

Vision Bridge Transformer at Scale

Zhenxiong Tan¹, Zeqing Wang¹, Xingyi Yang², Songhua Liu³, and
Xinchao Wang^{1*}

¹ National University of Singapore

² The Hong Kong Polytechnic University

³ Shanghai Jiao Tong University

Abstract. We introduce **Vision Bridge Transformer (ViBT)**, a large-scale instantiation of Brownian Bridge Models designed for conditional generation. Unlike traditional diffusion models that transform noise into data, Bridge Models offer an efficient data-to-data translation paradigm by directly modeling the trajectory between inputs and outputs. We first demonstrate the advantages of this paradigm through a controlled apples-to-apples comparison against diffusion models, revealing improved efficiency and performance in this setting. Building on this, we successfully scale the Bridge paradigm up to a 20B-parameter model to tackle complex image and video translation tasks. To enable stable training at this scale, we adopt a Transformer architecture equipped with a variance-stabilized velocity matching objective. Together, these advances unlock the potential of scaling Bridge Models, establishing strong performance for instruction-based image editing and video stylization.

1 Introduction

Generative models have advanced remarkably over the past few years. Diffusion models [18, 34, 51] and large-scale Transformer architectures [14, 44] have together produced systems capable of generating high-quality images [8, 63] and videos [55, 56, 58, 61, 76] with strong fidelity. Extending these models to conditional generation tasks, where outputs are shaped by a given source image or video, has naturally followed [53, 54, 67, 73]. Typically, these approaches inject visual conditions into the generative process by incorporating source images or videos as auxiliary inputs [67, 73].

Despite their success, the underlying *noise-to-vision* modeling paradigm widely adopted by these models [14, 15] can be unnatural for conditional generation tasks. In this paradigm, models start from noise and gradually refine it toward the target [8, 29, 58, 63]. However, in many conditional scenarios (e.g., image editing [19, 47, 70], colorization, and frame interpolation), inputs already closely resemble the desired outputs, making this process unintuitive. Moreover, incorporating additional conditioning tokens introduces substantial computational overhead under transformer architectures [53, 54], especially in video settings [1, 22, 58].

* Corresponding author (xinchao@nus.edu.sg)

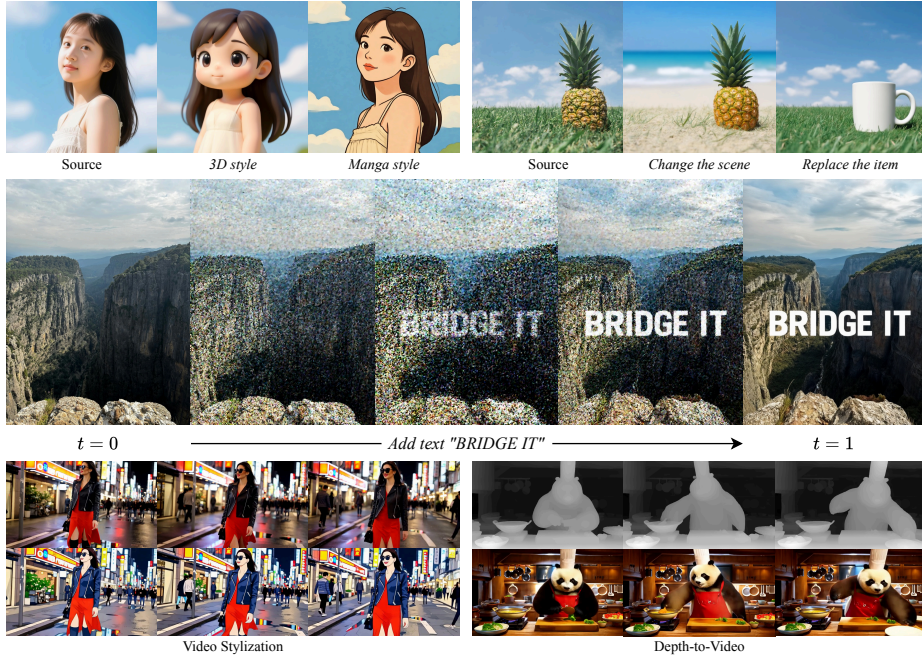


Fig. 1: Results of vision bridge transformer on various vision translation tasks.

A more direct alternative is the *data-to-data* paradigm, which models the transformation from source to target directly, without starting from pure noise [31, 35, 77]. Brownian Bridge based formulations provide a clean stochastic framework for this, and have shown early promise [31]. However, prior work has been confined to small architectures and simple tasks [5, 71, 75], leaving their potential for complex vision translation scenarios underexplored. To explore this potential, we first conduct a small-scale apples-to-apples comparison on image editing, training a Conditional DiT (noise-to-data) and a Bridge model (data-to-data) from the identical base model under identical conditions. This validates the Bridge model’s advantages in learning dynamics under this controlled setting: faster convergence and stronger performance at lower computational cost.

Motivated by this, we introduce *Vision Bridge Transformer* (ViBT), which scales the Bridge paradigm to large Transformer architectures initialized from state-of-the-art generative models [58, 63]. To train stably at this scale, we observe that conventional displacement-style training objectives [31] disproportionately bias training toward early generation steps, while naive velocity-based objectives [5, 38] exhibit severe numerical instability, negatively affecting convergence and performance. To address these issues, we propose a *variance-stabilized velocity matching* objective, which maintains numerical stability and balances target scales across timesteps, facilitating efficient training at scale.

We validate ViBT across a range of complex conditional generation tasks. ViBT achieves strong performance on video stylization and depth-to-video translation while maintaining temporal consistency. Applied to instruction-based image editing, ViBT scales seamlessly to a 20B-parameter base model through paradigm adaptation alone, achieving competitive results with a $2.3\times$ inference speedup. Ablation studies further confirm the importance of our training objective and the role of noise scale in the Bridge formulation.

2 Related Works

Generative models Early generative models, such as Variational Autoencoders (VAEs) [46] and Generative Adversarial Networks (GANs) [15, 25], enabled high-quality synthesis through latent modeling and adversarial training. Diffusion models [18, 49] later introduced iterative denoising processes, significantly advancing image and video generation. Flow-matching models [14, 34] further re-framed generation as learning deterministic or stochastic paths between distributions. More recently, transformer-based architectures trained at scale have further enhanced the fidelity and diversity of generative models [44, 58, 63], with linear-attention variants further improving their efficiency [37].

Conditional generation Conditional diffusion models incorporate auxiliary signals such as text, images, poses, or depth maps through additional encoders, auxiliary branches, or cross-attention mechanisms. Representative methods include ControlNet [73], IP-Adapter [67], and T2I-Adapters [42]. With the emergence of Diffusion Transformers (DiT) [44], recent approaches [29, 39, 53, 63] incorporate conditions directly into transformer attention layers for stronger guidance. However, these methods introduce substantial computational overhead, especially in video tasks.

Bridge models Bridge Models [31, 35, 77] construct stochastic processes directly connecting source and target distributions, providing an alternative to noise-driven generation. Early approaches employed Schrödinger bridges [12], stochastic interpolants [2], and entropic optimal transport [11]. Recent diffusion-based variants, such as Denoising Diffusion Bridge Models (DDBM) [77] and Brownian Bridge methods [31], demonstrated promising results for conditional generation and image translation tasks. Several recent works have highlighted the potential of Bridge Models in vision tasks, demonstrating improved structural and stylistic consistency in exemplar-guided image translation [30], enhanced temporal coherence in video synthesis [57], and increased efficiency during training and inference for basic image translation tasks [5, 21].

3 Preliminaries

Probability path modeling [12, 34, 38] defines a class of generative models that describe continuous-time processes transporting mass from a prior distribution

p_0 to a target distribution p_1 . Generally, these models are represented by a stochastic differential equation (SDE):

$$dX_t = v(X_t, t) dt + \sigma(t) dW_t, \quad t \in [0, 1], \quad (1)$$

with boundary conditions $X_0 \sim p_0$ and $X_1 \sim p_1$, velocity field $v : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$, diffusion coefficient $\sigma : [0, 1] \rightarrow \mathbb{R}_{\geq 0}$, and standard Brownian motion W_t .

In practice, the velocity field is typically parameterized by a neural network v_θ , trained via a velocity-matching objective [34, 38]:

$$\mathcal{L}(\theta) = \mathbb{E}_{(x_0, x_1), t, X_t} [\|v_\theta(X_t, t) - u_t(X_t | x_0, x_1)\|^2], \quad (2)$$

where $u_t(\cdot | x_0, x_1)$ denotes the instantaneous velocity induced by a chosen teacher trajectory (deterministic or stochastic), and X_t is sampled accordingly. Throughout, we use uppercase X_t for stochastic states and lowercase x_t for deterministic trajectories.

Rectified Flow Rectified Flow [14, 38] is a deterministic realization of probability path modeling obtained by setting $\sigma(t) \equiv 0$ in Eq. (1). It defines linear deterministic trajectories connecting endpoints $x_0 \sim p_0$, typically sampled from a standard Gaussian distribution $\mathcal{N}(0, I)$, and $x_1 \sim p_1$, drawn from the data distribution:

$$x_t = (1 - t)x_0 + tx_1. \quad (3)$$

Under this linear interpolation, the instantaneous velocity target simplifies to a constant vector:

$$u_t \equiv x_1 - x_0. \quad (4)$$

Brownian Bridge In contrast to the deterministic Rectified Flow, the standard Brownian Bridge [2, 31] incorporates stochasticity via a constant diffusion coefficient $\sigma(t) \equiv 1$. Given fixed endpoints (x_0, x_1) , its conditional intermediate states follow a Gaussian distribution:

$$X_t | (x_0, x_1) \sim \mathcal{N}\left(\underbrace{(1 - t)x_0 + tx_1}_{\text{linear interpolation}}, \underbrace{t(1 - t)I}_{\text{maximal variance at } t=0.5}\right). \quad (5)$$

Brownian Bridges are particularly suited to data-to-data transport tasks, such as denoising corrupted samples or translating data between structured image and video domains. Pairs of endpoints (x_0, x_1) are sampled from their respective source and target distributions. Under this stochastic formulation, the instantaneous velocity target used in velocity matching is expressed as:

$$u_t(X_t | x_0, x_1) = \frac{x_1 - X_t}{1 - t}. \quad (6)$$

4 Methodology

Our method leverages a transformer-based architecture to model vision translation tasks in latent space. Given paired source and target data (images or videos), we encode them into latent representations $x_0 \sim p_{\text{source}}$ and $x_1 \sim p_{\text{target}}$ using a pre-trained VAE encoder [27], and apply the Brownian Bridge formulation to directly model the transformation from x_0 to x_1 .

4.1 Stabilized velocity matching

During training, given latent endpoint pairs $(x_0, x_1) \sim p_{\text{source, target}}$, we uniformly sample a time $t \in [0, 1]$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. According to the conditional distribution of the Brownian bridge (5), the intermediate latent state x_t is constructed as:

$$x_t = (1 - t)x_0 + tx_1 + \sqrt{t(1 - t)}\epsilon. \quad (7)$$

The velocity-based training target at this noisy state, derived from Eq. (6), is given by:

$$u_t(x_t|x_1) = \frac{x_1 - x_t}{1 - t} = (x_1 - x_0) - \sqrt{\frac{t}{1 - t}}\epsilon. \quad (8)$$

Accordingly, the training objective is given by the mean squared error between the predicted and target velocities:

$$\mathcal{L}_{\text{velocity}} = \mathbb{E}_{t, \epsilon, x_0, x_1} \left[\|v_\theta(x_t, t) - u_t(x_t|x_1)\|^2 \right]. \quad (9)$$

However, this objective faces critical issues as $t \rightarrow 1$: the target velocity $u_t(x_t|x_1)$ diverges as $O\left(\frac{1}{\sqrt{1-t}}\right)$, causing instability, and the loss excessively focuses on these states, neglecting intermediate ones (Fig. 2).

An alternative approach adopted in previous works [31] is to use a displacement-based training target defined as

$$d_t(x_t|x_1) = x_1 - x_t. \quad (10)$$

Accordingly, the displacement-based training objective is given by the mean squared error:

$$\mathcal{L}_{\text{displacement}} = \mathbb{E}_{t, \epsilon, x_0, x_1} \left[\|d_\theta(x_t, t) - d_t(x_t|x_1)\|^2 \right]. \quad (11)$$

This displacement formulation naturally avoids numerical divergence, as it remains stable across all timesteps. However, its magnitude diminishes as $t \rightarrow 1$ at the rate $O(\sqrt{1-t})$, causing the training loss to be dominated by samples at smaller values of t .

Motivated by the above numerical instability and imbalanced loss across timesteps, we propose *stabilized velocity matching*, which introduces a normalization factor $\alpha(x_0, x_1, t)$ to balance the conditional target scale across timesteps. We rescale the original velocity target as:

$$\tilde{u}_t(x_t|x_1) = \frac{u_t(x_t|x_1)}{\alpha(x_0, x_1, t)}. \quad (12)$$

Specifically, for non-degenerate endpoint pairs with $\|x_1 - x_0\| > 0$, we define $\alpha(x_0, x_1, t)$ based on the normalized root mean square magnitude of the velocity:

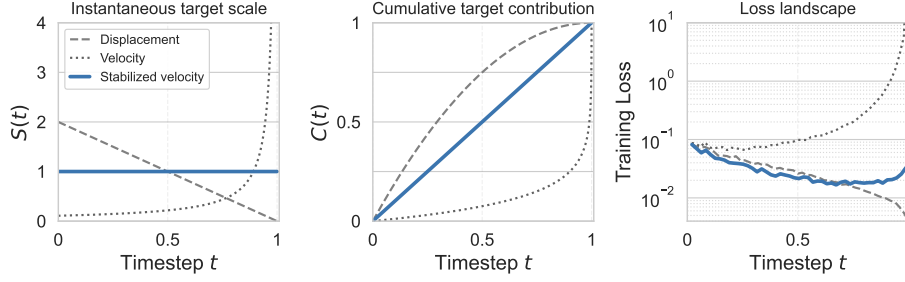


Fig. 2: Theoretical target contributions and empirical loss landscape. Left & Middle: Theoretical instantaneous target scale $S(t) = \mathbb{E}\|\tau_t\|^2$ and cumulative contribution $C(t) = \frac{\int_0^t S(s) ds}{\int_0^{0.999} S(s) ds}$ for $\tau_t \in \{d_t, u_t, \tilde{u}_t\}$. **Right:** Actual training loss across timesteps. While predicting vanilla velocity (u_t) or displacement (d_t) leads to loss explosion or vanishing as $t \rightarrow 1$, stabilized velocity ($\tilde{u}_t$) yields a time-uniform conditional target scale and prevents the training loss from vanishing or exploding.

$$\alpha(x_0, x_1, t)^2 = \frac{\mathbb{E}[\|u_t(x_t|x_1)\|^2]}{\|x_1 - x_0\|^2} \quad (13)$$

$$= 1 + \frac{tD}{(1-t)\|x_1 - x_0\|^2}, \quad (14)$$

where D denotes the latent dimensionality.⁴ Thus $\mathbb{E}_\epsilon\|\tilde{u}_t\|^2 = \|x_1 - x_0\|^2$, so the conditional target scale is independent of t . As shown in Fig. 2, this choice reduces divergence and balances target magnitudes.

The resulting stabilized velocity-matching objective is:

$$\mathcal{L}_{\text{velocity}} = \mathbb{E}_{t, \epsilon, x_0, x_1} \left[\|\tilde{v}_\theta(x_t, t) - \tilde{u}_t(x_t|x_1)\|^2 \right], \quad (15)$$

where $\tilde{v}_\theta(x_t, t) = v_\theta(x_t, t)/\alpha(x_0, x_1, t)$ normalizes the network prediction for loss calculation only, while the network still directly predicts velocity. Equivalently, this objective is a sample-wise weighted velocity-matching loss; Appendix C makes this explicit. The complete training procedure is summarized in Algorithm 1.

4.2 Variance-corrected sampling

To sample from the trained Brownian Bridge model with stabilized velocity matching, we discretize the continuous-time SDE defined in Eq. (1) using the Euler-Maruyama discretization [40]. Given a schedule $0 = t_0 < t_1 < \dots < t_N = 1$, the sampling process starts from the source x_0 and iteratively updates the latent state towards the target x_1 .

⁴ Derivations of factor $\alpha(\cdot)$ are in the Appendix C.

Algorithm 1: Training

Input: data pairs $(x_0, x_1) \sim p_{\text{source}, \text{target}}$, model v_θ , latent dimension D

- 1 **repeat**
- 2 Sample latent pair (x_0, x_1) , interpolation time $t \sim U(0, 1)$, and noise $\epsilon \sim \mathcal{N}(0, I)$;
- 3 Construct intermediate state $x_t = (1 - t)x_0 + tx_1 + \sqrt{t(1 - t)} \epsilon$;
- 4 Compute normalization factor $\alpha = \sqrt{1 + tD/[(1 - t)\|x_1 - x_0\|^2]}$;
- 5 Compute velocity target $\tilde{u}_t = (x_1 - x_t)/[\alpha(1 - t)]$;
- 6 Compute stabilized velocity loss $\mathcal{L}_{\text{velocity}} = \|\tilde{v}_\theta(x_t, t) - \tilde{u}_t\|^2$;
- 7 Update model parameters θ by gradient descent on $\mathcal{L}_{\text{velocity}}$;
- 8 **until** *convergence*;

The standard Euler–Maruyama discretization yields:

$$x_{k+1}^{\text{standard}} = x_k + \Delta t_k v_\theta(x_k, t_k) + \sqrt{\Delta t_k} \epsilon_k, \quad (16)$$

where $\Delta t_k = t_{k+1} - t_k$ and $\{\epsilon_k\}_{k=0}^{K-1}$ are i.i.d. samples drawn from $\mathcal{N}(0, I)$. This scheme assumes a locally constant variance structure, i.e., the stochastic term scales purely with $\sqrt{\Delta t_k}$.

However, in the Brownian Bridge process, the variance should gradually shrink as the trajectory approaches the target x_1 , reflecting decreasing uncertainty near $t = 1$. Consequently, the noise magnitude in the naive discretization becomes overly large at late steps, leading to biased trajectories and degraded sample quality.

To correct this mismatch, a scaling factor can be applied to continuously modulate the variance across timesteps. In practice, the diffusion term is rescaled by the ratio $\frac{1-t_{k+1}}{1-t_k}$ ⁵, resulting in a variance-corrected stochastic update [2, 31]:

$$x_{k+1}^{\text{corrected}} = x_k + \underbrace{\Delta t_k v_\theta(x_k, t_k)}_{\text{velocity toward target}} + \underbrace{\sqrt{\Delta t_k \frac{1-t_{k+1}}{1-t_k}} \epsilon_k}_{\text{variance-corrected noise}}. \quad (17)$$

This correction ensures that the variance decays smoothly as $t \rightarrow 1$, aligning the discrete sampling dynamics with the intrinsic structure of the Brownian Bridge. The complete inference procedure is summarized in Algorithm 2.

5 Experiments

We conduct comprehensive experiments to evaluate the effectiveness, efficiency, and scalability of our scaled Brownian Bridge formulation across various complex conditional generation tasks. Specifically, we begin with an apples-to-apples comparison in Section 5.1 against the standard Noise-to-Data diffusion paradigm,

⁵ Derivations of the scaling ratio are in the Appendix C.

Algorithm 2: Inference

Input: source-target latent pair (x_0, x_1) , trained model v_θ , latent dimension D , discretization steps N , discretization schedule $0 = t_0 < t_1 < \dots < t_N = 1$

- 1 Initialize $x \leftarrow x_0$;
- 2 **for** $k = 0, 1, \dots, N - 1$ **do**
- 3 Compute step size $\Delta t \leftarrow t_{k+1} - t_k$;
- 4 Compute scaling factor $\eta \leftarrow \sqrt{\Delta t \frac{1-t_{k+1}}{1-t_k}}$;
- 5 Sample noise $\epsilon \sim \mathcal{N}(0, I)$;
- 6 Update latent state:

$$x \leftarrow x + \Delta t v_\theta(x, t_k) + \eta \epsilon$$
- 7 **end**

Output: Final state x approximating the target x_1

establishing the advantages of the Data-to-Data approach in learning dynamics under a controlled setup. Next, we scale up to complex video conditional tasks in Section 5.2, exploring video stylization and depth-to-video translation, to demonstrate its strong capability in maintaining temporal consistency and structural fidelity. Returning to the instruction-based image editing task, Section 5.3 demonstrates that our proposed Bridge formulation can seamlessly scale to a 20B-parameter base model and achieve competitive performance purely through paradigm adaptation. Finally, we provide detailed ablation studies in Section 5.4 to analyze key properties like noise scales and validate our proposed stabilized velocity matching objective. Together, these evaluations show that ViBT achieves strong performance-efficiency trade-offs, establishing it as a capable and extensible foundation for complex visual conditional generation.

5.1 Apples-to-apples comparison

In this section, we first conduct an apples-to-apples comparison between the Noise-to-Data paradigm (Conditional DiT) and the Data-to-Data Bridge paradigm. Both approaches are initialized from the identical **SD3.5-Large** base model with no task-specific priors. We fine-tune them using the exact same image editing dataset, optimization strategy, and training iterations.

Observation on learning dynamics. Under this identical setup, we observe a clear difference in their learning dynamics. As shown in Figure 3, the Noise-to-Data approach learns to generate the entire target image from scratch. During the learning process, the whole image undergoes massive changes. In contrast, the Data-to-Data Bridge learns to edit. During its learning process, the regions that do not need to change remain unaffected. It only focuses on learning the specific areas that require modification. This allows it to rapidly learn the editing task while preserving the original image structure.

Quantitative results. The quantitative results are summarized in Table 1. Given identical training iterations, the Bridge approach achieves higher scores

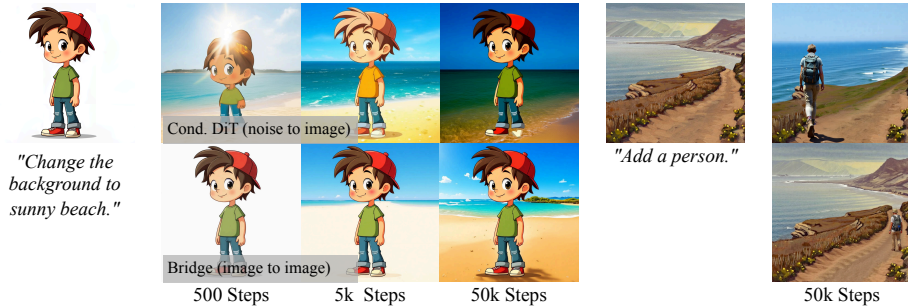


Fig. 3: Apples-to-Apples comparison on training dynamics. **Left:** Visualizing the learning progression of conditional DiT (noise-to-data) versus our Bridge paradigm (data-to-data) under identical training conditions. **Right:** An additional editing example. These results demonstrate that under the Bridge paradigm, the model not only rapidly learns the editing logic, but also effectively preserves the unedited regions.

Table 1: Apples-to-Apples quantitative comparison. The Data-to-Data Bridge paradigm improves performance and reduces training/inference costs compared to the Noise-to-Data Conditional DiT paradigm.

Method	ImgEdit Benchmark Score $\uparrow$										Human	Train	Inf.
	Add	Adjust	Extract	Replace	Remove	Bg.	Style	Hybrid	Action	Avg.	Pref. $\uparrow$	(GPU hr) $\downarrow$	(s) $\downarrow$
Cond. DiT. (noise-to-data)	3.26	2.67	1.72	2.19	2.03	2.37	3.17	1.29	2.07	2.43	40.7%	162	14.5
Bridge (data-to-data)	3.92	3.29	1.85	2.93	2.94	3.25	4.19	2.11	3.31	3.18	59.3%	96	6.2

than the Conditional DiT approach in this controlled setting. Specifically, the Bridge paradigm obtains a higher average score across all editing sub-tasks and is significantly preferred by human evaluators (59.3% vs. 40.7%). For efficiency, because the Bridge paradigm no longer requires the concatenation of extra conditional tokens, it reduces the overall computational overhead, requiring less training time to reach convergence and running faster during inference.

These comparisons demonstrate the potential of the Bridge paradigm. By shifting from noise reconstruction to data-to-data transport, it offers an effective and efficient foundation for complex visual editing tasks.

5.2 Complex video tasks

Having observed the advantage of the Data-to-Data Bridge paradigm over the Noise-to-Data approach in the controlled comparison, we now implement ViBT to explore a range of complex conditional generation tasks that have not been explored under the bridge paradigm, such as video stylization, depth-to-video translation, and instruction-based image editing. We initialize our models from the state-of-the-art video diffusion model Wan 2.1 1.3B.

Video stylization We first consider the instruction-based video stylization task, which aims to modify the visual style of an input video according to a

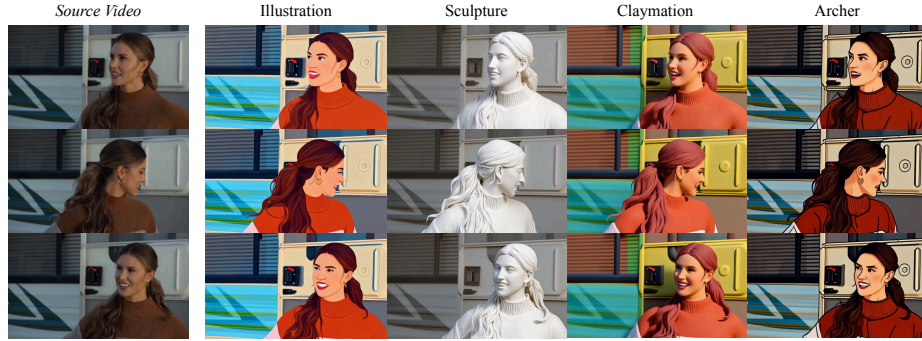


Fig. 4: Qualitative results of ViBT on the video stylization task with different styles.

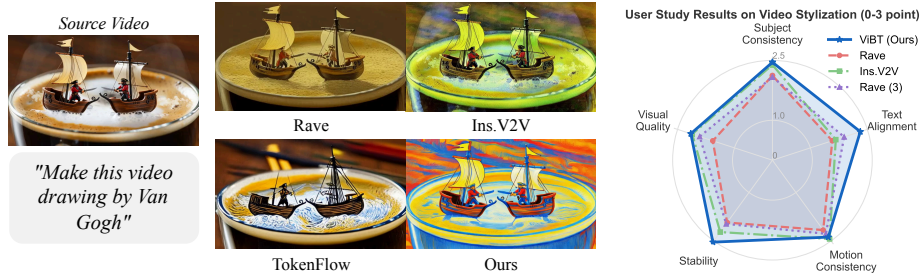


Fig. 5: **Left:** Qualitative comparison of stylized videos under the *Van Gogh* style. **Right:** User study results (26 participants, 5 dimensions, scored 1–3). Both qualitative results and human preference consistently favor ViBT.

Table 2: Quantitative comparison on video stylization. ViBT achieves the best overall performance across perceptual quality and aesthetic metrics.

Method	NIQE↓ [41]	TOPIQ _{NR} ↑ [7]	MUSIQ↑ [26]	MANIQA↑ [66]	CLIPQA↑ [59]	CLIP Score↑ [17]
TokenFlow	4.767	0.376	55.186	0.267	0.378	0.683
Ins.V2V	4.268	<u>0.467</u>	<u>60.621</u>	<u>0.306</u>	<u>0.440</u>	0.827
RAVE	6.514	0.351	50.595	0.269	0.377	0.683
ViBT	<u>4.328</u>	0.503	64.045	0.348	0.486	<u>0.782</u>

given textual instruction while preserving its original content and motion dynamics. We train our bridge model on 10,000 video samples from the *Ditto-1M* dataset [3]. For evaluation, we construct a benchmark of 100 videos generated by Wan 2.2 14B from *MovieGen Bench* [45] prompts, each paired with a random style instruction. We compare our ViBT model against several video stylization methods, including InsV2V [10], RAVE [23], and TokenFlow [48].

As shown in Table 2, ViBT achieves the best performance on most perceptual quality and aesthetic metrics, confirming its effectiveness in generating high-quality stylized videos. Beyond automatic metrics, we conduct a user study with 26 participants who rate each method across five dimensions on a scale of 1–3.



Fig. 6: Qualitative comparison on the *depth-to-video* task.

Table 3: Quantitative comparison on depth-to-video synthesis. ViBT achieves strong results across perceptual quality, ground-truth similarity, and VBench Score.

Method	Base Model	Perceptual quality						Ground truth similarity			CLIP Score $\uparrow$	VBench Score $\uparrow$
		NIQE $\downarrow$	TOPIQ $\uparrow$	MUSIQ $\uparrow$	MANIQA $\uparrow$	CLIPQA $\uparrow$		SSIM $\uparrow$	PSNR $\uparrow$	DISTS $\downarrow$		
ControlVideo	SD 1.5	6.641	0.443	50.735	0.354	0.436		0.385	9.067	0.465	0.732	0.62
Control A Video	SD 1.5	<u>5.102</u>	0.374	52.391	0.254	0.334		0.276	8.510	0.348	0.715	0.59
VideoComposer	SD 2.1	6.750	0.305	43.691	0.276	0.237		0.329	9.656	0.457	0.722	0.59
Wan Fun Control	Wan 2.1	5.346	0.477	<u>59.086</u>	<u>0.335</u>	<u>0.459</u>		<u>0.427</u>	<u>10.899</u>	<u>0.281</u>	<u>0.776</u>	0.69
ViBT	Wan 2.1	4.896	0.477	59.625	0.331	0.477		0.429	11.403	0.230	0.781	0.71

Table 4: VBench attribute breakdown on depth-to-video synthesis. ViBT achieves the highest average score and leads on the majority of individual attributes.

Method	Subj. Cons.	Bkgd. Cons.	Aesth. Qual.	Img. Qual.	Obj. Class	Multi. Objs.	Color	Spatial Rel.	Scene	Temp. Style	Overall Cons.	Human Action	Temp. Flicker	Motion Smooth.	Dyn. Degree	Appear. Style	Avg. Score
Control Video	0.899	0.94	0.54	0.52	0.57	0.26	0.706	0.46	0.29	0.20	0.24	0.80	0.991	0.990	0.11	0.229	0.55
Control A Video	0.791	0.88	0.48	<u>0.59</u>	0.59	0.25	0.799	0.44	0.43	0.21	0.24	0.83	0.982	0.976	0.72	<u>0.235</u>	0.59
Video Composer	0.873	0.92	0.44	0.48	0.67	0.23	0.854	0.32	0.29	0.22	0.24	0.91	0.963	0.949	0.88	0.222	0.59
Wan Fun	0.913	0.93	<u>0.60</u>	0.57	<u>0.87</u>	<u>0.65</u>	<u>0.848</u>	<u>0.70</u>	<u>0.46</u>	<u>0.24</u>	<u>0.26</u>	1.00	<u>0.989</u>	<u>0.978</u>	<u>0.86</u>	0.211	<u>0.69</u>
ViBT	<u>0.907</u>	<u>0.93</u>	0.63	0.63	0.91	0.71	0.835	0.74	0.54	0.25	0.27	1.00	0.990	0.976	0.82	0.221	0.71

As shown in Figure 5 (right), ViBT is consistently preferred over all baselines, aligning with the quantitative findings. Qualitative comparisons in Figure 5 (left) and Figure 4 further demonstrate that ViBT effectively applies diverse artistic styles while preserving the original content. More examples and user study details are provided in the Appendix.

Depth-to-video translation We further explore depth-to-video synthesis to verify the versatility of the bridge paradigm. We generate 1,003 training videos using Wan 2.2 14B [58] from *MovieGen Bench* [45] prompts, and extract corresponding depth maps via Depth Anything V2 [65] to form training pairs. For evaluation, we use 946 videos generated from VBench [20] prompts. We compare our ViBT against ControlVideo [74], Control-A-Video [9], VideoComposer [60], and Wan-Fun Control [1], which is based on the same base model as our ViBT.

Table 3 presents quantitative comparisons on perceptual quality, ground-truth similarity (SSIM [62], PSNR [6], DISTS [13]), and the overall VBench Score [20]. ViBT achieves the best or comparable results on most metrics, indicating strong video generation quality and reliable conditioning behavior. To further examine specific aspects, Table 4 reports fine-grained attribute evaluations under VBench, where ViBT achieves leading performance on most attributes. Figure 6 provides qualitative examples, showing that ViBT produces richer and more detailed visuals that align more closely with the depth conditions.

Table 5: Model ranking on *ImgEdit-Bench* based on average score.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Average
MagicBrush	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.83
Ins.Pix2Pix	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
Step1X-Edit	3.88	3.14	1.76	3.40	2.41	3.16	<u>4.63</u>	2.64	2.52	3.06
UniWorld-V1	3.82	3.64	2.27	3.47	<u>3.24</u>	2.99	4.21	2.96	2.74	3.26
FLUX Kontext	3.82	3.64	2.27	3.47	<u>3.24</u>	2.99	4.21	2.96	2.74	3.71
ViBT ⁶	<u>4.14</u>	<u>4.20</u>	2.64	<u>3.72</u>	3.03	<u>4.06</u>	4.87	<u>3.19</u>	<u>3.95</u>	<u>3.76</u>
Qwen-Image-Edit	4.17	4.29	<u>2.44</u>	4.30	3.90	4.15	4.00	3.32	4.51	3.90

**Fig. 7: Qualitative comparison on image editing.** ViBT achieves competitive visual quality compared to existing baselines.

To verify that these gains transfer beyond synthetic evaluation videos, we additionally evaluate ViBT on 500 real-world clips sampled from *OpenVid* [43]; the results are broadly consistent with Table 3 and are reported in Appendix A (Table S3 and Figure S2). Additional experimental results, including qualitative evaluations on extra video translation tasks such as video interpolation and video colorization, can be found in the Appendix A.

5.3 Instruction-based image editing

In this section, we transform a large-scale conditional DiT into the bridge paradigm. Specifically, we take Qwen-Image-Editing [63], a 20B-parameter image editing model, remove its conditional input channel, and convert it into a data-to-data bridge model via post-training. This paradigm transformation is lightweight: we post-train the 20B model with a rank-128 LoRA on 8 NVIDIA H100 GPUs for 20,000 iterations with a per-GPU batch size of 1, amounting to only 128 GPU-hours in total. We train on 5,000 synthetic editing pairs constructed from the *Open Images Dataset* [28] and evaluate on *ImgEdit-Bench* [68]. We compare against InstructPix2Pix [4], Step1X-Edit [36], FLUX.1 Kontext [29], and others [32, 69, 72].

Results and analysis Table 5 reports the quantitative results on *ImgEdit-Bench*⁷. After paradigm transformation from the base model (Qwen-Image-

⁶ The noise scale is set to $s = 0.5$; details are given in Section 5.4.

⁷ Results of baselines are reported from *ImgEdit Bench* [32, 33, 68].



Fig. 8: Qualitative results of ViBT on image editing. ViBT handles diverse editing tasks, including style transfer, object manipulation, and content modification, while maintaining the structural integrity of the original image.

Table 6: Comparison across different noise scales (s) on depth-to-video task.

Noise Scale (s)	SSIM $\uparrow$	PSNR $\uparrow$	NIQE $\downarrow$	DISTS $\downarrow$	CLIP Score $\uparrow$	VBench Score $\uparrow$
$s = 0$	0.347	9.808	5.432	0.3103	0.717	0.604
$s = 0.1$	0.331	9.206	5.413	0.3452	0.675	0.536
$s = 0.5$	0.398	10.227	5.185	0.2617	0.752	0.666
$s = 1$ (default)	0.429	11.403	4.896	0.2304	0.781	0.709
$s = 2$	0.396	11.305	4.499	0.2295	0.784	0.711
$s = 4$	0.394	10.146	5.912	0.3820	0.670	0.482

Table 7: Comparison across different noise scales (s) on image editing task.

Noise Scale (s)	Add	Adjust	Extract	Replace	Remove	Bg.	Style	Hybrid	Action	Avg.
$s = 0$	3.91	4.29	2.01	2.45	1.60	3.35	4.65	2.56	3.07	3.10
$s = 0.1$	3.43	4.00	2.04	2.31	1.61	3.53	4.46	2.58	3.29	3.03
$s = 0.5$	4.15	4.20	2.64	3.72	3.03	4.06	4.87	3.19	3.95	3.76
$s = 1$ (default)	4.20	3.70	2.31	3.86	2.91	3.92	4.85	2.72	3.52	3.55
$s = 2$	4.14	3.49	2.36	3.94	3.16	3.64	4.82	2.46	2.98	3.44
$s = 4$	3.70	2.67	2.24	3.60	2.88	2.93	4.43	1.78	2.50	2.97

Editing), ViBT retains competitive editing ability with only a 0.14 drop in average score, while achieving a $2.3\times$ inference speedup. Notably, ViBT surpasses the base model on several sub-tasks, particularly Style (4.87 vs. 4.00) and Extract (2.64 vs. 2.44), suggesting that the bridge paradigm is especially effective for structure-preserving edits. We also note that ViBT trails the base model on Remove and Action, indicating that edits requiring complete object deletion or large action changes still benefit from the original conditional input channel and remain challenging for the current bridge adaptation. The qualitative results in Figure 7 and 8 further confirm this: compared to the base model, ViBT significantly improves the preservation of unedited regions while maintaining faithful instruction-following. Detailed speed comparisons are provided in the Appendix A.

5.4 Ablation and analysis

Noise scale Several previous works [5, 31] further extend the Brownian Bridge formulation by modifying the diffusion term in Eq. (1). Instead of fixing the

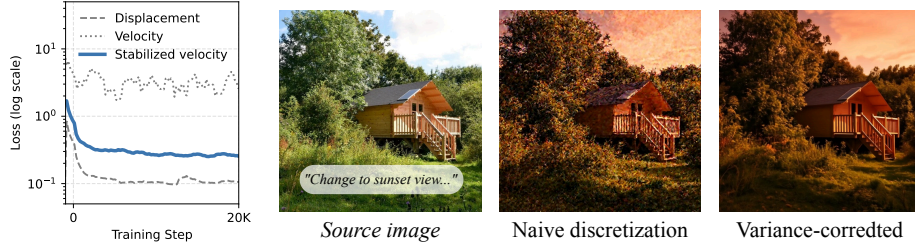


Fig. 9: Loss curves of different objectives. **Fig. 10:** Ablation study on variance-corrected sampling. Variance-corrected sampler avoids visible artifacts.

diffusion coefficient as a constant $\sigma(t) \equiv 1$, they introduce a global noise scale parameter s such that $\sigma(t) \equiv s$, leading to the generalized SDE:

$$dX_t = v_\theta(X_t, t) dt + s dW_t. \quad (18)$$

To investigate its impact, we conduct experiments across different values of s , summarizing results in Table 6 and Table 7. The corresponding training and inference modifications for this generalized formulation are detailed in the Appendix, Section C. Our findings indicate that moderate noise scales ($s = 1$ or $s = 2$) achieve better performance for the depth-to-video task, with $s = 2$ showing strong overall scores. For image editing tasks, a smaller noise scale ($s = 0.5$) surprisingly achieves the highest average performance, notably outperforming the default $s = 1$ setting. However, excessively small ($s < 0.5$) or large ($s > 2$) noise scales significantly degrade quality on both tasks. These observations highlight that optimal noise scales differ across tasks, contrasting with previous work [5] advocating an extremely small noise scale ($s = 0.005$).

Training objectives We compare three training objectives to validate our proposed stabilized velocity matching objective defined in Eq. (15), along with displacement matching in Eq. (11) and velocity matching in Eq. (2). Table 8 shows that stabilized velocity matching consistently achieves the best performance on both depth-to-video and image editing tasks. Specifically, it surpasses other objectives on all evaluated metrics for depth-to-video generation, and it also attains the highest average scores in diverse image editing scenarios. Moreover, Figure 9 highlights its superior training stability. We further verify that this advantage cannot be reproduced by simply reweighting the vanilla velocity loss across timesteps (*e.g.*, Min-SNR [16] or EDM-style [24] weighting), which act in the opposite direction to our stabilization; the detailed comparison is provided in Appendix A (Table S4).

Ablation study on variance-corrected sampling To validate the effectiveness of the variance-corrected sampling strategy described in Eq. (17), we perform an ablation study by comparing it with the standard Euler-Maruyama discretization method without variance correction. Figure 10 provides qualitative results for this comparison on the image editing task. We observe that the

Table 8: Quantitative comparison of different training objectives.

Training Objective	Depth-to-Video						Image Edit
	SSIM↑	PSNR↑	NIQE↓	DISTS↓	CLIP Score↑	VBench Score↑	ImgEdit Score
Displacement	0.409	11.04	4.91	0.26	0.772	0.695	3.50
Velocity	0.428	10.81	5.45	0.27	0.773	0.698	3.36
Stabilized velocity	0.429	11.40	4.90	0.23	0.78	0.71	3.55

naive discretization method (without variance correction) introduces noticeable artifacts, leading to degraded visual quality. In contrast, the variance-corrected sampling generates a cleaner and visually coherent image. To complement this qualitative evidence with a quantitative measurement, we further conduct a paired comparison on 200 image-editing samples, evaluating both samplers under identical settings with no-reference image-quality metrics; variance-corrected sampling improves quality consistently across all four metrics, confirming that the gains are measurable and not merely visual (see Appendix A, Table S5).

6 Conclusion, limitations, and future work

This paper introduced ViBT, which successfully scales the data-to-data Bridge paradigm to billion-scale parameters for complex vision generation tasks. By proposing a variance-stabilized velocity matching objective, we overcome the numerical instabilities of prior methods to enable stable training at scale. Extensive experiments validate that ViBT offers strong performance-efficiency trade-offs across diverse conditional editing and translation scenarios.

While ViBT demonstrates strong capabilities, its performance depends on the tuning of the noise scale parameter s . Moreover, the current 20B image-editing adaptation remains competitive but does not uniformly surpass its base model, with object removal and action edits remaining important failure cases. Future work will explore adaptive methodologies to automatically select this parameter, further enhancing the versatility of Bridge Models in varying tasks. Beyond the visual translation tasks studied here, the data-to-data formulation is naturally suited to modeling transitions between states, suggesting a promising extension of the Bridge paradigm toward world and action models [50] for embodied and interactive generation.

Acknowledgment

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00), the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002). It is also supported by NUS IT’s Research Computing group under grant number NUSREC-HPC-00001. The authors thank Ruonan Yu and Sicheng Feng for the helpful discussions.

References

1. aigc-apps: VideoX-Fun: A more flexible framework that can generate videos at any resolution and create videos from images. <https://github.com/aigc-apps/VideoX-Fun> (2025), videoX-Fun is an open-source video generation pipeline supporting AI image and video generation and training baseline and LoRA models.
2. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
3. Bai, Q., Wang, Q., Ouyang, H., Yu, Y., Wang, H., Wang, W., Cheng, K.L., Ma, S., Zeng, Y., Liu, Z., Xu, Y., Shen, Y., Chen, Q.: Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742 (2025)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
5. Chadebec, C., Tasar, O., Sreetharan, S., Aubin, B.: Lbm: Latent bridge matching for fast image-to-image translation. arXiv preprint arXiv:2503.07535 (2025)
6. Chen, C., Mo, J.: IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: <https://github.com/chaofengc/IQA-PyTorch> (2022)
7. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing **33**, 2404–2418 (2024)
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-*alpha*: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
9. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models (2023)
10. Cheng, J., Xiao, T., He, T.: Consistent video-to-video transfer using synthetic dataset. arXiv preprint arXiv:2311.00213 (2023)
11. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
12. De Bortoli, V., Thornton, J., Heng, J., Doucet, A.: Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in neural information processing systems **34**, 17695–17709 (2021)
13. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
16. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via Min-SNR weighting strategy. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 7407–7417. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00684>
17. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
19. Hu, Y., Liu, S., Tan, Z., Yang, X., Wang, X.: Image editing as programs with diffusion models. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2025)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
21. Ji, H., Lin, T.J., Li, H.: Dpbridge: Latent diffusion bridge for dense prediction. *ArXiv abs/2412.20506* (2024), <https://api.semanticscholar.org/CorpusID:275134211>
22. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598* (2025)
23. Kara, O., Kurtkaya, B., Yesiltepe, H., Reh, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024)
24. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) *Advances in Neural Information Processing Systems*. vol. 35, pp. 26565–26577. Curran Associates, Inc. (2022)
25. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4401–4410 (2019)
26. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5148–5157 (2021)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
28. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* **128**(7), 1956–1981 (2020)
29. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontekst: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742* (2025)
30. Lee, E., Jeong, S., Sohn, K.: Ebdom: exemplar-guided image translation with brownian-bridge diffusion models. In: *European Conference on Computer Vision*. pp. 306–323. Springer (2024)
31. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdom: Image-to-image translation with brownian bridge diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition* (2023)
32. Li, Z., Liu, Z., Zhang, Q., Lin, B., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888* (2025)
33. Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147* (2025)

34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
35. Liu, G.H., Vahdat, A., Huang, D.A., Theodorou, E.A., Nie, W., Anandkumar, A.: I²sb: Image-to-image Schrödinger bridge. arXiv preprint arXiv:2302.05872 (2023)
36. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
37. Liu, S., Yu, W., Tan, Z., Wang, X.: LinFusion: 1 GPU, 1 Minute, 16K Image. CoRR **abs/2409.02097** (2024). <https://doi.org/10.48550/ARXIV.2409.02097>
38. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
39. Liu, Y., Tan, Z., Hu, Y., Liu, S., Wang, X.: Gated condition injection without multimodal attention: Towards controllable linear-attention transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
40. Maruyama, G.: Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo* **4**(1), 48–90 (1955)
41. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
42. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024)
43. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025. OpenReview.net (2025), <https://openreview.net/forum?id=j7kdXSrISM>
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
45. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
46. Pu, Y., Gan, Z., Hénao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems* **29** (2016)
47. Qin, Z., Tan, Z., Wang, Z., Liu, S., Wang, X.: SpotEdit: Selective region editing in diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
48. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069 (2024)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
50. Shen, Q., Zhang, S., Liao, Y., Li, Q., Tan, Z., Wang, S., Yan, S., Wang, X.: World Action Models: A survey (2026), <https://arxiv.org/abs/2606.20781>
51. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

52. Song, Y., Liu, C., Shou, M.Z.: Omniconsistency: Learning style-agnostic consistency from paired stylization data. arXiv preprint arXiv:2505.18445 (2025)
53. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
54. Tan, Z., Xue, Q., Yang, X., Liu, S., Wang, X.: Ominicontrol2: Efficient conditioning for diffusion transformers. arXiv preprint arXiv:2503.08280 (2025)
55. Tan, Z., Yang, X., Liu, S., Wang, X.: Video-Infinity: Distributed long video generation. CoRR **abs/2406.16260** (2024). <https://doi.org/10.48550/ARXIV.2406.16260>
56. Team, M.L., Cai, X., Huang, Q., Kang, Z., Li, H., Liang, S., Ma, L., Ren, S., Wei, X., Xie, R., et al.: Longcat-video technical report. arXiv preprint arXiv:2510.22200 (2025)
57. Vasilev, V., Ivanov, A., Gushchin, N., Kovaleva, M., Korotin, A.: Time-correlated video bridge matching. ArXiv **abs/2510.12453** (2025), <https://api.semanticscholar.org/CorpusID:282064795>
58. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
59. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
60. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
61. Wang, Z., Zheng, B., Yang, X., Tan, Z., Xu, Y., Wang, X.: Minute-Long videos with dual parallelisms. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 10358–10366. AAAI Press (2026). <https://doi.org/10.1609/AAAI.V40I12.38006>
62. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
63. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
64. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
65. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024)
66. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqua: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1191–1200 (2022)
67. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
68. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)

69. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
70. Yu, R., Tan, Z., Chen, Z., Liu, S., Wang, X.: ViFeEdit: A video-free tuner of your video diffusion transformer. CoRR **abs/2603.15478** (2026). <https://doi.org/10.48550/ARXIV.2603.15478>
71. Yue, C., Peng, Z., Ma, J., Du, S., Wei, P., Zhang, D.: Image restoration through generalized ornstein-uhlenbeck bridge. arXiv preprint arXiv:2312.10299 (2023)
72. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems **36**, 31428–31449 (2023)
73. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
74. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
75. Zheng, K., He, G., Chen, J., Bao, F., Zhu, J.: Diffusion bridge implicit models. arXiv preprint arXiv:2405.15885 (2024)
76. Zheng, L., Zhang, Y., Guo, H., Pan, J., Tan, Z., Lu, J., Tang, C., An, B., Yan, S.: MEMO: memory-guided diffusion for expressive talking video generation. Trans. Mach. Learn. Res. **2026** (2026), <https://openreview.net/forum?id=uBcHcM7Kzi>
77. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. ArXiv **abs/2309.16948** (2023), <https://api.semanticscholar.org/CorpusID:263310775>