply a single-turn editor, conditioning each turn on the previous output and the current instruction [27, 39, 101, 112]. However, this formulation does not explicitly model the full interaction history, limiting its ability to retain earlier constraints and resolve context-dependent instructions [91]. Recent approaches build on UMMs to incorporate previous instructions and generated images into a shared context for subsequent edits [11, 92]. Multi-turn

Consistent Image Editing [112] uses flow-based inversion and dual-objective guidance to stabilize sampling, while FreqEdit [39] preserves high-frequency features to reduce accumulated degradation. VINCIE [61] learns in-context editing from video-derived image-text sequences, and AnchorEdit [86] introduces causal self-rollout training and memory to preserve subject identity across turns. MT-EditFlow [27] optimizes rewards aggregated over editing trajectories, while Edit-R2 [92] reconstructs session intent and jointly optimizes reasoning and image generation. Most recently, MT-OPSD [101] constructs degraded inputs through identity rollouts and distills clean-conditioned teacher predictions into the student. However, the mismatch between training on reference histories and inference on self-generated outputs can cause errors to accumulate, making visual quality and content consistency difficult to maintain over successive turns [27, 101].

Diffusion Distillation aims to reduce iterative diffusion sampling to one or a few steps with minimal loss in generation quality. Teacher-trajectory regression learns the teacher's sampling dynamics through direct noise-to-image regression [50, 106] or progressive step compression [57, 62], while InstaFlow [46] straightens trajectories to facilitate one-step distillation. Consistency-based methods constrain predictions along a diffusion trajectory to share a common endpoint [52, 53, 70], with extensions to interval mappings [24, 32, 77, 107] and transitive or bootstrapped supervision [2, 21]. Continuous-time formulations improve scalability [47], while rCM [109] combines consistency distillation with score regularization. Distribution-level methods supervise generated samples through adversarial objectives [31, 40, 63, 64, 87] or score-based matching [54, 58, 111]. DMD [95] matches distributions using the difference between target and student-distribution scores, while DMD2 [94] stabilizes online score estimation and incorporates adversarial supervision. Subsequent work improves distillation quality and scalability [10, 19, 48, 110], refines multi-step matching and guidance [15, 43, 81], and explores alternative divergence objectives [88]. Auxiliary student-score estimation can also be replaced by generator-derived surrogates [33] or discriminative density-ratio estimation [98]. We use DMD to transfer single-turn editing quality to outputs under imperfect multi-turn histories.

Autoregressive Diffusion generates visual content in a causal autoregressive manner. For images, UMMs [11, 13, 14, 17, 20, 66, 82, 83] extend sequence modeling to visual understanding, generation, and editing, often combining autoregressive context modeling with diffusion-based image synthesis. For videos, a common strategy distills pretrained bidirectional diffusion models into few-step autoregressive generators, reducing latency and enabling streaming beyond the training clip length [28, 96]. CausVid [96] uses asymmetric distribution matching between a bidirectional teacher and a causal student. Diffusion Forcing [5] trains with independently noised sequence elements, whereas Self Forcing [28] trains on autoregressive self-rollouts to align training with inference. Other studies improve causal student initialization [108, 113] and scalable few-step or one-step distillation [16, 102]. Long-horizon stability is further strengthened through streaming training and error control [4, 8, 37, 44, 90, 114], long-range context modeling [7, 85, 89], and KV-cache management [51, 68, 93, 104]. Further work improves distillation quality and diversity [49, 105] and reduces inference costs through sparse or multi-scale attention [38, 55, 84]. However, directly transferring these methods to multi-turn image editing is challenging, as bidirectional teachers capable of jointly generating complete editing trajectories are unavailable.

3 Method

MT Forcing combines distribution matching on self-generated histories with supervision on reference trajectories. Sec. 3.1 defines the multi-turn editing problem and its training–inference mismatch. Sec. 3.2 introduces Single-Turn DMD to supervise student rollouts with a single-turn teacher. Sec. 3.3 presents Teacher-Forcing Guidance to enhance Single-Turn DMD through reference supervision. Algorithm 1 summarizes the overall training procedure.

3.1 Problem Formulation

A multi-turn editing session consists of a source image x_{src} and N sequential instructions $c_{1:N}$. At turn i, the model G_θ generates x_i from the current instruction c_i and the preceding history $H_{<i} = (x_{\text{src}}, c_1, x_1, \ldots, c_{i-1}, x_{i-1})$.

$$p_\theta(x_{1:N} \mid x_{\text{src}}, c_{1:N}) = \prod_{i=1}^{N} p_\theta(x_i \mid H_{<i}, c_i). \tag{1}$$

The causal model caches key–value (KV) vectors of previous instructions and images for subsequent edits [41, 75].



Figure 2. Overview of the MT Forcing framework. Single-Turn DMD supervises student rollouts using a single-turn teacher, while Teacher-Forcing Guidance applies reference-trajectory supervision through warmup and regularization.

In multi-turn image editing, UMMs encounter different histories during teacher-forcing training and rollout inference. A reference trajectory $\tau = (x_{\text{src}}, \{(c_i, x_i^\star)\}_{i=1}^N) \sim \mathcal{D}_{\text{ref}}$ supplies $H_{<i}^\star = (x_{\text{src}}, c_1, x_1^\star, \ldots, c_{i-1}, x_{i-1}^\star)$ during teacher forcing, whereas during inference the model uses its own preceding outputs:

$$\begin{aligned} \text{Teacher-forcing training:} \quad & p_\theta(x_i \mid H_{<i}^\star, c_i), \\ \text{Self-rollout inference:} \quad & p_\theta(x_i \mid H_{<i}, c_i). \end{aligned} \tag{2}$$

Therefore, errors in generated histories can accumulate across turns, causing editing results to drift progressively.

3.2 Single-Turn Distribution Matching Distillation

Recent video-generation methods [8, 28] mitigate rollout drift by distilling from pretrained bidirectional teachers that jointly model multiple frames. Multi-turn editing lacks a corresponding bidirectional UMM that jointly generates instruction-conditioned editing trajectories. We therefore propose Single-Turn DMD, in which a single-turn editing teacher supervises each generated turn; we next specify its conditioning and training signal.

At turn i, we reorganize all editing instructions up to that turn, $c_{1:i}$, into a single-turn editing prompt, termed the *bundled prompt* $\bar{c}_i = \text{Reorganize}(c_{1:i})$. This prompt expresses the cumulative editing intent as one edit to be applied directly to the source image. For example, "add a vase" followed by "make it blue" can be reorganized as "add a blue vase." We use $\mathcal{C}_i^{\text{S}} = (H_{<i}, c_i)$ and $\mathcal{C}_i^{\text{T}} = (x_{\text{src}}, \bar{c}_i)$ as the student and teacher conditions, respectively. The student recursively samples $x_i \sim p_\theta(\cdot \mid \mathcal{C}_i^{\text{S}})$ and appends (c_i, x_i) to its history. Since $\mathcal{C}_i^{\text{S}}$ contains the source image and all instructions up to turn i, prompt reorganization maps it to $\mathcal{C}_i^{\text{T}}$, but this mapping is not invertible:

$$\mathcal{C}_i^{\text{S}} \Rightarrow \mathcal{C}_i^{\text{T}}, \qquad \mathcal{C}_i^{\text{T}} \not\Rightarrow \mathcal{C}_i^{\text{S}}. \tag{3}$$

Thus, the student has a *stronger condition* and the frozen teacher T_ϕ has a *weaker condition*.

Fig. 3 shows clear teacher outputs at Turns 1, 5, and 8 for bundled prompts formed from accumulated editing instructions. By realizing cumulative editing intent in a single turn while maintaining visual quality across editing stages, the teacher therefore can provide effective per-turn distributional supervision for the student.

How can a teacher with weaker condition train the student? The connection is provided by the distribution induced by the student's causal rollout. The student generates each image from $p_\theta(\cdot \mid \mathcal{C}_i^{\text{S}})$, using the accumulated

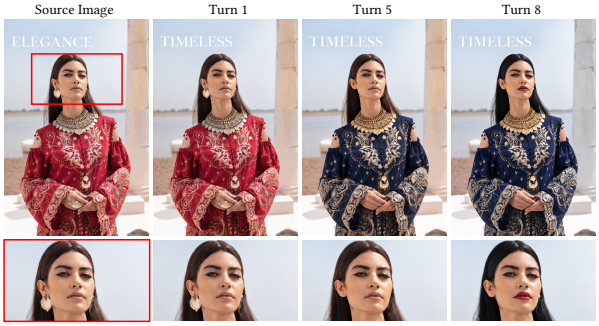


Figure 3. Visual quality across turns with a single-turn teacher. Each bundled prompt is applied independently to the source image. Enlarged crops show stable facial details.

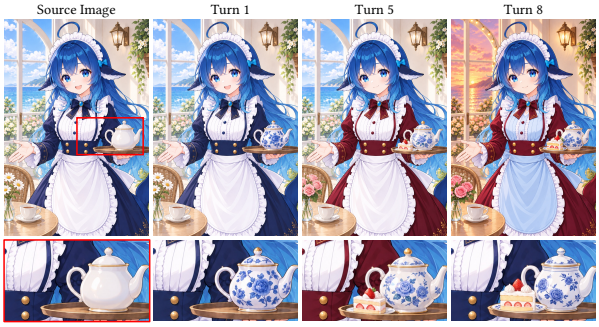


Figure 4. Content inconsistency of the single-turn teacher. The teapot's floral pattern changes across independent edits with bundled prompts, as highlighted in the enlarged crops.

history and the current instruction. For fixed $\mathcal{C}_i^{\mathrm{T}}$, let $p_{\theta,i}(\cdot \mid \mathcal{C}_i^{\mathrm{T}})$ denote the marginal distribution of x_i under the student's rollout, averaged over instruction sequences consistent with $\mathcal{C}_i^{\mathrm{T}}$.

$$p_{\theta,i}(x_i \mid \mathcal{C}_i^{\mathrm{T}}) = \mathbb{E}_{c_{1:i}|\mathcal{C}_i^{\mathrm{T}}}\left[\int p_\theta(x_i \mid \mathcal{C}_i^{\mathrm{S}}) \prod_{j<i} p_\theta(x_j \mid \mathcal{C}_j^{\mathrm{S}})\, dx_{1:i-1}\right]. \tag{4}$$

Following DMD [54, 94, 95], we apply the same forward process $x_i^t = \alpha_t x_i + \sigma_t \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, to both distributions. Image variables may denote pixels or a shared latent representation. Writing the resulting distributions as $p_{\theta,i,t}$ and $p_{\phi,i,t}$, we define the per-turn loss for a fixed $\mathcal{C}_i^{\mathrm{T}}$ as

$$\mathcal{L}_{\mathrm{DMD}}^{(i)} = \mathbb{E}_t\left[D_{\mathrm{KL}}\left(p_{\theta,i,t}(\cdot \mid \mathcal{C}_i^{\mathrm{T}}) \,\middle\|\, p_{\phi,i,t}(\cdot \mid \mathcal{C}_i^{\mathrm{T}})\right)\right], \tag{5}$$

where t follows a fixed distribution over nondegenerate noise levels. The KL compares the two noisy-image distributions under the same source image and bundled prompt. Through Eq. 4, optimizing this loss updates the student that actually generates under $\mathcal{C}_i^{\mathrm{S}}$, using a distribution-matching target expressed under $\mathcal{C}_i^{\mathrm{T}}$. The full objective averages these losses over turns and task conditions induced by sampled instruction sequences.

In practice, we optimize Eq. 5 through a score-based gradient estimate [94, 95]:

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}}^{(i)} \approx \mathbb{E}\left[\left(\frac{\partial x_i^t}{\partial \theta}\right)^\top \left(\hat{\mathbf{s}}_i^{\mathrm{F}} - \hat{\mathbf{s}}_i^{\mathrm{T}}\right)\right]. \tag{6}$$

Here $\hat{\mathbf{s}}_i^{\mathrm{T}}$ is provided by the frozen teacher T_ϕ, while $\hat{\mathbf{s}}_i^{\mathrm{F}}$ is estimated by the online fake model F_ψ, trained to denoise detached rollout outputs; both are conditioned on $\mathcal{C}_i^{\mathrm{T}}$. The student is updated by backpropagating this score difference through its generated image, treating the score estimates as fixed.

For efficiency, we truncate gradients across turns by detaching historical KV embeddings and retain gradients only through the sampled denoising step s_i of the current turn [28]. Thus, the partial derivative in Eq. 6 holds the history and incoming denoising state fixed. We denote the single-sample update direction inside the expectation by $\widehat{\mathbf{g}}_i$ and average these directions over turns for one student update after each rollout (Algorithm 1).

3.3 Teacher-Forcing Guidance

Why is Single-Turn DMD not enough? Single-Turn DMD trains the history-conditioned student through its rollout marginal, but matching this marginal does not uniquely determine the student's response to each history. The objective constrains the output distribution given $\mathcal{C}_i^{\mathrm{T}}$, leaving different dependencies on the additional visual history in $\mathcal{C}_i^{\mathrm{S}}$ indistinguishable. To see this, consider a student that ignores preceding generated images and independently samples from the teacher at each turn:

$$p_\theta(x_i \mid H_{<i}, c_i) = p_{\phi,i}(x_i \mid \mathcal{C}_i^{\mathrm{T}}). \tag{7}$$

Under Eq. 7, Eq. 4 gives $p_{\theta,i} = p_{\phi,i}$ and zero ideal DMD loss at every turn. Nevertheless, independent sampling can change contents across turns, even when every per-turn marginal distribution matches the teacher's. Thus, zero DMD loss does not guarantee consistent trajectories. The changing teapot pattern in Fig. 4 illustrates this ambiguity.

We introduce Teacher-Forcing Guidance to promote cross-turn consistency using reference trajectories $\tau \sim \mathcal{D}_{\text{ref}}$. Given the reference history $H^{\star}_{<i}$ and current instruction c_i, we supervise the target $x_i^{\star}$ through flow matching [42]:

$$\mathcal{L}_{\text{TF}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\tau,t_i,\epsilon_i}\left[\left\| v_\theta(x_i^{\star,t_i}, t_i, H^{\star}_{<i}, c_i) - v_i^{\star,t_i}\right\|_2^2\right], \tag{8}$$

where v_θ is the student's velocity prediction, $x_i^{\star,t_i} = \alpha_{t_i} x_i^{\star} + \sigma_{t_i}\epsilon_i$, and $v_i^{\star,t_i} = \dot{\alpha}_{t_i} x_i^{\star} + \dot{\sigma}_{t_i}\epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, I)$ and dots denoting time derivatives. Pairing each target with its reference history encourages the model to preserve contents across turns while following the current instruction. This complements Single-Turn DMD's marginal distribution matching by supervising history–output dependencies. We use $\mathcal{L}_{\text{TF}}$ in two stages: teacher-forcing warmup before distillation and teacher-forcing regularization during distillation.

Teacher-Forcing Warmup. Before distillation, we optimize $\mathcal{L}_{\text{TF}}$ to adapt the pretrained editor to reference multi-turn contexts. The checkpoint θ_{w} initializes the student G_θ, frozen teacher T_ϕ, and online fake model F_ψ. Warmup establishes history-dependent behavior, but an initialization does not constrain subsequent distribution-matching updates to retain it.

Teacher-Forcing Regularization. We retain reference supervision during distillation to penalize changes to reference-conditioned transitions that the marginal objective cannot distinguish. The combined objective is

$$\mathcal{L}_{\text{MT}} = \mathcal{L}_{\text{DMD}} + \lambda_{\text{TF}}\mathcal{L}_{\text{TF}}, \tag{9}$$

The update combines the averaged truncated DMD directions with $\lambda_{\text{TF}}\nabla_\theta\mathcal{L}_{\text{TF}}$. DMD supervises the weaker-conditioned output distribution on student rollouts, while Teacher-Forcing Regularization anchors the stronger history-dependent behavior on reference trajectories.

Algorithm 1 MT Forcing training procedure

Require: Denoise timesteps $\{t_1, \ldots, t_T\}$
Require: Reference editing trajectories $\mathcal{D}_{\text{ref}}$
Require: Pretrained editing model G_{θ_0}
Require: Number of editing turns N

Warm up G_{θ_0} on $\mathcal{D}_{\text{ref}}$ using $\mathcal{L}_{\text{TF}}$ (Eq. 8) to obtain θ_{w}
Initialize G_θ, T_ϕ, and F_ψ with weights θ_{w}
Freeze teacher T_ϕ
loop
  Sample $\tau = (x_{\text{src}}, \{(c_i, x_i^{\star})\}_{i=1}^{N}) \sim \mathcal{D}_{\text{ref}}$
  Initialize model output $\mathbf{X}_\theta \leftarrow [\,]$
  Disable gradient computation
  Initialize KV cache $\mathbf{KV} \leftarrow [G_\theta^{\text{KV}}(x_{\text{src}}; 0)]$
  for $i = 1, \ldots, N$ **do**
    Reorganize $c_{1:i}$ into a single-turn prompt $\bar{c}_i$ and set $\mathcal{C}_i^{\text{T}} = (x_{\text{src}}, \bar{c}_i)$
    Sample $s_i \sim \text{Uniform}(1, 2, \ldots, T)$
    Initialize $x_i^{t_T} \sim \mathcal{N}(0, I)$
    for $j = T, \ldots, s_i$ **do**
      if $j = s_i$ **then**
        Enable gradient computation
        Set $\hat{x}_i^0 \leftarrow G_\theta(x_i^{t_j}; t_j, \mathbf{KV}, c_i)$
        $\mathbf{X}_\theta$.append($\hat{x}_i^0$)
        Disable gradient computation
        Cache $\mathbf{kv}_i \leftarrow G_\theta^{\text{KV}}((c_i, \hat{x}_i^0); 0, \mathbf{KV})$
        $\mathbf{KV}$.append($\mathbf{kv}_i$)
      else
        Disable gradient computation
        Set $\hat{x}_i^0 \leftarrow G_\theta(x_i^{t_j}; t_j, \mathbf{KV}, c_i)$
        Sample $\epsilon \sim \mathcal{N}(0, I)$
        Set $x_i^{t_{j-1}} \leftarrow \Psi(\hat{x}_i^0, \epsilon, t_{j-1})$
      end if
    end for
    Compute truncated $\widehat{\mathbf{g}}_i$ using T_ϕ, F_ψ under $\mathcal{C}_i^{\text{T}}$ (Eq. 6)
  end for
  Update θ using $N^{-1}\sum_i \widehat{\mathbf{g}}_i + \lambda_{\text{TF}}\nabla_\theta\mathcal{L}_{\text{TF}}$
  Update ψ by denoising detached $\mathbf{X}_\theta$ with the corresponding $\mathcal{C}_i^{\text{T}}$
end loop

4 Experiments

In this section, we present experimental results for MT Forcing on multi-turn image editing. We first describe the baselines, benchmark, and evaluation metrics in Sec. 4.1. We then present quantitative and qualitative comparisons in Sec. 4.2. Finally, in Sec. 4.3, we conduct ablation studies to assess the contribution of each component in MT Forcing.

4.1 Experimental Setup

For baselines, we evaluate MT Forcing on two UMMs: SenseNova-U1 [14] and SenseNova-U1.5 [13]. We compare each editor's performance with and without MT Forcing. Ablation studies use SenseNova-U1.5 [13] as the baseline.

For benchmarking, we primarily evaluate on the step-wise sub-bench of MTEdit-Bench [76], which contains 100 sessions, each comprising a source image and eight sequential editing instructions.

For metrics, we follow MTEdit-Bench [76] and report Instruction Following (IF), Content Consistency (CC), and Drift Robustness Score (DRS). IF measures fulfillment of the current instruction, while CC evaluates preservation

Table 1. Quantitative results on the step-wise sub-bench. Results compare each base editor with MT Forcing across eight turns. Higher IF, CC, and DRS, and lower Pixel, indicate better performance.

Method	Turn 1				Turn 2				Turn 3				Turn 4			
	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓
SenseNova-U1 [14]†	9.29	**9.06**	4.07	6.63	8.95	8.51	4.14	13.19	8.53	7.61	4.12	12.28	7.37	5.95	3.90	3.37
+ MT Forcing	**9.71**$_{+.42}$	8.85$_{-.21}$	**4.25**$_{+.18}$	**1.23**$_{-5.40}$	**9.78**$_{+.83}$	**8.55**$_{+.04}$	**4.23**$_{+.09}$	**0.39**$_{-12.80}$	**9.65**$_{+1.12}$	**8.57**$_{+.96}$	**4.23**$_{+.11}$	**0.38**$_{-11.90}$	**9.56**$_{+2.19}$	**8.47**$_{+2.52}$	**4.25**$_{+.35}$	**0.35**$_{-3.02}$
SenseNova-U1.5 [13]	**9.90**	**9.49**	4.09	4.74	**9.85**	**9.16**	4.14	3.10	9.76	**8.76**	4.16	3.86	9.46	8.04	4.09	7.30
+ MT Forcing	9.74$_{-.16}$	8.82$_{-.67}$	**4.20**$_{+.11}$	**1.07**$_{-3.67}$	9.81$_{-.04}$	8.71$_{-.45}$	**4.25**$_{+.11}$	**0.54**$_{-2.56}$	**9.77**$_{+.01}$	8.64$_{-.12}$	**4.29**$_{+.13}$	**0.79**$_{-3.07}$	**9.81**$_{+.35}$	**8.49**$_{+.45}$	**4.32**$_{+.23}$	**0.72**$_{-6.58}$
Method	**Turn 5**				**Turn 6**				**Turn 7**				**Turn 8**			
	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓	IF↑	CC↑	DRS ↑	Pixel ↓
SenseNova-U1 [14]†	6.21	4.41	3.55	11.06	4.50	3.06	3.14	2.42	2.96	1.96	2.84	0.72	1.67	1.12	2.58	1.17
+ MT Forcing	**9.54**$_{+3.33}$	**8.19**$_{+3.78}$	**4.23**$_{+.68}$	**0.32**$_{-10.74}$	**9.55**$_{+5.05}$	**8.14**$_{+5.08}$	**4.20**$_{+1.06}$	**0.34**$_{-2.08}$	**9.16**$_{+6.20}$	**7.95**$_{+5.99}$	**4.21**$_{+1.37}$	**0.35**$_{-.37}$	**9.12**$_{+7.45}$	**7.94**$_{+6.82}$	**4.22**$_{+1.64}$	**0.37**$_{-.80}$
SenseNova-U1.5 [13]	8.93	7.09	3.99	8.58	8.72	6.18	3.72	9.44	6.64	4.75	3.40	6.84	4.52	3.22	3.03	6.89
+ MT Forcing	**9.85**$_{+.92}$	**8.36**$_{+1.27}$	**4.33**$_{+.34}$	**0.62**$_{-7.96}$	**9.66**$_{+.94}$	**8.32**$_{+2.14}$	**4.35**$_{+.63}$	**0.69**$_{-8.75}$	**9.49**$_{+2.85}$	**8.19**$_{+3.44}$	**4.34**$_{+.94}$	**0.76**$_{-6.08}$	**9.17**$_{+4.65}$	**7.68**$_{+4.46}$	**4.35**$_{+1.32}$	**0.71**$_{-6.18}$

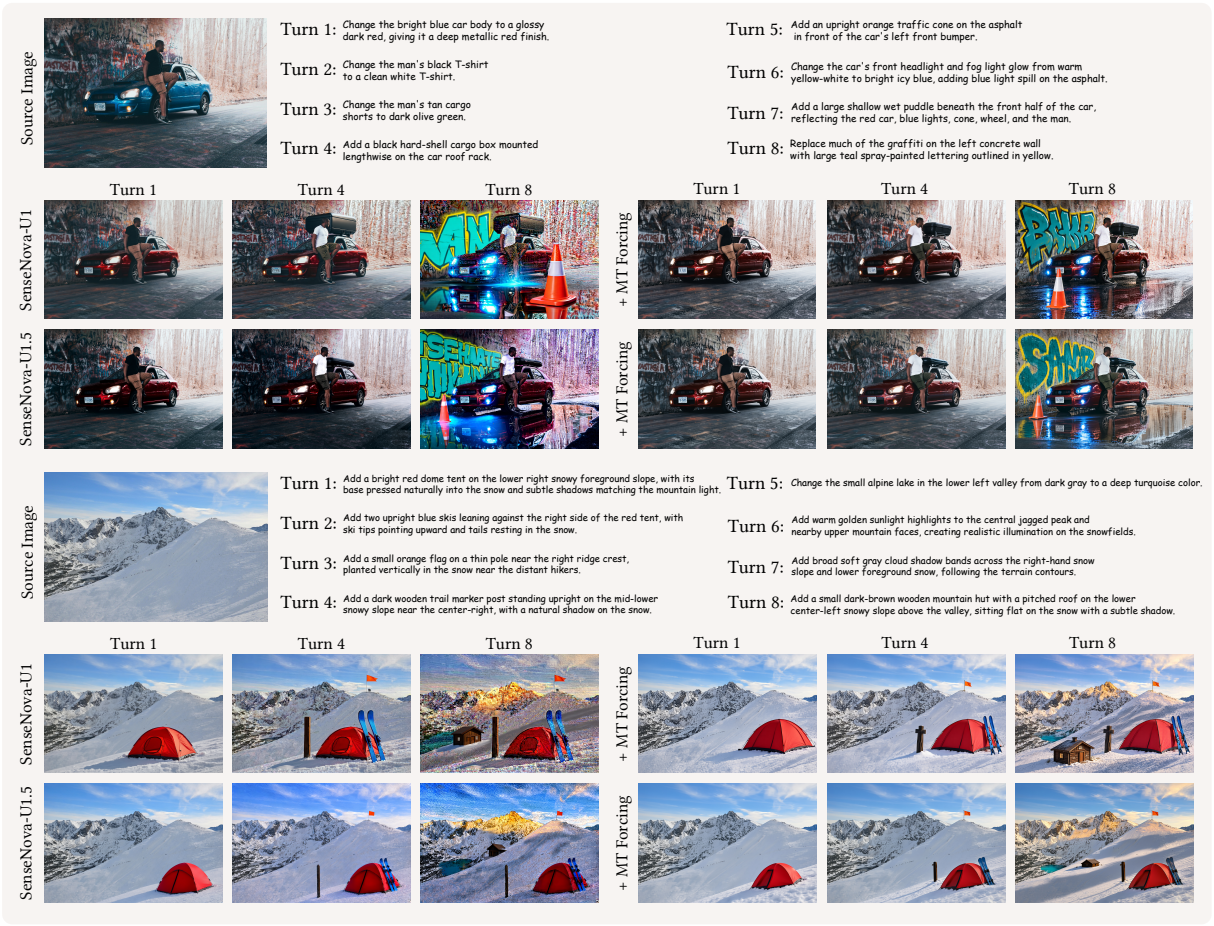


Figure 5. Qualitative results on the step-wise sub-bench. SenseNova-U1 and SenseNova-U1.5 are compared before (left) and after (right) MT Forcing at Turns 1, 4, and 8. MT Forcing reduces artifacts while preserving accumulated edits.

of the intended cumulative content under all instructions up to the current turn. Both are scored from 0 to 10 by Qwen3.5-Plus [73]. DRS measures robustness against visual-quality drift across turns, computed as the mean of the VisualQuality-R1 [80] and RALI [103] scores, and ranges from 0 to 5. Higher values indicate better performance for all three metrics. Moreover, we introduce Pixel to quantify geometric alignment between successive editing outputs. Using MatchAnything-RoMa [23], we compute the mean matched-point displacement in pixels as $\text{Pixel}_i = |\mathcal{M}_i|^{-1} \sum_{(u,v)\in\mathcal{M}_i} \|u - v\|_2$, where $\mathcal{M}_i$ contains valid matches between x_{i-1} and x_i, with $x_0 = x_{\text{src}}$. Lower values indicate less geometric change in matched regions.

4.2 Main Results

Quantitative comparison. Table 1 shows that MT Forcing outperforms both base editors on all four metrics from Turn 4 onward, with increasingly pronounced gains in IF, CC, and DRS at later turns. At Turn 8, IF and CC improve by 7.45 and 6.82 on SenseNova-U1, and by 4.65 and 4.46 on SenseNova-U1.5, respectively. On SenseNova-U1.5,

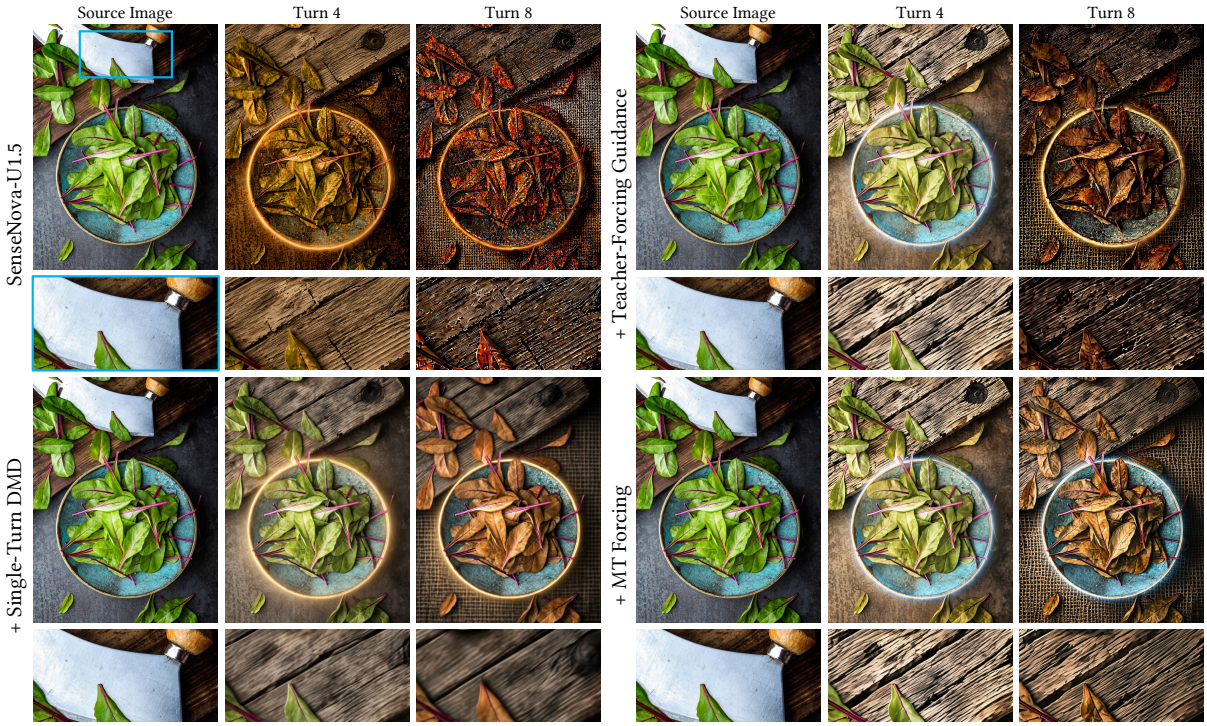


Figure 6. Qualitative ablation of MT Forcing. SenseNova-U1.5 is compared with Single-Turn DMD only, Teacher-Forcing Guidance only, and the full method at Turns 4 and 8. Enlarged crops highlight differences in texture detail.

MT Forcing maintains DRS between 4.20 and 4.35 throughout the session, while the baseline declines from 4.09 to 3.03. Its eight-turn mean Pixel also decreases from 6.34 to 0.74, an 88.4% reduction in matched-point displacement.

Qualitative comparison. Fig. 5 compares SenseNova-U1 and SenseNova-U1.5 with their MT Forcing counterparts at Turns 1, 4, and 8. By Turn 8, the baselines exhibit oversaturated lighting and harsh textures in the car scene, and speckled snow and exaggerated contrast in the mountain scene. MT Forcing preserves cleaner textures and coherent scene structure across turns while retaining earlier edits, including the red car body, roof box, tent, skis, and marker post. These examples visually support the improved drift robustness and content consistency measured in Table 1.

4.3 Ablation Studies

Table 2 and Fig. 6 compare the pretrained baseline, Single-Turn DMD only, Teacher-Forcing Guidance only, and full MT Forcing. The DMD-only variant initializes all three models from the pretrained editor and omits reference supervision. The guidance-only variant uses teacher-forcing warmup followed by continued reference supervision, without DMD. The full method combines warmup with DMD and teacher-forcing regularization.

Table 2. Quantitative ablation results on the step-wise sub-bench. Eight-turn mean scores on SenseNova-U1.5.

Single-Turn DMD	Teacher-Forcing Guidance	IF↑	CC↑	DRS ↑	Pixel ↓
×	×	8.47	7.09	3.83	6.34
✓	×	9.02	7.64	4.12	2.51
×	✓	9.10	7.83	3.92	2.74
✓	✓	9.66	8.40	4.30	0.74

Effect of Single-Turn DMD. DMD alone improves DRS from 3.83 to 4.12 over the pretrained baseline. Adding Single-Turn DMD to the guidance-only variant improves IF, CC, and DRS by 0.56, 0.57, and 0.38, respectively, while reducing Pixel from 2.74 to 0.74 (Table 2). In Fig. 6, guidance alone still produces harsh wood-grain textures and strong edge contrast by Turn 8, whereas the full method retains cleaner surfaces and leaf details. These results support the benefit of distribution matching on self-generated histories beyond teacher-forced supervision alone.

Effect of Teacher-Forcing Guidance. Compared with DMD alone, full MT Forcing improves IF and CC by 0.64 and 0.76, raises DRS from 4.12 to 4.30, and reduces Pixel from 2.51 to 0.74. Fig. 6 further shows that DMD alone reduces the baseline's severe texture artifacts, but the wood grain becomes blurred and fine detail is lost. Combining DMD with Teacher-Forcing Guidance retains sharper wood grain across the displayed turns while avoiding the baseline's exaggerated textures. This qualitative comparison illustrates the complementary roles of distribution matching and reference supervision in preserving visual quality and local detail.

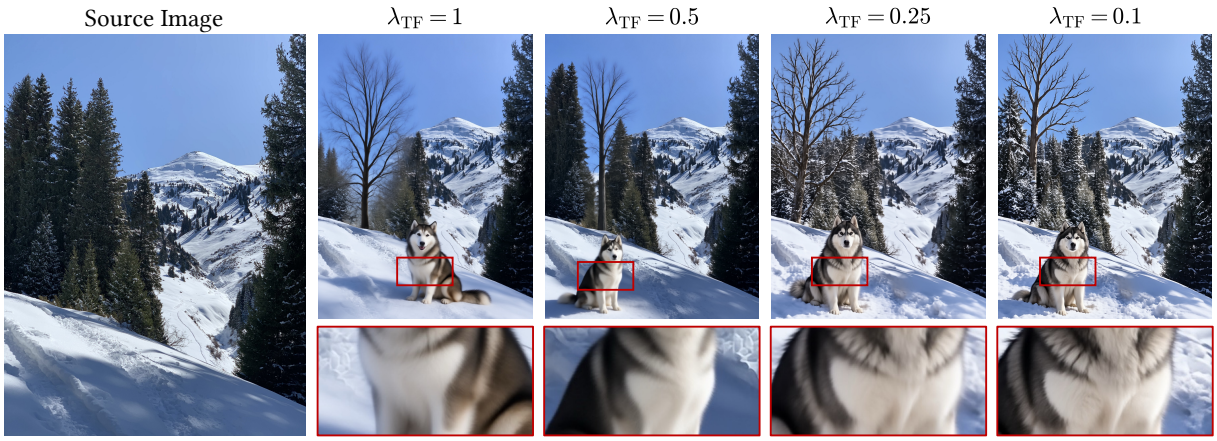


Figure 7. Visual quality comparison under different regularization weights. From left to right: the source image and outputs with $\lambda_{\text{TF}} = 1, 0.5, 0.25, 0.1$. Enlarged crops show differences in fur detail.

Effect of Teacher-Forcing Regularization. Fig. 7 compares outputs with $\lambda_{\text{TF}} = 1, 0.5, 0.25$, and 0.1. Although all four settings produce the husky in the snowy scene, the enlarged crops reveal smoother, less distinct fur at $\lambda_{\text{TF}} = 1$ and 0.5, and more clearly defined strands and coat boundaries at 0.25 and 0.1. This example suggests that excessive reference-loss weighting can limit fine-detail rendering, motivating a balance between teacher-forcing regularization and distribution matching.

5 Conclusion

We presented MT Forcing, a training paradigm for stable and consistent multi-turn image editing with autoregressive diffusion in unified multimodal models. Single-Turn DMD trains the student on self-generated editing histories using a single-turn teacher conditioned on the source image and a single-turn prompt reorganized from the accumulated instructions. Teacher-Forcing Guidance complements this supervision through warmup and regularization on reference trajectories to preserve cross-turn consistency. Experiments on the eight-turn step-wise split of MTEdit-Bench show improved visual quality, content consistency, and failure-aware instruction following relative to the base editor. An additional image-matching-based metric provides complementary evidence of improved geometric alignment across successive edits.

Our approach relies on reorganizing accumulated instructions into a single-turn prompt that the teacher can execute, and the current quantitative evaluation is limited to eight-turn step-wise sessions. Future work will explore longer editing sequences and more complex instructions that refer to earlier interaction states.

References

[1] 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 *ICLR*, 2024.

[2] David Berthelot, Arnaud Autef, Jierui Lin, Dian Ang Yap, Shuangfei Zhai, Siyuan Hu, Daniel Zheng, Walter Talbott, and Eric Gu. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[4] Tri Cao, Hung Nguyen, Phong Nguyen, and Khoi Nguyen. Recency forcing: Bridging the long-horizon gap in autoregressive video generation. *arXiv preprint arXiv:2609.19729*, 2026.

[5] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NIPS*, 2024.

[6] Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

[7] Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

[8] Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *ICLR*, 2026.

[9] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, Yueze Wang, Chengyuan Wang, Fan Zhang, Yingli Zhao, Ting Pan, Xianduo Li, Zecheng Hao, Wenxuan Ma, Zhuo Chen, Yulong Ao, Tiejun Huang, Zhongyuan Wang, and Xinlong Wang. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.

[10] Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. In *ECCV*, 2024.

[11] 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.

[12] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *ICLR*, 2025.

[13] Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. Sensenova-u1.5: Towards native unified visual intelligence, 2026.

[14] Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

[15] Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *CVPR*, 2026.

[16] Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

[17] Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy, 2026.

[18] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024.

[19] Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. Senseflow: Scaling distribution matching for flow-based text-to-image distillation. *arXiv preprint arXiv:2506.00523*, 2025.

[20] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024.

[21] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML Workshop*, 2023.

[22] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *ICCV*, 2025.

[23] Xingyi He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. In *Arxiv*, 2025.

[24] Jonathan Heek, Emiel Hoogeboom, and Tim Salimans. Multistep consistency models. *arXiv preprint arXiv:2403.06807*, 2024.

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[26] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024.

[27] Jiahui Huang, Yasi Zhang, Tianyu Chen, Shu Wang, Jianwen Xie, Oscar Leong, Mingyuan Zhou, Nanzhu Wang, and Ying Nian Wu. Mt-editflow: Reinforcement learning for multi-turn image editing with flow matching. *arXiv preprint arXiv:2606.01985*, 2026.

[28] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *NIPS*, 2025.

[29] 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 *CVPR*, 2024.

[30] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025.

[31] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling diffusion models into conditional gans. In *ECCV*, 2024.

[32] Dongjun Kim, Chieh-Hsin Lai, WeiHsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *ICLR*, 2024.

[33] Youngjoong Kim, Deokyeong Lee, and Jaesik Park. Distribution matching distillation without fake score network. *arXiv preprint arXiv:2605.19256*, 2026.

[34] 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.

[35] Mingsong Li, Lin Liu, Hongjun Wang, Haoxing Chen, Xijun Gu, Shizhan Liu, Dong Gong, Junbo Zhao, Zhenzhong Lan, and Jianguo Li. Multiedit: Advancing instruction-based image editing on diverse and challenging tasks. *arXiv preprint arXiv:2509.14638*, 2025.

[36] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NIPS*, 2024.

[37] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025.

[38] Zekun Li, Xiaoyan Cong, Hongyu Li, Zhiyang Dou, Chuan Guo, Abhay Mittal, Sizhe An, and Srinath Sridhar. Ms. forcing: Efficient streaming video generation with multi-scale patchification and attention. *arXiv preprint arXiv:2607.20940*, 2026.

[39] Yucheng Liao, Jiajun Liang, Kaiqian Cui, Baoquan Zhao, Haoran Xie, Wei Liu, Qing Li, and Xudong Mao. Freqedit: Preserving high-frequency features for robust multi-turn image editing. In *CVPR*, 2026.

[40] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

[41] Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *NIPS*, 2025.

[42] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[43] Dongyang Liu, Gao Peng, David Liu, DU DU, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *ICLR*, 2026.

[44] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026.

[45] 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.

[46] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *ICLR*, 2023.

[47] Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *ICLR*, 2025.

[48] Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *ICCV*, 2025.

[49] Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *CVPR*, 2026.

[50] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

[51] Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, et al. Future forcing: Future-aware training-free kv cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

[52] 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.

[53] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick Von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[54] Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *NIPS*, 2023.

[55] Chengtao Lv, Yumeng Shi, Yushi Huang, Ruihao Gong, Shen Ren, and Wenya Wang. Light forcing: Accelerating autoregressive video diffusion via sparse attention. *arXiv preprint arXiv:2602.04789*, 2026.

[56] Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. *arXiv preprint arXiv:2508.15772*, 2025.

[57] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

[58] Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In *CVPR*, 2024.

[59] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *ICLR*, 2024.

[60] OpenAI. Introducing ChatGPT Images 2.5. https://openai.com/index/introducing-chatgpt-images-2-5/, 2026.

[61] Leigang Qu, Feng Cheng, Ziyan Yang, Qi Zhao, Shanchuan Lin, Yichun Shi, Yicong Li, Wenjie Wang, Tat-Seng Chua, and Lu Jiang. Vincie: Unlocking in-context image editing from video. In *ICLR*, 2026.

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

[63] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia*, 2024.

[64] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024.

[65] Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019.

[66] SenseNova. Neo-unify: Building native multimodal unified models end to end, 2026.

[67] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *CVPR*, 2024.

[68] Shuwei Shi, Zhen Li, Muyao Niu, Chuanhao Li, Bo Zheng, Kaipeng Zhang, and Yinqiang Zheng. Surprise forcing: What to remember, when to skip in long video generation. *arXiv preprint arXiv:2607.18436*, 2026.

[69] Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, Guohui Zhang, Wenhu Zhang, Hang Xu, Nan Jiang, Xin Han, Haoze Sun, Maoquan Zhang, Haoyang Huang, and Nan Duan. Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

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

[71] Kenan Tang, Praveen Arunshankar, Andong Hua, Anthony Yang, and Yao Qin. Banana100: Breaking nr-iqa metrics by 100 iterative image replications with nano banana pro. *arXiv preprint arXiv:2604.03400*, 2026.

[72] Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

[73] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026.

[74] Super Intelligence Team. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

[75] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[76] TODO: authors. MTEdit: TODO: full paper title. Unpublished manuscript. TODO: complete bibliographic information, TODO.

[77] Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *NIPS*, 2024.

[78] Zimo Wang, Junkun Yuan, Angtian Wang, Haotian Yang, Canyu Zhang, Siyuan Yuan, Xingchang Huang, Bo Liu, Yizhi Wang, Yiding Yang, et al. Pdmd: Projected distribution matching distillation for video diffusion models. *arXiv preprint arXiv:2609.35768*, 2026.

[79] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025.

[80] Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *NIPS*, 2025.

[81] Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

[82] 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 *ICLR*, 2025.

[83] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *NIPS*, 2026.

[84] Boxun Xu, Yuming Du, Zichang Liu, Siyu Yang, Ziyang Jiang, Siqi Yan, Rajasi Saha, Albert Pumarola, Wenchen Wang, and Peng Li. Sparse forcing: Native trainable sparse attention for real-time autoregressive diffusion video generation. *arXiv preprint arXiv:2604.21221*, 2026.

[85] Haiyang Xu, Zheng Ding, and Zhuowen Tu. Recap-forcing: Retaining content appearances for long video generation. *arXiv preprint arXiv:2608.26671*, 2026.

[86] Hang Xu, Xiaoxiao Ma, Guohui Zhang, Yu Hu, Siming Fu, Jie Huang, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. Anchoredit: Maintaining temporal consistency in multi-turn image editing via causal memory. *arXiv preprint arXiv:2606.11751*, 2026.

[87] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *CVPR*, 2024.

[88] Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

[89] Bowen Xue, Brandon Y Feng, Chenguo Lin, Yuchen Lin, Yujia Zeng, Lvmin Zhang, Maneesh Agrawala, Honglei Yan, and Panwang Pan. Ring forcing: Towards precise long-term memory for autoregressive video diffusion. In *ECCV*, 2026.

[90] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[91] Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *NIPS*, 2025.

[92] Yuxiao Ye, Haoran He, Fangyuan Kong, Xintao Wang, Pengfei Wan, Kun Gai, and Ling Pan. Edit-r2: Context-aware reinforcement learning for multi-turn image editing. *arXiv preprint arXiv:2606.05950*, 2026.

[93] Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

[94] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *NIPS*, 2024.

[95] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

[96] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

[97] Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. *arXiv preprint arXiv:2603.21864*, 2026.

[98] Zhengming Yu, Junkun Yuan, Haotian Yang, Gordon Guocheng Qian, Yizhi Wang, Angtian Wang, Yiding Yang, Bo Liu, Xin Li, Wenping Wang, and Chongyang Ma. DMAD: Distribution matching as adversarial distillation for fast visual generation. *arXiv preprint arXiv:2610.02188*, 2026.

[99] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William Freeman, and Hao Tan. Test-time training done right. In *ICLR*, 2026.

[100] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. *NIPS*, 2025.

[101] Liangbing Zhao, Le Zhuo, and Mohamed Elhoseiny. On-policy self-distillation for multi-turn image editing, 2026.

[102] Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026.

[103] Shijie Zhao, Xuanyu Zhang, Weiqi Li, Junlin Li, Li Zhang, Tianfan Xue, and Jian Zhang. Reasoning as representation: Rethinking visual reinforcement learning in image quality assessment. *arXiv preprint arXiv:2510.11369*, 2025.

[104] Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026.

[105] Zhuoran Zhao, Shengju Qian, Tongtong Liang, Xianghao Kong, Songchun Zhang, Junchao Huang, Guian Fang, Xin Wang, Pan Hui, and Anyi Rao. Mask forcing: Improving autoregressive video diffusion distillation via dual-noise masking rollout. *arXiv preprint arXiv:2609.09123*, 2026.

[106] Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *ICLR*, 2023.

[107] Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation: Improved latent consistency distillation by semi-linear consistency function with trajectory mapping. *arXiv preprint arXiv:2402.19159*, 2024.

[108] Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026.

[109] Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *ICLR*, 2026.

[110] Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. *arXiv preprint arXiv:2410.14919*, 2024.

[111] Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *ICML*, 2024.

[112] Zijun Zhou, Yingying Deng, Xiangyu He, Weiming Dong, and Fan Tang. Multi-turn consistent image editing. In *ICCV*, 2025.

[113] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

[114] Yueting Zhu, Yuehao Song, Kaicheng Zhang, Bao Tang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Stream forcing: Constructing unified training trajectory for robust streaming video generation. *arXiv preprint arXiv:2608.10439*, 2026.