Learning Zero-Shot Subject-Driven Video Generation Using 1% Compute

Daneul Kim[1,3,§] Jingxu Zhang[3] Wonjoon Jin[2] Sunghyun Cho[2]
Qi Dai[3] Jaesik Park[1,†] Chong Luo[3,†]

[1]Seoul National University, Seoul, Republic of Korea
[2]POSTECH, Pohang, Republic of Korea
[3]Microsoft Research Asia, Beijing, China
{carpedkm,jaesik.park}@snu.ac.kr, chong.luo@microsoft.com

Abstract. Subject-driven video generation (SDV-Gen) aims to produce videos of a specific subject by adapting a pretrained video model, enabling personalized and application-driven content creation. To achieve this goal, per-subject tuning methods require approximately 200 A100 GPU hours to generate a customized video, whereas zero-shot methods avoid per-subject tuning but typically rely on millions of subject-video pairs for the supervision, incurring massive network fine-tuning costs (10K–200K A100 GPU hours). We propose a *data- and compute-efficient* zero-shot SDV-Gen framework that avoids test-time per-subject tuning and the use of large-scale subject-video pairs. Our key idea decomposes SDV-Gen into (i) identity injection learned from subject-image pairs and (ii) motion-awareness preservation maintained by a small set of arbitrary videos. We optimize the two tasks with stochastic switching, using random reference-frame sampling and image-token dropout to prevent trivial first-frame copying. Our gradient analysis shows that the two objectives rapidly evolve toward nearly orthogonal update subspaces, explaining the stable optimization. Using CogVideoX-5B, we adapt a single model with 200K subject-image pairs and 4,000 arbitrary videos in 288 A100 GPU hours. This yields about 1% of compute compared to prior zero-shot baselines (*i.e.*, 0.4% of VACE and 2.8% of Phantom) while using no subject-video pairs, yet remaining competitive in subject fidelity and motion quality. We show that the same recipe transfers to Wan 2.1-1.3B and Wan 2.2-5B.

Keywords: Zero-Shot · Video Customization · Video Personalization · Subject-Driven Video Generation

1 Introduction

Recent advances in video diffusion models [4, 23, 48, 56] have expanded controllable text-to-video (T2V) synthesis and video customization, enabling condition-

[§]Work done during internship at Microsoft Research Asia.
[†]Corresponding authors.

Fig. 1: Our results. We extend video generation models to enable *zero-shot* subject-driven video generation (SDV-Gen). We use small datasets and compute for the supervision. **(First row)** Results of extended CogVideoX-5B [56] supervised with 200K subject-image pairs and 4,000 arbitrary videos. **(Second row)** The same model extended with only 4,000 subject-image pairs and 4,000 videos. **(Third row)** Results of extended Wan 2.2-5B [48] with 200K subject-image pairs and 4,000 arbitrary videos.

ing on signals such as keypoints, edges, and reference images [1, 26, 33, 46, 57]. Building on this progress, *subject-driven video generation* [27, 29, 52, 53, 60, 63] (also known as subject-to-video customization or video personalization, *i.e.*, SDV-Gen) seeks to synthesize videos that preserve a target subject's identity while allowing variations in scene, motion, and context, an ability essential for customized content creation. In practice, this is typically achieved by adapting a pretrained T2V backbone to the target subject.

Early SDV-Gen approaches fine-tune a separate model or LoRA per subject [7, 43, 52, 55], limiting scalability. More recent methods train a single model that generalizes to unseen subjects [12,25,27,29,30,34,60], through training with large-scale subject-video pairs, but this implicitly assumes that both identity and motion-awareness must be learned from subject-video pairs. More importantly, these zero-shot methods rely on large subject-video datasets and substantial compute (Table 1; 10K–210K A100 GPU-hours for VACE [30] and Phantom [34]), raising the barrier to entry in this field. Reducing dependence on subject-video pairs and lowering computational cost are thus crucial for making subject-driven, customized video generation more accessible.

A potential alternative is to learn identity from subject-driven image generation (SDI-Gen) data (*i.e.*, subject-image pairs) while inheriting motion priors from the pretrained T2V backbone. In principle, subject-image pairs provide strong identity cues, whereas the backbone encodes motion priors (*i.e.*, temporal dynamics) from large-scale pretraining. However, naïve fine-tuning on subject-images pairs severely degrades inter-frame temporal coherence, often causing motion to collapse into inconsistent dynamics or near-static (frozen) movement.

Table 1: Comparison of subject-driven video generation methods. We compare per-subject tuning methods' required inputs and per-subject tuning time, along with zero-shot methods' dataset size and required finetuning time.

Per-subject Tuning	**Zero-shot**	**Required Inputs**	**Base Model (Param.)**	**A100 hours***
CustomCrafter [55]	✗	200 regularizing imgs per subject	VideoCrafter2 [9] (1.4B)	200‡
Still-Moving [7]	✗	Few reference images + 40 videos	Lumiere [3] (1.2B)	–
Zero-shot Methods	**Zero-shot**	**Dataset Size**	**Base Model (Param.)**	**A100 hours***
VideoBooth [29]	△	48,724 subject-video pairs	SD-based VDM [22, 42] (1.08B)	775–1,938‡
VACE [30]	✓	53M source videos	LTX [20] & Wan [48] (1.3–14B)	70K–210K‡
Phantom [34]	✓	1M subject-video pairs	Wan [48] (1.3–14B) + Seed [49]	10K–30K‡
Ours-*tiny/mini*	✓	**2,000**/**4,000** subject-image pairs + **4,000** arbitrary videos	CogVideoX [56] (5B)	**288**
Ours	✓	**200K** subject-image pairs + **4,000** arbitrary videos	CogVideoX [56] (5B) & Wan [48] (2.1-1.3B, 2.2-5B)	**288**

*Total GPU hours. ‡Estimated from reported batch sizes and wall-clock training time. Please see the supplement for details.

This suggests that identity and motion modeling interfere with each other, indicating that treating SDV-Gen as a single objective has a fundamental issue.

We consider subject adaptation into two complementary components: identity injection from subject-image pairs and motion-awareness preservation from a small set of arbitrary videos. Based on this, we propose to use a stochastic task-switching schedule that alternates between these objectives, predominantly sampling subject-image pairs while using arbitrary videos only to maintain temporal coherence. This separation allows identity and motion to be learned from different types of data sources, avoiding the collapse observed in naïve SDI-Gen fine-tuning and removing the need for subject-video pairs.

Across CogVideoX-5B [56], Wan 2.1-1.3B and Wan 2.2-5B [48], we observe that the conflict between the two gradients from the two-objectives diminishes rapidly and becomes nearly orthogonal, leading to zero-shot SDV-Gen result as in Figure 1. To understand why this two-objective training remains stable under task switching, we analyze the gradients of the two objectives during fine-tuning. The gradient analysis suggests that identity and motion updates concentrate on largely disjoint parameter subspaces, reducing interference.

This yields a data- and compute-efficient recipe to extend a T2V backbone into a zero-shot (*i.e.*, tuning-free at test time) SDV-Gen model without subject-video pairs. We fine-tune CogVideoX-5B on 200K subject-image pairs and 4,000 Pexels videos [31] for 4,000 steps (288 A100-hours), using ∼**1%** of the compute (2.8% of Phantom [34]; 0.4% of VACE [30]. See supplement Sec. **F** for the detailed cost computation.) while remaining competitive. By showing that SDV-Gen can be learned with substantially less computation than prior baselines, we outline a practical path toward scalable video customization. Our approach even applies in a strong low-data regime (4,000 subject-image pairs with 4,000 arbitrary videos) and transfers to Wan 2.1-1.3B and Wan 2.2-5B.

Our contributions are summarized as follows:

- We introduce a recipe to extend T2V models into a compelling zero-shot subject-driven video generation framework. We separate identity injection from motion preservation, enabling training without any subject-video pairs.
- We provide a gradient analysis showing that the identity and motion-aware objectives update independent parameter subspaces, explaining why the proposed formulation remains stable without subject-video pairs.
- We fine-tune multiple T2V backbones using only 200K subject-image pairs and 4,000 arbitrary videos, trained for 288 A100-hours, which is 1% of compute compared to other zero-shot baselines.

2 Related Work

2.1 Subject-driven Image Generation

Diffusion models have substantially advanced text-to-image synthesis [11,17,42], and extensive work studies how to inject novel subjects while preserving identity. Early approaches introduce explicit spatial control (*e.g.*, pose or edge guidance) [39, 62]. More recent methods employ cross-attention-based visual encoders such as IP-Adapter and SSR-Encoder [57, 64], as well as personalization techniques including DreamBooth [43], Textual Inversion [19], CustomDiffusion [32], DisenBooth [10], and subsequent multi-subject or zero-shot variants [5, 16, 35, 47, 51, 61]. These methods typically learn subject-specific representations from subject-image pairs, enabling subject-driven image generation (SDI-Gen). We follow this line by using subject-image pairs (Subject-200K from OminiControl [46]) as the primary source of identity supervision, and additionally leverage a small number of arbitrary videos to preserve temporal coherence.

2.2 Subject-driven Video Generation

Subject-driven video generation (SDV-Gen) extends SDI-Gen to the temporal domain. Per-subject tuning methods such as DreamVideo [52], MotionBooth [54], Still-Moving [7], and CustomCrafter [55] and DualReal [50] adapt a model (or LoRA) for each identity, incurring non-trivial test-time compute per subject. In contrast, zero-shot approaches including VideoBooth [29], Concept-Master [27], DreamVideo-2 [53], MagicMirror [63], ID-Animator [21], and Consis-ID [60] train a single model but often rely on subject-video pairs and may target restricted domains (*e.g.*, faces). Large-scale DiT-based systems such as VACE [30], Phantom [34], HunyuanCustom [25], and VideoAlchemist [12] further improve quality by training on million-scale datasets (*e.g.*, Phantom-Data [13] and OpenS2V-Nexus [59]), but require substantial compute budgets (Table 1). In contrast, we study a *data- and compute-efficient* alternative that avoids subject-video pairs during adaptation by decomposing identity learning and motion preservation, adapting pretrained backbones using subject-image pairs and a small set of arbitrary videos. We demonstrate efficient zero-shot SDV-Gen on modern video diffusion backbones, such as CogVideoX-5B [56] and Wan 2.2-5B [48].

2.3 Multi-task Optimization

Our approach trains a single model with two objectives: (1) identity injection on subject-image pairs and (2) motion preservation on arbitrary videos. This relates to multi-task learning, where gradient conflicts can hurt performance. Methods such as Gradient Surgery [58] mitigate interference by modifying updates. Related issues such as catastrophic forgetting [37] motivate continual learning approaches that revisit past samples via memory buffers [6, 36] or replay [18, 38, 41, 44]. In contrast, since we keep a fixed, small pool of videos, we proposed to adopt a lightweight stochastic task-switching schedule between subject-to-image and image-to-video updates. This design is computationally more efficient than Gradient Surgery and continual learning approaches because it updates a single objective per step, avoiding the simultaneous computation of multiple task gradients, explicit gradient projections, and memory expansion. We analyze training dynamics and show that gradient conflict diminishes rapidly. This supports stable optimization under limited data and compute.

3 Method

3.1 Preliminaries

We fine-tune a pretrained diffusion transformer (DiT) [40] for zero-shot subject-driven video generation. We instantiate our framework with CogVideo X-5B [56] and show that the same adaptation recipe transfers to Wan 2.1-1.3B and Wan 2.2-5B [48].

At each denoising step, the backbone processes noisy visual tokens $X \in \mathbb{R}^{N\times d}$ and text tokens $C_T \in \mathbb{R}^{M\times d}$ that share embedding dimension d. Each block applies LayerNorm followed by an attention operation that updates the visual tokens X under text conditioning from C_T, with spatial and temporal positions encoded by RoPE [45] on the visual tokens, $X_{i,j} \leftarrow X_{i,j} \cdot R(i,j)$. The text conditioning is realized as joint multi-modal attention over the concatenated sequence $[X; C_T]$ in CogVideoX and as cross-attention from X to C_T in the Wan family, and our adaptation is agnostic to this choice.

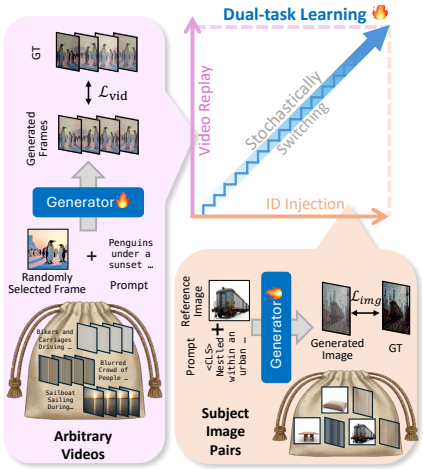

Fig. 2: Dual-task learning strategy. We formulate SDV-Gen as a *dual-task* problem: (i) identity injection (bottom) from subject-image pairs, and (ii) motion-awareness preservation (left) using arbitrary videos via stochastically switched learning.

Because attention cost scales quadratically with the number of tokens, video-centric fine-tuning is substantially more expensive than image-only training. Our

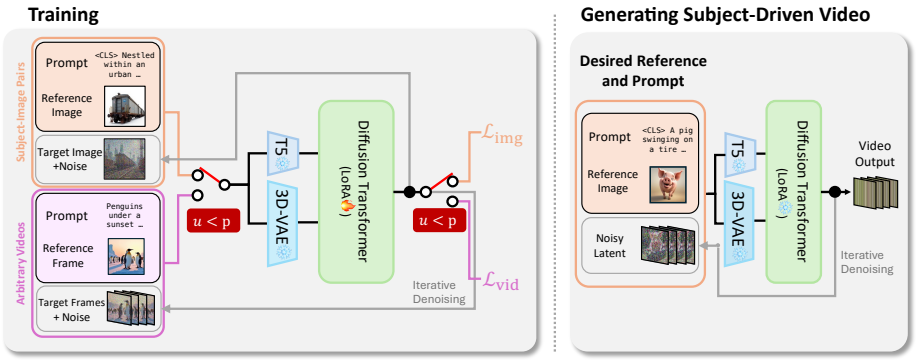

Fig. 3: Training and Inference Details. Left: During training, we stochastically alternate between two objectives: identity injection using subject-image pairs and motion-awareness preservation using a small set of arbitrary videos. **Right:** At inference time, *no additional per-subject tuning is required*. The model generates a video conditioned on the reference image and text prompt in a *zero-shot* manner.

goal is to adapt the backbone without any subject-video pairs by using subject-image pairs as the primary identity supervision and a small set of arbitrary videos to preserve temporal dynamics.

Let $\mathcal{D}_{\text{img}}$ denote a dataset of subject-image pairs $\{(I_{\text{ref}}, I_{\text{out}}), P\}$, where I_{ref} and I_{out} depict the same subject under different poses, viewpoints, or contexts, and P is the text prompt. Let $\mathcal{D}_{\text{vid}}$ denote a text-video dataset $\{(P_{\text{vid}}, V)\}$, where P_{vid} is the text prompt and V is the corresponding video. We define two training objectives, $\mathcal{L}_{\text{img}} = \mathcal{L}_{\text{subject-image}}(I_{\text{ref}}, I_{\text{out}}, P)$ for identity injection and $\mathcal{L}_{\text{vid}} = \mathcal{L}_{\text{text-video}}(P_{\text{vid}}, V)$ for motion-awareness preservation. The corresponding gradients are $g_{\text{img}} = \nabla_\theta \mathcal{L}_{\text{img}}$ and $g_{\text{vid}} = \nabla_\theta \mathcal{L}_{\text{vid}}$ with respect to the trainable parameters θ.

3.2 Dual-Task Formulation

We formulate a dual-task problem (Figure 2): (i) identity injection from subject-image pairs to learn subject appearance, and (ii) motion-awareness preservation from arbitrary videos. We optimize via stochastic task switching (Figure 3). At each iteration, we draw $u \sim \mathcal{U}(0, 1)$ and select:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{vid}}, & \text{if } u < p, \\ \mathcal{L}_{\text{img}}, & \text{otherwise.} \end{cases} \tag{1}$$

yielding expected update $\mathbb{E}[g] = (1 - p)\, g_{\text{img}} + p\, g_{\text{vid}}$. We use $p = 0.2$, so 80% of updates come from subject-image pairs and 20% from arbitrary videos during fine-tuning. p is chosen empirically. We demonstrate more in detail in supplementary material.

3.3 Identity Injection with Subject-Image Pairs

Following OminiControl [46], given subject-image pairs $(I_{\text{ref}}, I_{\text{out}})$ of the same subject and prompt P, we encode $X_{\text{ref}} = \text{VAE}(I_{\text{ref}})$, $X_{\text{out}} = \text{VAE}(I_{\text{out}})$, $C_T = \text{T5}(P)$, and train MM-DiT for X_{out} conditioned on (X_{ref}, C_T). For single-frame input, we truncate RoPE embeddings to match the token count of one frame.

We apply LoRA [24] to a subset of attention and normalization layers that interact with reference-image tokens, while keeping the rest of the backbone frozen. To anchor identity in the conditioning stream, we prepend a special `<CLS>` token to prompts (e.g., "`An <CLS> armchair in the living room`"). We provide ablations in the supplement showing that `<CLS>` improves identity metrics (CLIP-I, DINO-I) without degrading motion quality.

3.4 Motion Awareness with Arbitrary Videos

Fine-tuning a pretrained video model on SDI-Gen pairs alone often erodes temporal modeling and yields near-static generations at test time. To preserve motion, we introduce an image-to-video (I2V) objective on a small set of arbitrary videos $(P_{\text{vid}}, V) \in \mathcal{D}_{\text{vid}}$. For each video, we sample a reference frame index i and set $I_{\text{ref}} = V_i$ as conditioning, while the full video V serves as the reconstruction target. We encode

$$X_{\text{ref}} = \text{VAE}(I_{\text{ref}}), \quad X_{\text{vid}} = \text{VAE}(V), \quad C_T^{\text{vid}} = \text{T5}(P_{\text{vid}}), \tag{2}$$

and train MM-DiT for X_{vid} conditioned on $(X_{\text{ref}}, C_T^{\text{vid}})$. We adopt I2V fine-tuning (instead of pure T2V) since it matches the inference-time modality (reference image→video) and isolates motion preservation from identity supervision.

Mitigating reference-frame copying. Conditioning on a single frame can admit a degenerate solution that simply replicates the reference appearance across time with minimal motion. We propose to apply two simple regularizers to discourage this behavior: (i) *random reference-frame sampling*, where i is sampled uniformly over frames rather than fixed to the first frame; (ii) *image-token dropout*, where we randomly drop a subset of tokens in X_{ref} with probability p_{drop}, encouraging the model to rely on temporal priors instead of copying. We ablate these regularizers in the supplement.

3.5 Training Procedure

Putting everything together, fine-tuning proceeds as follows. At each training step we sample $u \sim \mathcal{U}(0, 1)$ and

1. if $u < p$, draw a mini-batch of arbitrary videos (T, V) from $\mathcal{D}_{\text{vid}}$, sample a random reference frame and apply image-token dropping, and update θ with $\mathcal{L}_{\text{vid}}(T, V)$;
2. otherwise, draw subject-image pairs $(I^{(1)}, I^{(2)}, P)$ from $\mathcal{D}_{\text{img}}$ and update θ with $\mathcal{L}_{\text{img}}(I^{(1)}, I^{(2)}, P)$.

This stochastic proxy replay maintains motion-awareness through occasional video updates while injecting identity from subject-image pairs in most steps.

SDI-Gen→I2V. An alternative way to generate subject-driven video generation is to sequentially apply SDI-Gen model and I2V model. However, this pipeline relies on clear subject visibility in the initial frame for identity propagation, which often fails in dynamic scenes (e.g., "a toy car comes closer to the camera on a wet pavement," Figure 4). When the subject is small or occluded, the image-to-video model may misidentify it and introduce inconsistent or fabricated details, causing a *progressive loss of identity* in later frames. We provide additional comparison with SDI-Gen→I2V as baselines. Please refer to the result in Sec. 5.

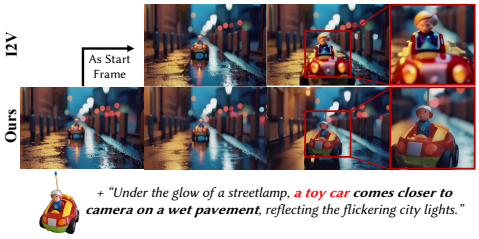

Fig. 4: Limitation of SDI-Gen→I2V. When the subject is small in the first frame, I2V fails to produce consistent results because it cannot reliably interpret low-resolution subjects.

Computational Efficiency. We quantify the computational efficiency of our stochastic task-switching scheme by analyzing the expected per-step cost. Let C_{img} and C_{vid} denote the per-step cost of SDI-Gen and I2V updates, respectively. Under MM-DiT attention, if a video contains T latent frames, the total token count grows approximately linearly with T, so the attention cost grows approximately as T^2. For T=13, we empirically observe $C_{\text{vid}} \approx 169\, C_{\text{img}}$. With task-switch probability p=0.2, the expected cost per step is

$$\mathbb{E}[C] = (1-p)C_{\text{img}} + pC_{\text{vid}} \approx 0.8C_{\text{img}} + 0.2 \cdot 169C_{\text{img}} \approx 34.6\, C_{\text{img}},$$

which is substantially lower than video-only fine-tuning. Using only 4,000 arbitrary videos, we fine-tune T2V models within 288 A100-hours while achieving competitive zero-shot SDV-Gen performance (Table 1).

4 Gradient Dynamics in Dual-Task Learning

In Sec. 3.2, we formulated SDV-Gen as a dual-task learning problem with two losses, $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{vid}}$, and their corresponding gradients $g_{\text{img}}(t)$ and $g_{\text{vid}}(t)$. To analyze gradient dynamics, we freeze the model at each checkpoint θ_t and compute $g_{\text{img}}(t)$ and $g_{\text{vid}}(t)$ on mini-batches from $\mathcal{D}_{\text{img}}$ and $\mathcal{D}_{\text{vid}}$ with respect to the trainable parameters θ_{train} (LoRA and task-specific normalization layers), matching the training setup. We log their cosine similarity $\phi(t)$ and the gradient norms $\|g_{\text{img}}(t)\|_2$ and $\|g_{\text{vid}}(t)\|_2$. Measurement protocols are provided in the supplement.

4.1 Do the Gradients Become Orthogonal?

Figure 5 plots the evolution of $\phi(t)$ under stochastic dual-task learning. We fine-tune pretrained CogVideoX-5B and Wan 2.2-5B using Subject-200K [46] as

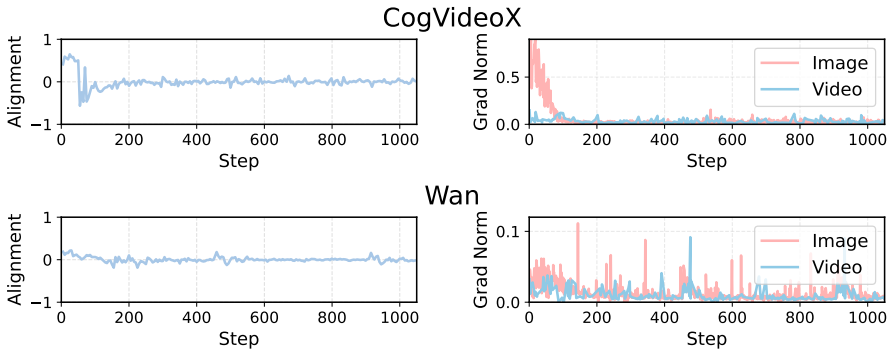

Fig. 5: Gradient analysis on alignment and norms during fine-tuning on CogVideoX-5B (Top) and Wan 2.2-5B (Bottom). (Left) Cosine similarity $\phi(t)$ between g_{img} and g_{vid} (over trainable parameters) quickly collapses to near zero under dual-task training, indicating emergent near-orthogonality. **(Right)** ℓ_2 norms $\|g_{\text{img}}(t)\|_2$ and $\|g_{\text{vid}}(t)\|_2$ remain non-negligible and similar in scale after 100 steps.

subject-image supervision and a small set of arbitrary videos from Pexels [31]. Starting from initialization, the cosine similarity rapidly converges to a narrow band around zero and remains there for the rest of fine-tuning regardless of the model used. This supports the hypothesis that, after a short transient, identity injection and motion-awareness preservation update nearly orthogonal directions in the trainable parameter subspace. (See supplement Sec. **B.1-2** for details and comparison and Sec. **B.3** for analysis on Wan 2.1-1.3B variant.)

To rule out the trivial explanation that $\phi(t)$ is close to zero only because both gradients vanish as training converges, we additionally track gradient norms. As shown in Figure 5, both $\|g_{\text{img}}(t)\|_2$ and $\|g_{\text{vid}}(t)\|_2$ remain well above the numerical noise floor throughout fine-tuning and stay within the same order of magnitude after the initial drop. The decay of $\phi(t)$ therefore reflects a genuine change in gradient *direction* rather than an artifact of vanishing magnitudes.

4.2 Why Does the Gradient Conflict Diminish?

The empirical behavior above suggests that stochastic switching between $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{vid}}$ tends to reduce their gradient inner product over training. To provide intuition for this behavior, we include a simplified analysis in supplement Sec. **C** based on a local second-order quadratic approximation. This analysis is intended only as an explanatory model, and should not be interpreted as a formal guarantee for DiT training, where the optimization landscape is highly non-convex.

Proposition 4.2.1 (Local decay of gradient inner product). *Under the local second-order assumptions, let*

$$\bar{\mathcal{L}}(\theta) = (1-p)\mathcal{L}_1(\theta) + p\mathcal{L}_2(\theta), \tag{3}$$

and consider gradient descent on the mixture loss. Then there exist constants $C > 0$ *and* $\rho \in (0, 1)$ *such that*

$$|\langle \nabla \mathcal{L}_1(\theta_t), \nabla \mathcal{L}_2(\theta_t) \rangle| \leq C\rho^t. \tag{4}$$

In particular,

$$\lim_{t \to \infty} \langle \nabla \mathcal{L}_1(\theta_t), \nabla \mathcal{L}_2(\theta_t) \rangle = 0. \tag{5}$$

4.3 Relation to Gradient-Surgery Method

Several works in multi-task learning explicitly manipulate gradients to resolve conflicts, such as Gradient Surgery [58], which projects each task gradient to remove components that conflict with other tasks. In our two-task setting, Gradient Surgery would replace g_{img} by $g_{\text{img}}^{\text{GS}} = g_{\text{img}} - \frac{\langle g_{\text{img}}, g_{\text{vid}} \rangle}{\|g_{\text{vid}}\|_2^2} g_{\text{vid}}$ if $\langle g_{\text{img}}, g_{\text{vid}} \rangle < 0$ and analogously for g_{vid}. This guarantees $\langle g_{\text{img}}^{\text{GS}}, g_{\text{vid}} \rangle \geq 0$ at each step, but requires computing *both* gradients and performing additional projections at every iteration, roughly doubling the gradient-computation cost compared to sampling a single task.

Our empirical analysis shows that the simpler stochastic switching scheme drives the gradient cosine $\phi(t)$ toward zero while keeping both norms non-vanishing, achieving a similar end effect (reducing gradient conflict) without explicit projections or extra gradient evaluations. For this reason, we use dual-task learning with stochastic switching as our default, and regard Gradient Surgery-style techniques as unnecessary refinements in this setting. We provide qualitative and quantitative comparisons between stochastic switching and Gradient Surgery in the supplement.

5 Experiments

5.1 Setup

Implementation. We build our approach on CogVideoX-5B, Wan 2.1-1.3B, and Wan 2.2-5B. CogVideoX-5B uses a 3D MM-DiT backbone [56] with a DDIM scheduler, whereas the Wan backbones use an X-DiT architecture [48] trained with flow matching. For all backbones, we optimize the two objectives with AdamW using a learning rate of 5×10^{-5} and a cosine-with-restarts schedule. For CogVideoX-5B, we fine-tune for 4,000 steps using BF16 mixed precision, with batch sizes of 256 for image updates and 32 for video updates. For Wan 2.1-1.3B and Wan 2.2-5B, we fine-tune each backbone for 5,000 steps, with batch sizes of 192 for image updates and 16 for video updates. The measured training cost is approximately 288 A100 GPU-hours for each backbone. We apply LoRA to the designated trainable layers, using rank 128 for CogVideoX and rank 32 for Wan, with dropout 0.2. Unless otherwise specified, we use stochastic task switching with probability $p = 0.2$, chosen empirically. We provide ablations on the LoRA rank and switching probability in the supplement with CogVideoX.

Table 2: VBench benchmark results. We use public Phantom [34] and VACE [30] results on Wan-2.1 1.3B [48].

SDI-Gen → I2V methods	**Motion Smooth.**	**Dynamic Degree**	**CLIP-T**	**CLIP-I**	**DINO-I**
OminiControl [46]	98.40	47.92	33.06	72.09	52.61
BLIP [33]	97.81	48.10	28.83	79.27	57.21
IP-Adapter [57]	97.74	51.90	28.94	76.60	51.41
SDV-Gen Methods	**Motion Smooth.**	**Dynamic Degree**	**CLIP-T**	**CLIP-I**	**DINO-I**
VideoBooth [29]	96.89	53.33	29.59	65.65	34.17
Phantom-1.3B [34]	98.70	67.08	33.50	73.00	53.55
VACE-1.3B [30]	98.74	43.75	33.70	73.54	54.47
MAGREF-480P [15]	98.86	65.09	33.00	71.15	48.50
Ours (CogVideoX-5B)	98.23	76.25	33.65	76.19	61.23
Ours (Wan 2.1-1.3B)	98.97	53.75	32.51	74.65	58.77
Ours (Wan 2.2-5B)	98.09	82.50	31.29	76.97	57.27

Colors: 1st, 2nd, 3rd.

Table 3: OpenS2V-Eval v1.1 benchmark results on Single-Domain. Total score is the normalized weighted sum of other scores.

Method	**Training Cost**	**Total↑**	**Aes↑**	**Smooth.↑**	**Amp.↑**	**FaceSim↑**	**Gme↑**	**Nexus↑**	**Natural↑**
Vidu 2.0	-	52.90	43.32	91.88	17.52	36.19	66.96	44.84	66.11
Pika 2.1	-	53.12	47.43	86.07	26.32	32.33	69.84	47.35	64.68
Kling 1.6	-	56.67	45.97	85.76	47.17	39.27	65.36	49.30	73.63
VACE-P1.3B [30]	≈70K hrs	49.20	48.93	95.68	11.91	18.04	70.78	36.24	66.85
VACE-1.3B [30]	≈70K hrs	51.13	49.41	95.42	22.51	22.37	70.87	38.34	68.33
VACE-14B [30]	≈210K hrs	**61.75**	48.94	93.16	19.69	64.65	65.86	50.82	70.56
Phantom-1.3B [34]	≈10K hrs	54.50	49.00	93.70	16.38	44.03	69.54	37.72	66.76
Phantom-14B [34]	≈30K hrs	**57.02**	47.46	94.86	41.55	51.82	70.07	35.30	71.11
SkyReels-A2-P14B [8]	-	55.06	40.85	85.54	26.41	54.42	61.81	48.60	61.85
MAGREF-480P [15]	-	53.44	46.31	92.63	27.43	33.77	69.02	42.45	68.33
Ours (CogVideoX-5B) [56]	288 hrs	50.05	45.40	93.90	19.38	18.05	70.53	41.23	68.52
Ours (Wan 2.1-1.3B) [48]	288 hrs	51.45	43.80	93.68	17.55	13.90	71.12	45.21	75.65
Ours (Wan 2.2-5B) [48]	288 hrs	**56.84**	44.70	84.97	14.38	48.78	61.98	49.97	71.30

Colors: 1st, 2nd, 3rd.

We use OminiControl's Subject200K dataset as our subject-image pairs. To study data efficiency, we additionally evaluate 2,000 and 4,000 subsets of Subject200K. We sample ∼4,000 arbitrary videos (1% of Pexels 400K [31]), providing diverse real-world motion patterns. We train two sampling-rate variants: an 8 FPS model aligned with the original CogVideoX setting, and a 16 FPS model further trained on Pexels videos to improve motion smoothness and temporal consistency at higher frame rates. All ablations are conducted on CogVideoX, with additional implementation details deferred to the supplementary material.

Baseline. For zero-shot baselines, we use the SDV-Gen models of VideoBooth [29], Phantom [34] and VACE [30] with Wan 2.1-1.3B, and MAGREF [15]. Also we additionally compare with state-of-the-art SDI-Gen models OminiControl [46], BLIP-Diffusion [33], and IP-Adapter [57], combined with the CogVideoX-5B I2V model: each first performs subject-driven image generation with its original setup, and we then generate videos via CogVideoX-5B. For per-subject tuning, we additionally compare with Still-Moving [7] and CustomCrafter [55] qualita-

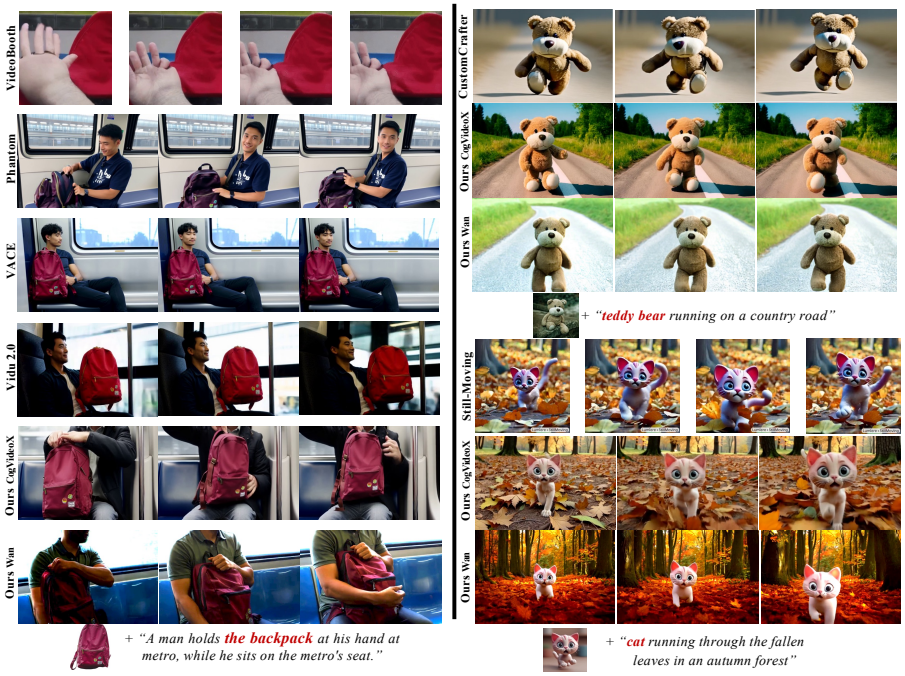

Fig. 6: Qualitative comparison with zero-shot methods (left) and per-subject tuning methods (right). Ours CogvideoX denotes our model fine-tuned on CogVideoX-5B, and Ours Wan denotes our model fine-tuned on Wan 2.2-5B. Note that ours is zero-shot, requiring no per-subject tuning at inference time.

tively, using the results reported in their paper due to unavailable code or costly per-subject tuning.

Evaluation Details. We randomly select 30 reference images from recent SDI-Gen benchmark sets [2, 7] and the standard DreamBooth dataset [43]. For each image, we use GPT to generate 8 prompts, resulting in 240 reference image–prompt pairs for video generation. Using the generated 240 videos, we report VBench metrics [28], including *Motion Smoothness* for temporal consistency, *Dynamic Degree* for the amount of motion, *CLIP-T* for text–video alignment, *CLIP-I* for CLIP-based image similarity, and *DINO-I* for DINO-based subject fidelity, and compare them with baseline methods. For ablation studies, we use a 120-video subset, corresponding to 4 prompts per image, as it shows similar trends. In addition, we report evaluation results on the OpenS2V benchmark version 1.1 [59].

5.2 Comparison with Baselines

Quantitative Analysis. Table 2 reports VBench metrics for zero-shot SDV-Gen, where Phantom and VACE use the publicly available Wan-2.1 1.3B ver-

sions. Our CogVideoX-5B and Wan 2.2-5B variants achieve strong motion dynamics, obtaining the second-highest and highest Dynamic Degree scores, respectively, while maintaining high Motion Smoothness. Our Wan 2.1-1.3B variant further achieves the best Motion Smoothness. Across CLIP-T, CLIP-I, and DINO-I, our variants remain competitive with Phantom and VACE, with Cog VideoX-5B achieving the best DINO-I score. Compared with SDI-Gen→I2V pipelines, which often obtain high CLIP-I but weaker motion dynamics, our method provides a stronger balance between subject fidelity and video motion.

Table 3 further evaluates OpenS2V-Eval v1.1 on the Single-Domain setting. Our Wan 2.2-5B variant achieves a competitive Total score of 56.84 with only 288 training hours, closely matching Phantom-14B (57.02) and outperforming commercial models such as Kling, Pika, and Vidu, as well as open-source models such as MAGREF and SkyReels. Although VACE-14B obtains the highest Total score, it requires substantially larger training cost. Our Wan 2.1-1.3B variant also remains competitive with 1.3B-scale baselines, surpassing VACE-1.3B in Total score despite using much less training compute. The relatively lower Amplification score of our Wan 2.2-5B variant is likely related to its TI2V backbone [14, 48]; in contrast, our Wan 2.1-1.3B model shows comparable motion scores to other 1.3B-family models such as VACE-1.3B and Phantom-1.3B. We additionally provide a *human preference study* in Supplement Sec. **E.1** for all CogVideoX and Wan-family variants.

Qualitative Analysis. As shown on the left of Figure 6, our method produces sharper subject details and more consistent identity than tuning-free baselines such as Vidu 2.0 and VideoBooth, while remaining competitive with recent SDV-Gen methods (Phantom and VACE). In the backpack example, VideoBooth largely fails to follow the prompt, collapsing to unrelated close-up frames. Vidu 2.0 preserves the backpack appearance but shows less stable motion across frames. In contrast, our CogVideoX and Wan variants better retain fine backpack textures and maintain coherent foreground motion, consistent with the VBench gains in Table 2.

We also compare against per-subject tuning methods CustomCrafter [55] and Still-Moving [7] using their official samples (Figure 6, right). For the teddy-bear sequence, CustomCrafter exhibits noticeable shape/texture drift across frames, whereas our results are more identity-faithful. For the cat sequence, Still-Moving tends to lose fine facial details (e.g., whiskers) and shows weaker detail preservation, while our results better maintain subject appearance under motion. We further find that *ours-mini*, trained with only 4,000 SDI-Gen pairs, remains competitive against these baselines. We discuss this further in the ablation below.

Table 4: Ablation study on training strategy. All variants are evaluated on the same 120-sample subset using VBench.

Method	Motion Smoothness	Dynamic Degree	CLIP-T	CLIP-I	DINO-I
Image-only	99.60	0.84	32.67	71.15	43.19
Two-stage	96.04	81.51	28.96	84.73	76.13
Ours	98.45	69.64	32.69	77.14	62.88

5.3 Ablation Study

Training Strategy. We compare three training strategies: (1) alternating optimization (Ours), (2) image-only, and (3) sequential two-stage (*i.e.*, image-only → image-to-video) fine-tuning approaches in the Table 4. Image-only fine-tuning produced high motion smoothness (99.60) because static videos (0.84 dynamic degree) lead to smooth motion, indicating a need for improved dynamics. Two-stage fine-tuning improved dynamics (81.51) and similarity (CLIP-I: 84.73) but introduced severe artifacts and identity forgetting. (See supplement Sec. **D.8**.) Our *dual-task learning* excelled, achieving a balance between motion smoothness (98.45) and dynamic degree (69.64). For additional training strategies comparison with gradient surgery [58] or with penalty, see supplement Sec. **B.4** and Sec. **D.9**.

Fig. 7: Ablation on the number of subject-image pairs. Using only 2,000 pairs (Ours-*tiny*) leads to weaker motion dynamics and more frequent failure cases.

Effect of Scaling the Image Dataset. While our method remains robust even with only 4,000 subject-image pairs (Ours *mini*), scaling the dataset to the full 200K pairs consistently improves fine-grained subject fidelity and temporal realism. With fewer subject-image pairs, we occasionally observe a *copy-paste* artifact: the generated subject preserves its identity, but appears overly rigid, often repeating a near-identical appearance or layout across frames.

Table 5: Ablation study on subject-image training-pair scale. All variants are evaluated on the same 120-sample subset. We report VBench metrics, Flow Entropy (FE), and copy score (CS).

Method	# Pairs	Motion Smooth.	Dynamic Degree	CLIP-T	CLIP-I	DINO-I	FE↑	CS↓
Ours-*tiny*	2,000	98.72	60.71	33.38	74.72	57.03	54.73	45.27
Ours-*mini*	4,000	97.99	**78.33**	**33.22**	75.87	58.86	58.91	41.09
Ours	200K	98.45	69.64	32.69	**77.14**	**62.88**	**63.77**	**36.32**

For example, in the *pig surfing* case in Figure 7, the full-set model better captures viewpoint and pose variations, including head turns and body articulation, rather than reusing a static subject template. This trend is also reflected in Table 5. As the number of subject-image pairs increases, CLIP-I and DINO-I improve, indicating stronger identity preservation. At the same time, flow entropy (FE), which measures the diversity of subject motion, increases, while copy score (CS), which captures near-rigid repetition of the reference subject, decreases. These results suggest that a small number of subject-image pairs is sufficient to learn the basic identity mapping, whereas broader subject-image coverage helps preserve high-frequency details and reduces copy-paste artifacts under complex motion. See the supplement Sec. **D.10** for implementation and measurement details on the FE and CS metrics.

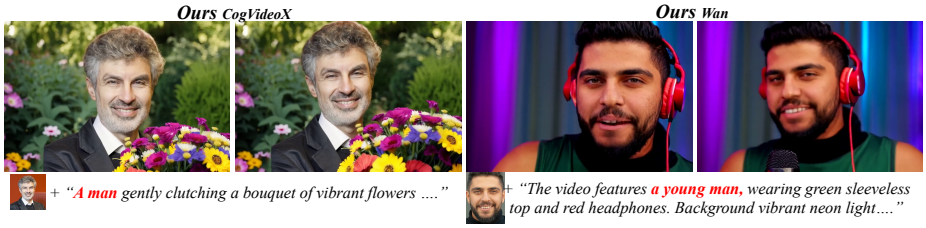

Fig. 8: Human-driven Video Generation. *Left:* CogVideoX. *Right:* Wan. Without training on human-centric datasets, our model still preserves facial identity well.

5.4 Limitations

Our primary fine-tuning dataset, Subject-200K [46], is largely object-centric and includes very few human faces. Despite this limited face coverage, our model generalizes reasonably well to facial subjects, preserving recognizable identity (Figure 8). Especially with Ours with Wan 2.2-5B, FaceSim is comparable to *commercial* models on OpenS2V (Table 3). However, failures can still occur under strong stylization/extreme appearance changes.

Regarding the Motion score in Table 3, we verify that our extended CogVideoX model preserves the motion magnitude of the vanilla model, indicating that motion variation is not degraded even after our recipe is applied (details in the supplement). However, it is lower than state-of-the-art subject-driven methods, suggesting room for improvement.

6 Conclusion

We recast subject-driven video generation as a dual-task problem that combines identity injection from subject-image pairs with motion-awareness preservation from a small set of arbitrary videos. Using a simple stochastic task-switching scheme, we train with 200K subject-image pairs and 4,000 arbitrary videos in 288 A100-hours, achieving competitive subject fidelity and motion quality while generalizing to unseen subjects in a zero-shot manner. Across backbones, gradient measurements show that the image and video objectives rapidly become nearly orthogonal, enabling stable task switching without explicit gradient surgery and delivering strong identity fidelity, motion quality, and text alignment compared to both per-subject-tuned and zero-shot SDV-Gen baselines.

Acknowledgement. This work was supported by IITP grant (RS-2021-II211343: AI Grad. School Program (SNU) (5%), RS-2019-II191906: AI Grad. School Program (POSTECH) (5%), RS-2026-25511821: Development of Personalized Video Automatic Generation and Insertion Technology (20%), RS-2024-00436680: Global Research Support Program in the Digital Field (40%)) and NRF (RS-2024-00405857 (30%)) funded by the Korea government (MSIT). We appreciate generous support from Microsoft Research Asia.

References

1. Atzmon, Y., Gal, R., Tewel, Y., Kasten, Y., Chechik, G.: Multi-shot character consistency for text-to-video generation (2024)
2. Avrahami, O., Hertz, A., Vinker, Y., Arar, M., Fruchter, S., Fried, O., Cohen-Or, D., Lischinski, D.: The chosen one: Consistent characters in text-to-image diffusion models. In: ACM SIGGRAPH 2024 conference papers. pp. 1–12 (2024)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Chan, K.C., Zhao, Y., Jia, X., Yang, M.H., Wang, H.: Improving subject-driven image synthesis with subject-agnostic guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6733–6742 (2024)
6. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
7. Chefer, H., Zada, S., Paiss, R., Ephrat, A., Tov, O., Rubinstein, M., Wolf, L., Dekel, T., Michaeli, T., Mosseri, I.: Still-moving: Customized video generation without customized video data. ACM Transactions on Graphics (TOG) **43**(6), 1–11 (2024)
8. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
9. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7310–7320 (2024)
10. Chen, H., Zhang, Y., Wu, S., Wang, X., Duan, X., Zhou, Y., Zhu, W.: Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In: International conference on learning representations. vol. 2024, pp. 23104–23126 (2024)
11. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
12. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6099–6110 (2025)
13. Chen, Z., Li, B., Ma, T., Liu, L., Liu, M., Zhang, Y., Li, G., Li, X., Zhou, S., He, Q., et al.: Phantom-data: Towards a general subject-consistent video generation dataset. arXiv preprint arXiv:2506.18851 (2025)
14. Choi, J.S., Lee, K., Yu, S., Choi, Y., Shin, J., Lee, K.: Improving motion in image-to-video models via adaptive low-pass guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40364–40375 (2026)
15. Deng, Y., Yin, Y., Guo, X., Wang, Y., Fang, J.Z., Yuan, S., Yang, Y., Wang, A., Liu, B., Huang, H., et al.: Magref: Masked guidance for any-reference video generation with subject disentanglement. arXiv preprint arXiv:2505.23742 (2025)

16. Ding, G., Zhao, C., Wang, W., Yang, Z., Liu, Z., Chen, H., Shen, C.: Freecustom: Tuning-free customized image generation for multi-concept composition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9089–9098 (2024)
17. 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)
18. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
19. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
20. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
21. He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., Zhang, J.: Id-animator: Zero-shot identity-preserving human video generation. arXiv preprint arXiv:2404.15275 (2024)
22. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
23. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
25. Hu, T., Yu, Z., Zhou, Z., Liang, S., Zhou, Y., Lin, Q., Lu, Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation. arXiv preprint arXiv:2505.04512 (2025)
26. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. arXiv preprint arXiv:2307.14073 (2023)
27. Huang, Y., Yuan, Z., Liu, Q., Wang, Q., Wang, X., Zhang, R., Wan, P., Zhang, D., Gai, K.: Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. arXiv preprint arXiv:2501.04698 (2025)
28. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
29. Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: Videobooth: Diffusion-based video generation with image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6689–6700 (2024)
30. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
31. jovianzm: Pexels-400k. `https://huggingface.co/datasets/jovianzm/Pexels-400k` (2025), accessed: 2026-03-05
32. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023)

33. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36**, 30146–30166 (2023)
34. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14951–14961 (2025)
35. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Customizable image synthesis with multiple subjects. Advances in neural information processing systems **36**, 57500–57519 (2023)
36. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
37. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
38. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. nature **518**(7540), 529–533 (2015)
39. 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)
40. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
41. Robins, A.: Catastrophic forgetting, rehearsal and pseudorehearsal. Connection Science **7**(2), 123–146 (1995)
42. 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)
43. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023)
44. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. Advances in neural information processing systems **30** (2017)
45. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
46. 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)
47. Tewel, Y., Kaduri, O., Gal, R., Kasten, Y., Wolf, L., Chechik, G., Atzmon, Y.: Training-free consistent text-to-image generation. ACM Transactions on Graphics (TOG) **43**(4), 1–18 (2024)
48. 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)
49. Wang, J., Lin, Z., Wei, M., Zhao, Y., Yang, C., Loy, C.C., Jiang, L.: Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2161–2172 (2025)

50. Wang, W., Huang, M., Tu, Y., Mao, Z.: Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16565–16575 (2025)
51. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. arXiv preprint arXiv:2406.07209 (2024)
52. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6537–6549 (2024)
53. Wei, Y., Zhang, S., Yuan, H., Wang, X., Qiu, H., Zhao, R., Feng, Y., Liu, F., Huang, Z., Ye, J., et al.: Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. arXiv preprint arXiv:2410.13830 (2024)
54. Wu, J., Li, X., Zeng, Y., Zhang, J., Zhou, Q., Li, Y., Tong, Y., Chen, K.: Motionbooth: Motion-aware customized text-to-video generation. Advances in Neural Information Processing Systems **37**, 34322–34348 (2024)
55. Wu, T., Zhang, Y., Wang, X., Zhou, X., Zheng, G., Qi, Z., Shan, Y., Li, X.: Customcrafter: Customized video generation with preserving motion and concept composition abilities. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8469–8477 (2025)
56. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations. vol. 2025, pp. 83048–83077 (2025)
57. 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)
58. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)
59. Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Luo, J., Yuan, L.: Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. arXiv preprint arXiv:2505.20292 (2025)
60. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12978–12988 (2025)
61. Zeng, Y., Patel, V.M., Wang, H., Huang, X., Wang, T.C., Liu, M.Y., Balaji, Y.: Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6786–6795 (2024)
62. 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)
63. Zhang, Y., Liu, Y., Xia, B., Peng, B., Yan, Z., Lo, E., Jia, J.: Magicmirror: Id-preserved video generation in video diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14464–14474 (2025)
64. Zhang, Y., Song, Y., Liu, J., Wang, R., Yu, J., Tang, H., Li, H., Tang, X., Hu, Y., Pan, H., et al.: Ssr-encoder: Encoding selective subject representation for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8069–8078 (2024)