# *DomainShuttle*: TOWARD GENERALIZABLE CROSS-DOMAIN SUBJECT-DRIVEN VIDEO GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Subject-driven text-to-video (S2V) generation has attracted growing interest in academia and industry. An ideal model should faithfully preserve reference appearance for in-domain generation while retaining intrinsic subject features under flexible cross-domain transformations. This generalizability should extend across generation paradigms, from bidirectional fixed-length to streaming autoregressive generation. Existing methods primarily optimize in-domain fidelity by indiscriminately preserving reference appearance, limiting cross-domain flexibility in real-to-fantasy, fantasy-to-real, and real-fantasy interaction scenarios. To address this limitation, we propose *DomainShuttle*, a novel framework for flexible and high-fidelity video personalization. Specifically, Domain-MoT decouples video and reference processing, and applies Domain-aware AdaLN for domain-specific modeling of reference images. Spatiotemporal Reference Anchor RoPE (STRA-RoPE) temporally anchors reference image tokens and spatially organizes reference tokens with subject-aware offsets, supporting both generation paradigms. For bidirectional training, Cross-Pair Consistent Loss aligns velocity predictions across reference sets to reduce interference from subject-irrelevant features. To further demonstrate its generalizability across generation paradigms, we extend the bidirectional model to streaming autoregressive generation through chunk-wise causal training. Extensive experiments demonstrate that *DomainShuttle* achieves strong video quality, text controllability, and subject consistency in both in-domain and cross-domain scenarios and on both bidirectional and streaming autoregressive architectures. The anonymous project page with videos is available at `https://DomainShuttle.github.io/`.

## 1 INTRODUCTION

Subject-driven video generation (S2V) aims to generate videos from textual instructions while preserving subjects specified by reference images, with broad applications in advertising, creative design, and AI filmmaking. Real-world applications require S2V models to generalize along two complementary dimensions. First, across visual domains, a model should faithfully reproduce a subject when the target video shares the visual domain of the reference images (in-domain), while preserving the subject's intrinsic identity when adapting the subject to a different visual domain in the generated video (cross-domain). Second, S2V models should preserve subject features in fixed-length clips under the bidirectional generation paradigm and generalize this capability to long videos under the streaming autoregressive generation paradigm. Together, these requirements motivate a unified framework for S2V generalization across both domain settings and generation paradigms.

Existing S2V studies primarily optimize subject fidelity in in-domain scenarios, *i.e.*, preserving as many subject features as possible without altering their attributes or style. Early works focus on single-subject identity preservation (He et al., 2024; Yuan et al., 2025b; Li et al., 2025b;a), while recent methods extend to multiple subjects through reference-feature injection (Liu et al., 2025; Jiang et al., 2025), image-to-video priors (Chen et al., 2026b; Deng et al., 2026; Chen et al., 2026a), multimodal understanding (Li et al., 2026b; Hu et al., 2025; Deng et al., 2025b), or representation alignment (Wang et al., 2026a). Overall, existing methods focus mainly on in-domain, fixed-length generation with bidirectional generation paradigms.



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




Figure 1: *DomainShuttle* supports in-domain and cross-domain scenarios. Beyond bidirectional generation, it can be applied to streaming autoregressive generation for consistent long videos.

However, existing methods pay limited attention to diverse cross-domain scenarios. Cross-domain scenarios include rendering real-world subjects in fantasy domains, mapping fantasy subjects to real-world objects, and generating complex interactions between real-world and fantasy subjects, as illustrated in Fig. 1. Indiscriminately preserving reference appearance entangles intrinsic subject features with subject-irrelevant properties, making it difficult to balance subject consistency with flexible adaptation. Moreover, reference-conditioning designs developed for bidirectional generation do not directly address stable subject preservation in streaming autoregressive generation. These limitations constrain S2V generalization across visual domains and generation paradigms.

We propose that an ideal S2V method should freely **shuttle** across both visual domains and generation paradigms, as illustrated in Fig. 1. Achieving this generalization requires jointly optimizing subject fidelity and flexibility in various scenarios. Specifically, reference conditioning should preserve recognizable subject features (*e.g.*, facial structure and unique markings) that remain compatible with the requested domain, while allowing appearance attributes (*e.g.*, style and material) to adapt to prompts. This principle should hold under both bidirectional and autoregressive modeling.

To this end, we propose ***DomainShuttle***, a novel S2V framework jointly optimizing subject consistency and generation flexibility across both domains and generation paradigms. Its core idea is decoupling video and reference processing to explicitly structure reference conditioning across both bidirectional and streaming autoregressive architectures. *DomainShuttle* includes three key components. (1) **Domain-MoT** (Mixture-of-Transformers) processes video and reference features through independent branches, and introduces domain-aware AdaLN in the reference branch to explicitly model target domain attributes, which enables cross-domain transformations with fine-grained text guidance. (2) **Spatiotemporal Reference Anchor RoPE** (STRA-RoPE) temporally anchors reference tokens at a fixed origin and spatially organizes them with subject-aware offsets. This structured reference coordinate system is shared by bidirectional and streaming autoregressive generation. (3) **Cross-Pair Consistent Loss** aligns velocity predictions under different reference sets at the same noise level for bidirectional generation, reducing sensitivity to subject-irrelevant reference features such as lighting, composition, or viewpoint. For streaming generation, we extend *DomainShuttle* to chunk-wise causal generation while retaining Domain-MoT and STRA-RoPE. Comprehensive evaluations show competitive video quality, text alignment, and subject consistency of our model in both in-domain and cross-domain settings, as well as robust subject preservation in streaming generation.

Our main contributions are summarized as follows:

- We propose *DomainShuttle*, a unified S2V framework that jointly supports subject consistency and generative flexibility, improving S2V generalization across different visual domains and generation paradigms.
- We introduce Domain-MoT with Domain-aware AdaLN to explicitly model target domain attributes through decoupled video and reference processing, alongside STRA-RoPE to temporally anchor and spatially organize reference tokens. For bidirectional training, we additionally introduce Cross-Pair Consistent Loss to reduce interference from subject-irrelevant features.
- We further extend *DomainShuttle* to streaming autoregression through chunk-wise causal generation with our reference modeling. Extensive experiments show our strong subject consistency and text controllability across both domain settings and generation paradigms.

## 2 Related Work

**Video Foundation Models**. Video diffusion models have substantially advanced video creation. Early approaches (Ho et al., 2022; Blattmann et al., 2023; Guo et al., 2024) augment UNet-based image diffusion models with temporal modules. DiT (Peebles & Xie, 2023) enables further scaling, supporting models such as CogVideoX (Yang et al., 2024), HunyuanVideo (Kong et al., 2024), Seedance (Team Seedance et al., 2025), and the Wan series (Wan et al., 2025). These advances enable diverse applications, including audio-visual generation (Yang et al., 2025) and reference video-based editing (Bai et al., 2025a; Wu et al., 2025c; Zhang et al., 2025). Beyond bidirectional fixed-length synthesis, recent methods explore streaming autoregressive generation. Self Forcing (Huang et al., 2025a) reduces the train-test gap through autoregressive self-rollout during training. Causal Forcing (Zhu et al., 2026) uses an autoregressive teacher for ODE initialization to address the architectural mismatch in distillation. Helios (Yuan et al., 2026) combines drift-aware training, context compression, and few-step sampling for efficient long-video generation. Together, these advances extend video generation from fixed-length synthesis to long-term streaming generation.

**Subject-driven Video Generation**. Subject-driven text-to-video generation synthesizes videos under textual guidance while preserving subjects specified by reference images. Early studies (Yuan et al., 2025b; Li et al., 2025b;a) focus on single-subject generation, while recent works extend to multi-subject personalization involving humans, objects, and backgrounds (Fei et al., 2025; Huang et al., 2025b; Chen et al., 2025; 2026b). Phantom (Liu et al., 2025) introduces a dynamic reference-feature injection scheme for multiple subjects. Other methods exploit the subject-preserving priors of I2V models (Deng et al., 2026; Chen et al., 2026a) or incorporate multimodal large language model features to enhance subject and instruction understanding (Li et al., 2026b; Hu et al., 2025; Deng et al., 2025b; Cai et al., 2026). RefAlign (Wang et al., 2026a) aligns reference features with the semantic space of a visual foundation model. MV-S2V (Song et al., 2026) uses temporally shifted RoPE to distinguish subjects and reference views. These methods primarily emphasize in-domain subject fidelity within bidirectional generation, leaving flexible cross-domain transformation and subject preservation in streaming generation less explored.

## 3 Methodology

*DomainShuttle* advances flexible and high-fidelity personalization through three modules: Domain-MoT, Spatiotemporal Reference Anchor RoPE, and Cross-Pair Consistent Loss, as shown in Fig. 2. During training, **Domain-MoT** processes video and reference image features through separate branches, using Domain-aware AdaLN in the reference branch to explicitly model target domain attributes. We then introduce **Spatiotemporal Reference Anchor RoPE** (STRA-RoPE), which temporally anchors references at a fixed temporal index and spatially processes them with subject-aware offsets. **Cross-Pair Consistent Loss** (CCL) aligns velocity predictions on two sets of reference images to reduce the interference of subject-irrelevant features. In this section, we first introduce preliminaries and present these modules in detail. Then we describe the data pipeline.

### 3.1 Preliminaries

Our method is trained on the DiT-based model (Wan et al., 2025). During training, the text encoder $\boldsymbol{\mathcal{E}}_t(\cdot)$ (Chung et al., 2023) encodes the text into text features $\boldsymbol{c}_t$, and the 3D VAE encoder $\boldsymbol{\mathcal{E}}_v(\cdot)$

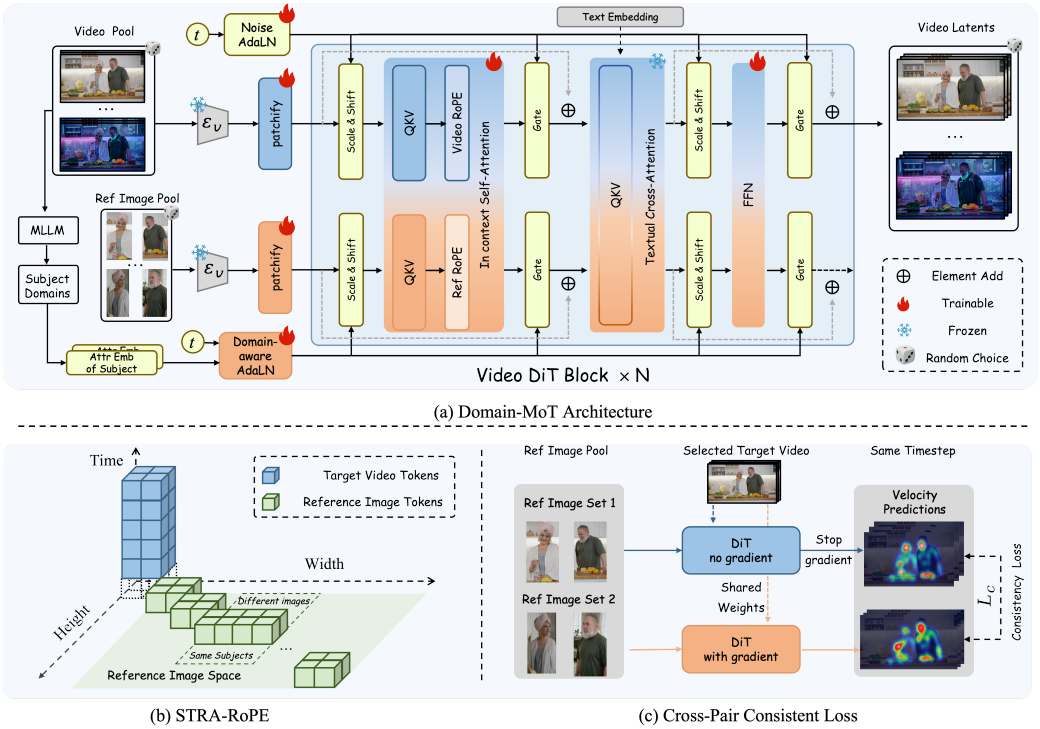



Figure 2: Overview of *DomainShuttle*. (a) The reference images and videos are injected into the decoupled branches of Domain-MoT, facilitating domain-aware AdaLN guidance in the reference branch. (b) STRA-RoPE temporally anchors reference image tokens at a fixed temporal index and spatially processes them with subject-aware offsets. (c) Cross-Pair Consistent Loss aligns the velocity predictions on two reference image sets, reducing the influence of subject-irrelevant features.

encodes the video and $N$ reference images $\mathcal{R} = \{I_0, I_1, \cdots, I_{N-1}\}$ into video latents $\boldsymbol{z}_1$ and reference features $\boldsymbol{c}_r$. The model is optimized by flow-matching loss (Lipman et al., 2023):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\boldsymbol{z}_0,\boldsymbol{z}_1} \| G_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_t, \boldsymbol{c}_r) - (\boldsymbol{z}_1 - \boldsymbol{z}_0) \|_2^2, \tag{1}$$

where $t \in [0, 1]$ denotes continuous time. $\boldsymbol{z}_0$ is a sample drawn from the prior distribution, and $\boldsymbol{z}_1$ denotes video latent encoded by 3D VAE. $G_\theta$ represents a learnable vector field parameterized by $\theta$.

## 3.2 Model Architecture

### 3.2.1 Domain-MoT

A key issue in cross-domain S2V is the entanglement between intrinsic subject features and domain-specific attributes, which hinders simultaneous subject preservation and flexible domain transformation. To address this issue, we propose Domain-MoT, as shown in Fig. 2 (a). Specifically, after 3D VAE encoding, the video latents and reference image features are separately patchified into $\boldsymbol{f}_v$ and $\boldsymbol{f}_r$. Domain-MoT processes them through independent pathways and injects domain attributes into the reference pathway via Domain-aware AdaLN, enabling explicit domain modeling.

Domain-MoT employs in-context self-attention with independent QKV projections and RoPE for the video and reference image features. This design preserves the pretrained capability of the video pathway while allowing the reference pathway to specialize in subject feature extraction. Formally,

$$\text{Softmax}\left(\frac{[R_r(\boldsymbol{Q}_r); R_v(\boldsymbol{Q}_v)] \cdot [R_r(\boldsymbol{K}_r); R_v(\boldsymbol{K}_v)]^\top}{\sqrt{d}}\right)[\boldsymbol{V}_r; \boldsymbol{V}_v], \tag{2}$$

where $R_v$ and $R_r$ denote the RoPE applied to the video branch and reference image branch, respectively, which will be discussed in the next subsection. The queries $\boldsymbol{Q}_v = \boldsymbol{W}_v^q \cdot \boldsymbol{f}_v$ and $\boldsymbol{Q}_r = \boldsymbol{W}_r^q \cdot \boldsymbol{f}_r$,

keys $\boldsymbol{K}_v = \boldsymbol{W}_v^k \cdot \boldsymbol{f}_v$ and $\boldsymbol{K}_r = \boldsymbol{W}_r^k \cdot \boldsymbol{f}_r$, values $\boldsymbol{V}_v = \boldsymbol{W}_v^v \cdot \boldsymbol{f}_v$ and $\boldsymbol{V}_r = \boldsymbol{W}_r^v \cdot f_r$. $\boldsymbol{W}^q$, $\boldsymbol{W}^k$ and $\boldsymbol{W}^v$ are learnable weight parameters. $[;]$ denotes feature concatenation. Self-attention and cross-attention masks are used only for streaming generation, as detailed in Appendix A. Then, for video chunk $s$, we apply cross-attention for reference and video features $\boldsymbol{f}_c^{(s)} = [\boldsymbol{f}_r; \boldsymbol{f}_v^{(s)}]$. The text features are combined as $\boldsymbol{f}_t^{(s)} = [\boldsymbol{f}_{t,B}; \boldsymbol{f}_{t,C}^{(s)}]$, where $\boldsymbol{f}_{t,B}$ encodes the global prompt and $\boldsymbol{f}_{t,C}^{(s)}$ encodes the global prompt concatenated with chunk-specific prompt $s$. Cross-attention is defined as:

$$\text{Softmax}\left(M_{\text{cross}}^{(s)} + \frac{\boldsymbol{Q}_c^{(s)}(\boldsymbol{K}_t^{(s)})^\top}{\sqrt{d}}\right)\boldsymbol{V}_t^{(s)}, \tag{3}$$

where $\boldsymbol{Q}_c^{(s)} = \boldsymbol{W}^q\boldsymbol{f}_c^{(s)}$, $\boldsymbol{K}_t^{(s)} = \boldsymbol{W}^k\boldsymbol{f}_t^{(s)}$, and $\boldsymbol{V}_t^{(s)} = \boldsymbol{W}^v\boldsymbol{f}_t^{(s)}$. $M_{\text{cross}}^{(s)}$ is 0 for permitted latents and $-\infty$ otherwise. For bidirectional generation, the chunk index $s$ is omitted and a single text condition controls all video frames.

**Domain-aware AdaLN.** Existing S2V methods typically modulate reference tokens and video latents indiscriminately, entangling reference-domain attributes with the generated video. Domain-aware AdaLN instead decouples their modulation. The reference AdaLN is conditioned on both the target domain attribute and time features, whereas the video-latent AdaLN uses only time features. Here, each subject's target domain attribute specifies its coarse target type in the generated video, rather than its domain in the reference images. Fine-grained styles are controlled by prompts through cross-attention. We use MLLM (Bai et al., 2025c) to annotate all subject attributes for both training and evaluation. This conditioning provides target-type priors for reference modeling, supporting in-domain subject preservation and cross-domain adaptation together with text guidance. The AdaLN mechanisms for video latent and reference image features are as follows:

$$\begin{cases}\hat{\boldsymbol{f}}_v \;=\; g_v(t) \;\odot\; \big[\mathcal{F}\big(\text{LN}(\boldsymbol{f}_v) \;\odot\; (1+\gamma_v(t)) \;+\; \beta_v(t)\big)\big] + \boldsymbol{f}_v, \\ \hat{\boldsymbol{f}}_{r,m} \;=\; g_r(t,a_m) \;\odot\; \big[\mathcal{F}\big(\text{LN}(\boldsymbol{f}_{r,m}) \;\odot\; (1+\gamma_r(t,a_m)) \;+\; \beta_r(t,a_m)\big)\big] + \boldsymbol{f}_{r,m},\end{cases} \tag{4}$$

where $t$ denotes time features and $a_m \in \{A_1, \ldots, A_K\}$ is domain attribute for subject $m$. LN is layer normalization and $\odot$ is Hadamard product. $\hat{\boldsymbol{f}}_v$ and $\hat{\boldsymbol{f}}_{r,m}$ denote the modulated video features and reference features, respectively. The reference modulation factors $\gamma_r$, $\beta_r$, and $g_r$ depend on $(t, a_m)$, while their video counterparts depend only on $t$. $\mathcal{F}(\cdot)$ is a residual function (*e.g.*, attention).

Domain-MoT is independent of the causal attention mask. Under causal attention, each video chunk directly accesses the reference pathway, which is modulated by the target domain attribute via Domain-aware AdaLN, rather than relying solely on historical chunks to propagate subject features. This provides stable subject and domain conditioning through autoregressive generation.

### 3.2.2 Spatiotemporal Reference Anchor RoPE

Video DiTs (Wan et al., 2025; Yang et al., 2024) commonly use 3D Rotary Positional Encoding (RoPE) to assign the video token a coordinate $(i, j, k)$ along the temporal, height and width axes. Existing S2V methods typically treat reference images as additional video frames and distinguish them only by temporal indices. This ignores both the absence of temporal continuity among reference subjects and the association between multiple images of the same subject.

We therefore propose **Spatiotemporal Reference Anchor RoPE** (STRA-RoPE), which can be applied to both bidirectional and streaming autoregressive generation, as shown in Fig. 2 (b). Temporally, all reference tokens are anchored at index 0, while video indices start from 1. In bidirectional generation, this anchor separates the references from the jointly denoised video frames. In streaming autoregression, it remains fixed as an stable anchor and provides each chunk with the same subject references. Spatially, subject-aware offsets separate different subjects while grouping images of the same subject. The video RoPE $R_v(i, j, k)$ and reference RoPE $R_r(j, k)$ are defined as follows:

$$\begin{cases}R_v(i,j,k) = \text{RoPE}_{3D}(i+1, j, k), \\ R_r(j,k) = \text{RoPE}_{3D}(0, j + h \times (m+1), k + w \times (n+1)),\end{cases} \tag{5}$$

where $h$ and $w$ denote the height and width of the reference latent grid, with spatial position $j \in \{0, \ldots, h-1\}$ and $k \in \{0, \ldots, w-1\}$. $m$ indexes the $m$-th reference subject, and $n$ indexes the $n$-th reference image across all subjects, both starting from 0. Height offsets distinguish subjects, while width offsets distinguish reference images.

#### 3.2.3 Cross-Pair Consistent Loss

To further enhance the flexibility of our model in various complex cross-domain scenarios, we introduce **Cross-pair Consistent Loss** (CCL) to reduce interference from subject-irrelevant features. We construct a pool of multiple reference image sets for each video (see Sec. 3.3 for dataset details). During training, two sets of reference images are randomly sampled, and the corresponding velocity predictions are aligned using the same latent and timestep as shown in Fig. 2 (c). Because the sets differ in viewpoint, occlusion, and illumination, matching a trainable branch to a stop-gradient branch encourages the model to capture shared subject features (*e.g.*, identity and texture) while suppressing frame-specific perturbations. Aligning predictions across reference sets at the same noise level reduces sensitivity to irrelevant subject features, supporting subject consistency and cross-domain flexibility. The loss $\mathcal{L}_{\mathrm{C}}$ is defined as:

$$\mathcal{L}_{\mathrm{C}} = \|G_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_t, \boldsymbol{c}_r) - G^*_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_t, \boldsymbol{c}^*_r)\|_2^2, \tag{6}$$

where $\boldsymbol{c}_r$ and $\boldsymbol{c}^*_r$ denote two different sets of reference image features. $G^*_\theta$ and $G_\theta$ share the same weights, with gradients stopped in $G^*_\theta$. The total loss is $\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{C}}\mathcal{L}_{\mathrm{C}}$, with $\lambda_{\mathrm{C}} = 0.1$.

### 3.3 Training Data Pipeline

We use three video personalization datasets: Phantom-Data (Chen et al., 2026c), OpenS2V (Yuan et al., 2025a), and Ditto-1M (Bai et al., 2025a). Some image personalization datasets are used for initialization. We use Grounding-DINO (Liu et al., 2024) for object detection and SAM2 (Ravi et al., 2024) for multi-frame segmentation to build reference image sets for CCL. Ditto-1M is a video editing dataset in which reference images can align with multiple edited and source videos. Only 20K pairs come from Ditto-1M as cross-domain data augmentation, including 10K reference-image/source-video pairs and **10K** reference-image/edited-video pairs. We ablate the use of Ditto-1M in Tab. 13 of Appendix C.7. In total, we collect 720K video samples covering humans, objects, fantasy subjects, and backgrounds. The dataset supports both "multiple reference sets $\rightarrow$ one video" and "one reference set $\rightarrow$ multiple videos" cross-pair. More dataset details are shown in Appendix B.

## 4 Experiments

### 4.1 Experimental Setups

**Implementation.** **(1) Bidirectional training.** We train bidirectional variants on Wan2.1-14B-T2V (Wan et al., 2025) and Wan2.2-14B-T2V. After initial fine-tuning on image-personalization data, we train for 10k steps on 720K video-personalization samples (batch size 64), freezing cross-attention to preserve text following. The Domain-MoT reference branch is initialized from the video branch, and the CCL weight is set to 0.1. We use $K = 4$ target domain attributes: real-world human, real-world object, background, and fantasy subject. Both bidirectional training stages use Adam with a learning rate of $1e-5$. Training costs are 14,000 and 28,000 A100 GPU-hours for Wan2.1 and Wan2.2, respectively. **(2) Streaming autoregressive training.** Following prior streaming methods (Yuan et al., 2026; Zhu et al., 2026), we initialize from the trained bidirectional Wan2.1 variant and continue training for 5,000 steps using Adam with a learning rate of $5e-6$. Training uses a fixed 77-frame window with four latent frames per chunk. For streaming generation with STRA-RoPE, the temporal indices of reference tokens are fixed at 0, while those of video latents span 1–20. During streaming generation, reference tokens and video chunks are processed by the reference and video branches of Domain-MoT, respectively. More details are shown in Appendix C.1.

**Test Dataset.** We construct a test set consisting of 120 in-domain samples and 120 cross-domain samples. In the test set, 120 in-domain samples are all from the open source OpenS2V-Eval (Yuan et al., 2025a) data, while the 120 cross-domain samples are self-constructed. The regular in-domain samples include multi-human, multi-object, human–object interactions, and background preserva-

Table 1: **Quantitative results**. The best scores are shown in **bold**, and the second-best are underlined. *DomainShuttle* achieves competitive performance across various metrics.

| Method | Video Quality | | Text Controllability | Cross-Domain | | | | | In-Domain | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AES↑ | MS↑ | GME↑ | CD-CLIP↑ | Style Score↑ | Style Consist↑ | CD-Score↑ | RA↑ | DINO-I↑ | CLIP-I↑ |
| Kling 1.6 (Kling, 2025) | 0.513 | 0.970 | 0.576 | 0.624 / 0.643 | 0.661 | 0.939 | 0.731 / 0.758 | 0.736 | 0.327 | 0.670 |
| MAGREF (Deng et al., 2026) | 0.484 | 0.961 | 0.677 | 0.613 / 0.635 | 0.684 | 0.945 | 0.669 / 0.722 | 0.634 | 0.346 | 0.673 |
| SkyReels-V3 (Li et al., 2026a) | 0.480 | 0.916 | 0.659 | 0.589 / 0.623 | 0.681 | 0.902 | 0.663 / 0.688 | 0.711 | 0.348 | 0.668 |
| Phantom (Liu et al., 2025) | 0.514 | 0.974 | 0.665 | 0.622 / 0.642 | 0.716 | 0.945 | 0.665 / 0.714 | 0.587 | 0.336 | 0.678 |
| HuMo (Chen et al., 2026b) | 0.469 | 0.976 | 0.665 | 0.604 / 0.633 | 0.708 | 0.943 | 0.636 / 0.685 | 0.487 | 0.337 | 0.662 |
| BindWeave (Li et al., 2026b) | 0.445 | 0.958 | 0.599 | 0.598 / 0.609 | 0.706 | 0.919 | 0.578 / 0.650 | 0.494 | 0.344 | 0.665 |
| RefAlign (Wang et al., 2026a) | 0.500 | 0.984 | 0.673 | 0.623 / 0.642 | 0.720 | 0.949 | 0.726 / 0.730 | 0.582 | 0.365 | 0.682 |
| FFGO-Wan2.2-14B (Chen et al., 2026a) | 0.406 | 0.945 | 0.655 | 0.588 / 0.605 | 0.691 | 0.801 | 0.632 / 0.674 | 0.705 | 0.303 | 0.646 |
| VACE-Wan2.2-14B (Jiang et al., 2025) | 0.472 | 0.973 | 0.682 | 0.605 / 0.629 | 0.729 | 0.933 | 0.674 / 0.713 | 0.672 | 0.339 | 0.670 |
| Ours (Wan2.1-14B) | 0.512 | 0.981 | 0.691 | 0.629 / 0.648 | 0.845 | 0.955 | 0.794 / 0.785 | 0.731 | **0.397** | **0.694** |
| Ours (Wan2.2-14B) | **0.518** | **0.988** | **0.706** | **0.639 / 0.661** | **0.883** | **0.965** | **0.854 / 0.837** | **0.827** | 0.381 | 0.681 |

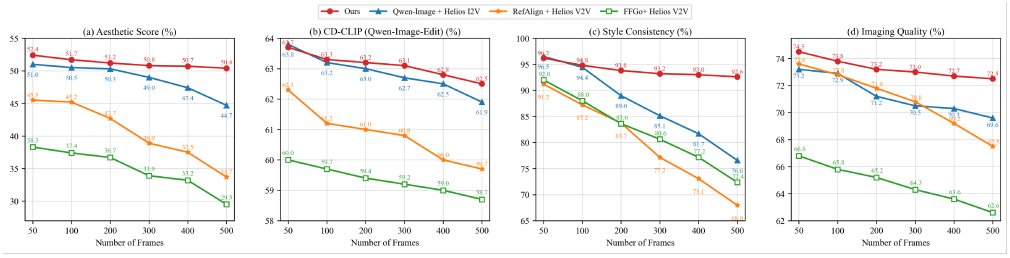



Figure 3: **Quantitative results** at different frames. *DomainShuttle* shows the least degradation, showing effectiveness for streaming autoregressive architectures.

tion. The cross-domain samples include 40 real-to-fantasy transformations, 40 fantasy-to-real transformations, and 40 complex interaction cases between real subjects and fantasy characters.

**Evaluation Metrics.** We evaluate effectiveness from three aspects. (1) Video Quality: The Aesthetic Score (AES) and Motion Smoothness (MS) (Yuan et al., 2025a) are used to evaluate video quality. Autoregressive evaluation additionally reports Imaging Quality (Huang et al., 2024). (2) Text Controllability: GMEScore (GME) (Zhang et al., 2024) is used to evaluate controllability. (3) Subject Consistency: For the In-Domain scenario, we adopt the standard DINO-I (Oquab et al., 2023) and CLIP-I (Radford et al., 2021) metrics. We segment the subjects and compute their similarity.

For **cross-domain**, we use five metrics. (1) **CD-CLIP** measures subject CLIP similarity to edited references generated separately by Nano Banana Pro (Google, 2025) and Qwen-Image-Edit (Wu et al., 2025a), yielding two scores under the same evaluation protocol. (2) **Style Score** uses Appearance Style metric of VBench (Huang et al., 2024) to assess style alignment. (3) **Style Consistency** measures CSD similarity (Somepalli et al., 2024) between subsequent frames and the first frame. (4) **CD-Score** evaluates overall cross-domain performance, including subject preservation, target domain fidelity, and subject interaction, using GPT-5.2 (OpenAI, 2025) and Qwen3-VL-8B-Instruct (Bai et al., 2025b) for two scores with the same rubric. (5) **Relation Alignment (RA)** uses VQAScore (Lin et al., 2024) to quantify confidence in the fulfillment of fantasy-to-real transformations and real-fantasy interactions. Tab. 1 reports both model variants of CD-CLIP and CD-Score separately. Appendices C.2 and C.3 show the evaluation details and human preference evaluation, respectively.

**Baselines.** For the **bidirectional** setting, we compare with three groups of baselines: (1) closed-source Kling1.6 (Kling, 2025); (2) Wan2.1-14B-based models: MAGREF (Deng et al., 2026), SkyReels-V3 (Li et al., 2026a), Phantom (Liu et al., 2025), HuMo (Chen et al., 2026b), BindWeave (Li et al., 2026b), and RefAlign (Wang et al., 2026a); (3) Wan2.2-14B-based models: FFGO (Chen et al., 2026a) and VACE-Wan2.2 (Jiang et al., 2025). For the **streaming autoregressive** setting in Fig. 3, we build three pipelines using the Wan2.1-based Helios (Yuan et al., 2026). Qwen-Image + I2V generates the first frame with Qwen-Image (Wu et al., 2025a), while RefAlign and FFGO generate initial clips. Then Helios extends these outputs autoregressively. Reference tokens remain at the start of memory for continual access by video chunks. See Appendix C.4 for details.

### 4.2 Main Results

**Quantitative Results.** As shown in Tab. 1, *DomainShuttle* achieves competitive motion smoothness and text controllability while substantially outperforming baselines in cross-domain scenarios. We further evaluate the autoregressive generation over increasing frames in Fig. 3, with an extended




Figure 4: **Qualitative comparison** with existing methods. *DomainShuttle* outperforms existing methods in cross-domain scenarios, achieving flexible text controllability (*e.g.*, the yellow bus printed with the character) while preserving the features of the reference subject.

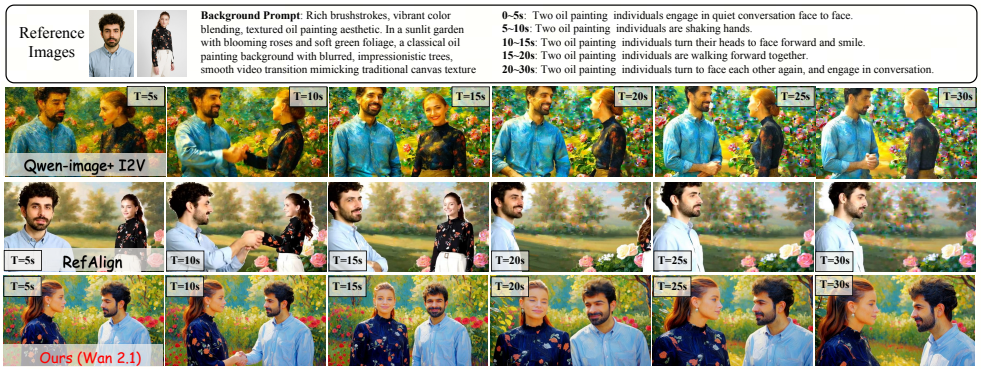



Figure 5: **Qualitative comparison** with streaming baselines for interactive long video generation. *DomainShuttle* further shows strong performance in streaming autoregression scenarios.

quantitative comparison shown in Tab. 12. Our method exhibits less degradation in these metrics than the baselines, showing robust autoregressive generation. Overall, *DomainShuttle* generates videos with flexibility and subject consistency in both bidirectional and autoregressive settings.

**Qualitative Results.** Fig. 4 shows that *DomainShuttle* preserves subject identity while accurately following cross-domain instructions, where baselines often sacrifice either fidelity or controllability. Fig. 5 further demonstrates streaming autoregressive generation. Our method maintains consistent subjects and visual style while accurately following interaction prompts, whereas baselines exhibit identity drift or incorrect interactions. More qualitative results are shown in Appendix C.6.

### 4.3 Ablation Study

All ablation experiments use the same training steps. Unless specified, the ablations use Wan2.2, while the streaming ablations use Wan2.1. More ablation studies are shown in Appendix C.7.

Table 2: **Ablation Studies** of each essential module.

(a) Bidirectional Architecture (Wan 2.2)

| ID | Setting | GMEScore↑ | CD-CLIP↑ | CD-Score↑ | DINO-I↑ | CLIP-I↑ |
|---|---|---|---|---|---|---|
| 0 | Naive Method | 0.662 | 0.600 | 0.696 | 0.347 | 0.667 |
| 1 | 0 + Dual Self-Attn | 0.672 | 0.609 | 0.712 | 0.364 | 0.678 |
| 2 | 0 + Domain-MoT | 0.686 | 0.628 | 0.781 | <u>0.380</u> | **0.687** |
| 3 | 2 + STRA-RoPE | 0.693 | 0.630 | <u>0.808</u> | 0.376 | 0.678 |
| 4 | 3 + CCL | **0.706** | **0.639** | **0.854** | **0.381** | <u>0.681</u> |

(b) Streaming Autoregressive Architecture (Wan 2.1)

| ID | Setting | GMEScore↑ | CD-CLIP↑ | CD-Score↑ | DINO-I↑ | CLIP-I↑ |
|---|---|---|---|---|---|---|
| 5 | Ours-Bi + Helios-V2V | 0.670 | 0.595 | 0.675 | 0.319 | 0.651 |
| 6 | Ours-AR w/o Domain-MoT | <u>0.672</u> | 0.608 | 0.713 | 0.347 | 0.670 |
| 7 | Ours-AR w/o STRA-RoPE | 0.668 | <u>0.614</u> | <u>0.747</u> | <u>0.375</u> | <u>0.681</u> |
| 8 | Ours-AR (full) | **0.681** | **0.621** | **0.773** | **0.382** | **0.688** |

Bi: *Bidirectional Architecture*; AR: *Streaming Autoregressive Architecture*.




(a) Effectiveness of the Domain-MoT (b) Effectiveness of the Remaining Modules

Figure 6: **Ablation Studies**. (a) Compared to Dual Self-Attn, Domain-MoT transfers all reference subjects into the fantasy domain. (b) Replacing STRA-RoPE with naive RoPE causes incorrect interactions, while removing CCL induces reference copying. Using both yields the best results.

**Effectiveness of Domain-MoT.** Ablation results of Domain-MoT are shown in Tab. 2 and Fig. 6 (a). The Naive Method directly concatenates reference image tokens with video tokens without additional operations. Dual Self-Attn uses separate video and reference branches without Domain-aware AdaLN. ID-0, ID-1, and ID-2 employ a naive RoPE scheme, in which the RoPE indices of the reference image tokens are concatenated to the video tokens along the temporal dimension. The Naive Method (ID-0) usually fails to transfer the reference subjects to the target domain. Dual Self-Attn (ID-1) partially improves domain transfer, but still cannot consistently transform all reference subjects. For instance, in Fig. 6 (a), only the woman is converted into the paper-cut style. In contrast, Domain-MoT (ID-2) successfully converts all humans into the target style. Quantitatively, Domain-MoT improves the CD-Score from 0.696 to 0.781 compared to the Naive Method.

**Effectiveness of the Remaining Modules.** The ablation results of STRA-RoPE and CCL are shown in Tab. 2 and Fig. 6 (b). Replacing STRA-RoPE with naive RoPE causes the reference image to be treated as a video frame, leading to incorrect subject interactions and weaker editability. For instance, the animated character appears at an incorrect spatial position and fails to attach to the cup. CCL mainly improves **controllability** in cross-domain scenarios, rather than fidelity. Removing CCL causes the model to directly copy the reference image. Combining STRA-RoPE and CCL, *DomainShuttle* (ID-4) achieves the best performance.

**Effectiveness for Autoregressive Generation.** The ablation results on 500-frame videos are shown in Tab. 2. Compared with the cascade of Ours-Bi and Helios-V2V (Yuan et al., 2026) (ID-5), Ours-AR retains Domain-MoT and STRA-RoPE throughout streaming generation. Removing either Domain-MoT (ID-6) or STRA-RoPE (ID-7) reduces subject consistency. The full model (ID-8) performs best, showing the effectiveness of both modules for autoregressive generation.

## 5 Conclusion

In this paper, we propose *DomainShuttle*, a novel framework for jointly optimizing subject consistency and generation flexibility. *DomainShuttle* introduces Domain-MoT to decouple video and reference processing while modeling target domain attributes through Domain-aware AdaLN, STRA-RoPE to temporally anchor and spatially organize reference tokens, and Cross-Pair Consistent Loss to reduce the influence of irrelevant subject features. Extensive experiments demonstrate that our model achieves competitive performance across both domain settings and generation paradigms.

### AI USE STATEMENT

We used generative AI tools to polish the language of the manuscript, improving clarity, grammar, and conciseness. The authors reviewed and revised all AI-assisted edits to ensure that they accurately reflected the intended scientific meaning. We take full responsibility for the final content, claims, and conclusions of this work.

### ETHICS STATEMENT

Human reference images used in our experiments are sourced from public datasets or generated by AI, reducing the need to collect personal images. We support responsible use of the proposed method, with respect for privacy and applicable data licenses, and discourage its use for deceptive or harmful content.

### REPRODUCIBILITY STATEMENT

The inference code for bidirectional generation is provided in the supplementary materials. We describe the architecture and training objectives of *DomainShuttle* in the methodology section and provide training configurations for both bidirectional and streaming autoregressive generation in the experimental setup. Data sources and processing procedures are documented in Sec. 3.3 and Appendix B. Additional inference settings and evaluation protocols are provided in Appendix C.1 and Appendix C.2, respectively.

## REFERENCES

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025a.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-as-prompt: Unified semantic control for video generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Yiyang Cai, Nan Chen, Rongchang Xie, Junwen Pan, Chunyang Jiang, Cheng Chen, Wen Zhou, Zhenbang Sun, Wei Xue, Wenhan Luo, and Yike Guo. Homie: Human-object centric video personalization via multimodal intelligent enhancement. *arXiv preprint arXiv:2607.18217*, 2026.

Jingxi Chen, Zongxia Li, Zhichao Liu, Guangyao Shi, Xiyang Wu, Fuxiao Liu, Cornelia Fermüller, Brandon Y Feng, and Yiannis Aloimonos. First frame is the place to go for video content customization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9243–9252, 2026a.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Human-centric video generation via collaborative multi-modal conditioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2939–2947, 2026b.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Multi-subject open-set personalization in video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6099–6110, 2025.

Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yunsheng Jiang, Gen Li, Xinghui Li, Liyang Chen, Siyu Zhou, Qian He, and Xinglong Wu. Phantom-data: Towards a general subject-consistent video generation dataset. In *The Fourteenth International Conference on Learning Representations*, 2026c.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

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

Yufan Deng, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Angtian Wang, Shenghai Yuan, Yiding Yang, Bo Liu, Haibin Huang, and Chongyang Ma. Cinema: Coherent multi-subject video generation via mllm-based guidance. *arXiv preprint arXiv:2503.10391*, 2025b.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked guidance for any-reference video generation with subject disentanglement. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Google. Nano banana pro. `https://gemini.google/au/overview/image-generation/?hl=en-AU`, 2025. Accessed: 2026-08-02.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zinan Guo, Pengze Zhang, Yanze Wu, Chong Mou, Songtao Zhao, and Qian He. Musar: Exploring multi-subject customization from single-subject dataset via attention routing. *arXiv preprint arXiv:2505.02823*, 2025.

Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in neural information processing systems*, 2025a.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21807–21818, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Daneul Kim, Jingxu Zhang, Wonjoon Jin, Sunghyun Cho, Qi Dai, Jaesik Park, and Chong Luo. Learning zero-shot subject-driven video generation using 1% compute. *arXiv preprint arXiv:2504.17816*, 2025.

Kling. Kling api. `https://klingai.com/global/`, 2025. Accessed: 2026-08-25.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, Mingshan Chang, Yuqiang Xie, Binjie Mao, Youqiang Zhang, Nuo Pang, Hao Zhang, Yuzhe Jin, Zhiheng Xu, Dixuan Lin, Guibin Chen, and Yahui Zhou. Skyreels-v3 technique report. *arXiv preprint arXiv: 2601.17323*, 2026a.

Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12737–12746, 2025a.

Hengjia Li, Haonan Qiu, Shiwei Zhang, Xiang Wang, Yujie Wei, Zekun Li, Yingya Zhang, Boxi Wu, and Deng Cai. Personalvideo: High id-fidelity video customization without dynamic and semantic degradation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19406–19416, 2025b.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. Bindweave: Subject-consistent video generation via cross-modal integration. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13146–13156, 2025a.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, volume 15067, pp. 366–384, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55, 2024.

OpenAI. Gpt-5.2. `https://platform.openai.com/docs/models/gpt-5.2`, 2025. Accessed: 2026-08-25.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Ziyang Song, Xinyu Gong, Bangya Liu, and Zelin Zhao. MV-S2V: multi-view subject-consistent video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers*, pp. 33:1–33:12, 2026.

Team Seedance, Siyan Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Xuyan Chi, Jian Cong, Qinpeng Cui, Qide Dong, Junliang Fan, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Lei Wang, Yuxin Song, Ge Wu, Haocheng Feng, Hang Zhou, Jingdong Wang, Yaxing Wang, and Jian Yang. Refalign: Representation alignment for reference-to-video generation. In *European Conference on Computer Vision*, pp. 647–664, 2026a.

Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *IEEE Transactions on Multimedia*, 28: 7137–7150, 2026b.

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

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18682–18692, 2025b.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16692–16701, 2025c.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *Advances in Neural Information Processing Systems*, 2025a.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12978–12988, 2025b.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

Zhongwei Zhang, Fuchen Long, Wei Li, Zhaofan Qiu, Wu Liu, Ting Yao, and Tao Mei. Region-constraint in-context generation for instructional video editing. *arXiv preprint arXiv:2512.17650*, 2025.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Forty-third International Conference on Machine Learning*, 2026.

In the Appendix, we detail streaming attention masks and KV caching in Appendix A, provide the construction of the training set in Appendix B, and present more experimental setups and results in Appendix C. We strongly recommend viewing the anonymous project page `https://DomainShuttle.github.io/` for a direct and clear demonstration of the capabilities of our model and improvements over previous methods.

## A Attention Masks and KV Caching for Streaming Generation

Fig. 7 shows the full attention mask for streaming generation only. Bidirectional generation uses unmasked joint self-attention (Eq. 2). The mask is shown with rows denoting queries and columns denoting keys. Reference queries attend to all reference tokens but no video tokens; video queries can attend to all reference tokens. For video-to-video attention, a clean chunk $x_s$ attends only to clean chunks $x_j$ with $j \leq s$. A noisy chunk $x_s^t$ attends to earlier clean chunks ($j < s$) and itself, but not to its clean counterpart or other noisy chunks. Attention within each permitted chunk is unrestricted. Applied at every layer, these masks block both direct access to future video chunks and indirect access through reference tokens. For cross-attention, $C_B$ denotes the encoded global prompt, while $C_s$ denotes the encoded concatenation of the global prompt and the specific description for chunk $s$. Reference queries attend only to $C_B$, whereas video queries for chunk $s$ attend only to $C_s$. Thus, changing the chunk-specific prompt does not change the reference branch's text condition.

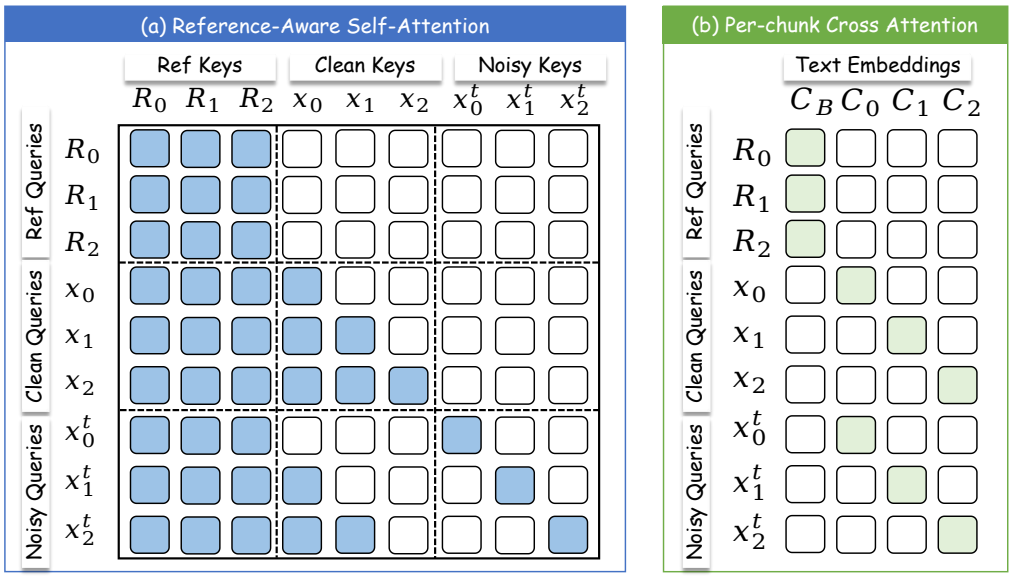



Figure 7: **Attention masks for streaming autoregressive generation.** Rows are queries and columns are keys. Colored cells indicate permitted attention and white cells indicate masked connections. (a) Reference queries cannot read video tokens, and noisy chunks access only references, earlier clean chunks, and their own tokens. (b) Reference queries attend only to $C_B$, which encodes the global prompt. Video queries in chunk $s$ attend only to $C_s$, which encodes the global prompt concatenated with the chunk-specific description.

For efficient inference, we cache the keys and values of fully denoised historical frames, storing keys before RoPE rotation. When denoising the newest chunk, we apply RoPE to the cached historical keys and the current chunk's queries and keys using their assigned window-relative indices, while keeping the reference and attention-sink indices fixed. This allows historical KV features to be reused as the window slides without retaining outdated temporal rotations.

# B CONSTRUCTION OF THE TRAINING DATASET

Our training datasets are all open-source datasets, primarily including open-source image personalization datasets and open-source video personalization datasets. Specifically, image personalization datasets consist of four parts: UNO (Wu et al., 2025b), Echo-4o (Ye et al., 2025), MUSAR (Guo et al., 2025), and Nano-Consistent-150K (Ye et al., 2025). We use single subject subsets of UNO and Nano-Consistent-150K, together with multi subject subsets of Echo-4o and MUSAR. We filter the datasets based on aesthetic quality and personalization quality scores using an MLLM-based evaluation system, ultimately obtaining 200K high-quality image samples to equip text-to-video models with basic personalization ability.

For video personalization datasets, we first filter low-quality videos using aesthetics and motion metrics. Then we use MLLM (Bai et al., 2025c) for visual-semantic alignment to remove low-quality references (*e.g.*, incomplete, blurred, or semantically mismatched subjects). We ultimately obtain 720K high-quality video samples. The video personalization datasets we use are also all open-source datasets. Each selected training sample typically contains no more than four reference images. The number of samples after filtering for each dataset is shown in Tab. 3.

Table 3: Detailed description of the training dataset.

| Dataset Name | Filtered Dataset Size | Modality | Category |
|---|---|---|---|
| UNO (Wu et al., 2025b) | 50K | Image | Single Subject |
| Nano-Consistent-150K (Ye et al., 2025) | 60K | Image | Single Subject |
| Echo-4o (Ye et al., 2025) | 60K | Image | Multi Subject |
| MUSAR (Guo et al., 2025) | 30K | Image | Multi Subject |
| Total Dataset | 200K | Image | - |
| Phantom-Data (Chen et al., 2026c) | 400K | Video | Single & Multi Subject |
| Opens2v (Yuan et al., 2025a) | 300K | Video | Multi Subject |
| Ditto-1M (Bai et al., 2025a) | 20K | Video | Single & Multi Subject |
| Total Dataset | 720K | Video | - |

# C ADDITIONAL EXPERIMENTS RESULTS

## C.1 IMPLEMENTATION DETAILS

**Implementation for Domain-aware AdaLN.** Domain attributes refer to the attributes in the generated video, rather than in the reference image. We use an MLLM (Bai et al., 2025c) to annotate the attributes of all subjects in videos for training and to annotate the attributes of all subjects in prompts for evaluation. During streaming generation, each subject's domain attribute remains fixed as prompts change. For a real person holding a cup printed with a cartoon character, the person's attribute is *real-world human*, while the printed character's attribute is *real-world object*. These four domain attributes support diverse visual styles through text conditioning. This separation allows stable target domain conditioning alongside flexible and fine-grained text control. Each subject's modulation parameters are broadcast only to its own reference tokens. This subject-wise modulation does not restrict the joint attention between reference and video tokens. In Tab. 2(a), ID-1 (Dual Self-Attn) and ID-2 (Domain-MoT) differ only in the introduction of Domain-aware AdaLN, with the remaining branch structure unchanged. ID-1 uses timestep-only reference modulation, whereas ID-2 additionally conditions it on the target domain attribute. Both variants use Temporal-Offset RoPE without CCL. This comparison therefore evaluates the contribution of Domain-aware AdaLN, rather than changes to the attention or feed-forward branches.

**Bidirectional Generation.** Before video personalization training, we update only the patch embedding and self-attention modules for 2,000 steps on 200K image-personalization samples with a batch size of 96. Tab. 4 summarizes the configurations for image initialization, video personalization, and the subsequent streaming stage. Both bidirectional training stages use a learning rate of $10^{-5}$ and a resolution of $480 \times 832$. As summarized in Tab. 5, training the bidirectional variants requires approximately 14,000 and 28,000 A100 GPU-hours for Wan2.1-14B and Wan2.2-14B, respectively. The table also lists reported training costs of representative baselines for reference. At inference, we use 50 sampling steps for Wan2.1 and 40 for Wan2.2. The classifier-free guidance

Table 4: Training details across different stages

| **Training Details** | 1. Image initialization | 2. Video personalization | 3. Streaming Training |
|---|---|---|---|
| Learning Rate | 1e-5 | 1e-5 | 5e-6 |
| Beta $(\beta_1, \beta_2)$ | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) |
| Training Steps | 2000 | 10000 | 5000 |
| Resolution | $480 \times 832$ | $480 \times 832$ | $480 \times 832$ |
| Reference | Multiple | Multiple | Multiple |

scale is set to 3 in Wan2.1. In Wan2.2, the high-noise classifier-free guidance scale is set to 4, while the low-noise guidance scale is set to 3. The flow shift parameter is set to 5 for both models.

Table 5: Comparison of training costs for bidirectional generation

| **Method** | **Base Model** | **Training Cost (Summary)** |
|---|---|---|
| Phantom (Liu et al., 2025) | Wan2.1-14B | 30k A100 hours |
| BindWeave (Li et al., 2026b) | Wan2.1-14B | 6k steps on 512 XPUs |
| *DomainShuttle* | Wan2.1-14B | 14k A100 hours |
| *DomainShuttle* | Wan2.2-14B | 28k A100 hours |

**Streaming Autoregressive Generation.** Following the anti-drifting strategy of Helios (Yuan et al., 2026), we add noise to clean historical latents during training to improve robustness to imperfect generated context. In inference, we use a sliding window with four latent frames per chunk. Under STRA-RoPE, reference tokens remain at temporal index 0, while the first video latent frame is retained throughout generation as an attention sink at index 1. The remaining historical latents and the current chunk slide within the window, with their temporal indices reassigned within 2–20 as the window advances. Cross-pair Consistent Loss (CCL) is not used during autoregressive training. The autoregressive stage builds on the reference-conditioning capabilities learned by the bidirectional model and focuses on adaptation to causal generation, without additional CCL computation.

**More Explanation about STRA-RoPE.** For example, with two reference images for subject 0 and one reference image for subject 1, STRA-RoPE assigns spatial offsets $(h, w)$, $(h, 2w)$, and $(2h, 3w)$, respectively. For tokens at the same spatial position $(j, k)$ in two consecutive reference images, the RoPE offset from the earlier image to the next image is $(0, h\Delta m, w)$, where $\Delta m = 0$ for the same subject and $\Delta m = 1$ when crossing to the next subject.

STRA-RoPE applies a uniform spatial offset to each reference image, preserving relative positions among its tokens while changing their positions relative to other references and video tokens. These offsets are used during fine-tuning and inference, allowing the model to adapt to the shifted coordinates rather than relying solely on the pretrained backbone's extrapolation ability. No additional positional interpolation or frequency rescaling is applied.

**More Explanation about CCL.** CCL uses two forward passes of the same model with shared parameters, with gradients stopped through the second prediction. The stopped-gradient prediction is recomputed from the current parameters at each step, rather than produced by a separate frozen teacher. Flow matching provides ground-truth supervision, while CCL serves as an auxiliary consistency regularizer.

### C.2 Evaluation Details

**Evaluation Dataset.** The 120 in-domain test cases are all from **open-source** OpenS2V-Eval (Yuan et al., 2025a), with 30 each for multi-human, multi-object, human-object interaction, and Hard_dev cases. Because some methods support at most four subjects, we use the first four subjects in each Hard_dev case as references. For the cross-domain evaluation dataset, we construct 40 multi-subject samples for real-world subjects in fantasy domains, 40 multi-subject samples to map the fantasy

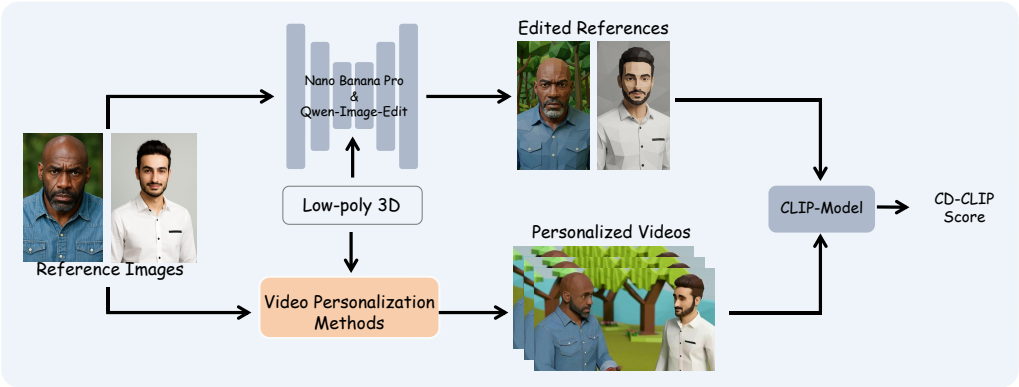



Figure 8: The CD-CLIP evaluation pipeline, illustrated with Nano Banana Pro. Both Nano Banana Pro and Qwen-Image-Edit generate edited references according to the domain transformation prompt. Then, each video personalization method generates videos conditioned on the prompt. Finally, CLIP computes the cosine similarity between the generated frames and the edited references. Both variants use the same Hungarian matching, segmentation, CLIP similarity, and aggregation protocol, assigning zero to confirmed missing subjects.

Table 6: The 32 target styles in the 40 real-to-fantasy test cases. Bold labels indicate styles seen during our fine-tuning. Non-bold labels indicate unseen styles. Names are shortened for readability.

| **Comic book** | **Illustration** | **Cyberpunk** | **Line art** |
|---|---|---|---|
| **Steampunk** | **Pop art** | **Bronze sculpture** | **Japanese anime** |
| **Minimalism** | **Borderlands** | Disney animation | Pencil sketch |
| Chocolate sculpture | Flickering translucent hologram | LEGO | Pixar |
| Charcoal drawing | Neon light art | Pixel art | Silver sculpture |
| Vector | Ghibli | Paper cutting | Low Poly |
| Adventure Time | Oil painting | Archer | 3D |
| Moe anime | Watercolor | Rusted iron sculpture | Ukiyo-e |

subjects to the real-world objects, and 40 multi-subject samples for interactions between real-world and fantasy subjects. Both bidirectional and streaming autoregressive settings are evaluated on these 240 cases. Streaming quantitative metrics are computed on generated videos of 500 frames. The training and test sets are strictly separated in reference subjects and prompts. No test reference image, target video, or near-duplicate video appears in the training data. Test characters and subject combinations are also excluded from training. At inference, *DomainShuttle* is conditioned only on reference images and text prompts, without any reference video input.

**Target Styles.** The 40 real-to-fantasy test cases cover 32 distinct style-prompt labels, listed in Tab. 6. These span drawing and painting, animation, material-based rendering, and digital aesthetics.

Only 10 styles are included in training through 10K edited videos from Ditto-1M (Bai et al., 2025a) as augmentation to preserve the base model's generative capabilities. The other two data sources are publicly available. OpenS2V is also used by RefAlign (Wang et al., 2026a) and BindWeave (Li et al., 2026b), while Phantom-Data is used by RefAlign and Phantom (Liu et al., 2025). These shared data sources help support fair comparisons with existing methods. All remaining test styles are excluded from fine-tuning. Among these 40 test cases, 10 cases evaluate styles seen during fine-tuning, while the remaining 30 cases evaluate unseen styles. The ablation in Tab. 13 shows that our method maintains strong cross-domain performance without this augmentation, suggesting that the gains primarily stem from the proposed method rather than the additional editing data.

We further compare normalized Style Scores on styles seen and unseen during fine-tuning. The Wan2.1 variant scores $0.857$ and $0.841$ on seen and unseen styles, respectively, while the Wan2.2 variant scores $0.898$ and $0.878$. These small gaps suggest that our method is not limited to styles seen during fine-tuning and maintains competitive performance on unseen styles.

**Evaluation Metrics.** For all subject similarity metrics under the bidirectional architecture, we uniformly sample 16 frames from the videos generated by each method, compute the score for each frame, and take the average score as the final result. For streaming autoregressive generation, we sample one frame every 10 frames to calculate scores.

**Video Quality.** We select aesthetic score and motion smoothness from the OpenS2V-Nexus benchmark [1] to evaluate the video quality generated by each method. We test all 240 videos in the test set and calculate the average aesthetic and smoothness scores of these videos.

**Text following.** We use GMEScore to evaluate alignment between generated videos and the prompt. We test all 240 videos in the test set and calculate the average GMEScore of these videos.

**DINO-I and CLIP-I.** We evaluate these metrics only on the 120 in-domain test cases. After subject segmentation, we compute cosine similarity between corresponding generated and reference subject regions using DINO-v2 (Oquab et al., 2023) and CLIP (Radford et al., 2021) features, respectively. Scores are averaged over sampled frames and subjects within each case, then across all 120 cases.

**Style Score.** We evaluate the 40 real-to-fantasy test cases using a subject-level adaptation of VBench's Appearance Style metric (Huang et al., 2024). Grounding DINO (Liu et al., 2024) localizes the reference subjects in generated frames, and their bounding boxes define the subject crops. We use the same OpenAI CLIP ViT-B/32 checkpoint and image preprocessing as VBench. We use bicubic resizing so that the shorter side is 224 pixels, followed by a $224 \times 224$ center crop.

For each subject crop $I$ and its target-style text $q$, we directly tokenize $q$ without an additional text template and compute the raw score $s(I, q) = \mathrm{logit}_{\mathrm{CLIP}}(I, q)/100$, following VBench. Here, $\mathrm{logit}_{\mathrm{CLIP}}$ is the model's image–text output including its learned logit scale. We average the raw scores over sampled frames for each subject, then over subjects within each case, and finally across the 40 cases to obtain $\bar{s}$. The reported Style Score follows VBench's fixed-bound normalization:

$$\mathrm{StyleScore} = \frac{\bar{s} - 0.0009}{0.2855 - 0.0009}. \tag{7}$$

The bounds are the predefined Appearance Style constants in VBench, not estimated from the compared methods. No clipping is applied. All methods use the same scoring and normalization protocol. The reported value is a normalized similarity score, not a style-correctness rate.

**CD-CLIP.** Fig. 8 illustrates the evaluation pipeline. We separately use Nano Banana Pro (Google, 2025) and Qwen-Image-Edit (Wu et al., 2025a) to edit references according to the requested domain transformation. Edited references are manually checked, and unsuccessful edits are regenerated with adjusted generation settings. Fewer than 10 reference images required re-editing. The approved references are fixed and shared across all evaluated methods. Subjects instructed to retain their original domain use the original references.

For multi-subject cases, we use Grounding-DINO Base and SAM2-Hiera-Large to extract subject regions, with box and text thresholds of $0.2$ and $0.3$, respectively. If detection fails, we lower the box threshold to $0.1$ and retry, keeping the text threshold unchanged. Candidate boxes with an IoU above $0.7$ are treated as duplicate detections and deduplicated before matching. We use Hungarian matching to maximize the total CLIP cosine similarity between reference subjects and detected instances in each sampled frame, with each subject and instance participating in at most one match. Matching is performed at the subject level.

Unmatched subjects and ambiguous correspondences are manually reviewed to distinguish subject absence from detection or matching errors. If a subject is present but localization fails, its region is manually corrected before scoring. Overall, approximately $4\%$ of all evaluated cases require this additional manual correction, keeping the manual workload limited. A subject confirmed absent from a sampled frame receives a score of $0$ for that frame, without full-image fallback. We average scores over all sampled frames for each subject, then over all reference subjects within each case,

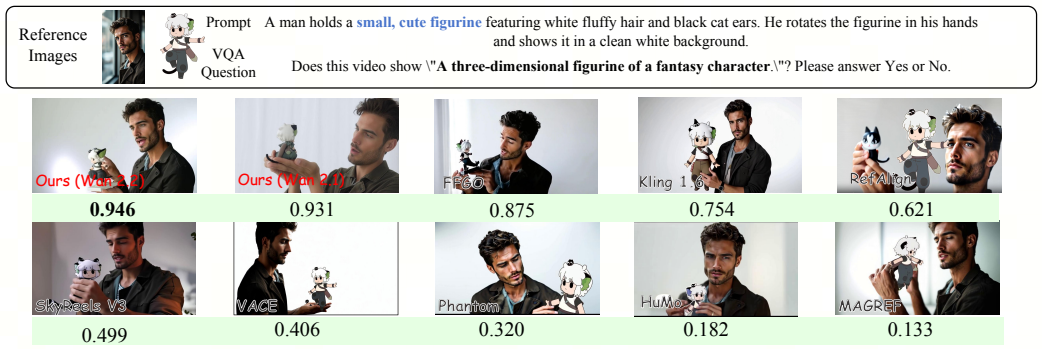
[1] https://github.com/PKU-YuanGroup/OpenS2V-Nexus/tree/main/eval

Table 7: MLLM prompt for Cross-Domain (CD) Score.

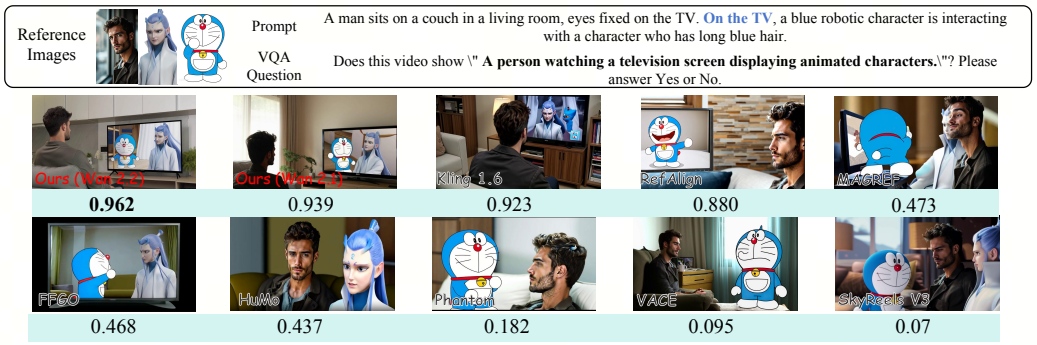
| |
|---|
| Guidelines: You are an AI evaluator for cross-domain subject-driven text-to-video generation. You will receive a reference image, a generated video frame, and a reference caption specifying the intended output. |
| Reference Image 1: the reference subject or subjects.<br>Image 2 from Generated Video: the generated frame to evaluate.<br>Reference Caption: the intended representation of each subject and the requested scene and relationships.<br><br>Task:<br>First, identify each reference subject and its intended representation from the caption (*e.g.*, a real person, a stylized character, a figurine, or a printed/painted character). Different subjects may require different representations; do not assume that every subject must undergo a domain change.<br>Then assess each subject separately along three aspects: (1) **Identity**: whether its distinctive, identity-bearing features remain recognizable; (2) **Representation**: whether its style, material, and form match the caption; and (3) **Relations**: whether its placement, attachment, and visible interactions with other subjects match the caption. For example, a real person holding a cup printed with a character should remain a real person, while the character should appear on the cup rather than as a separate figure.<br>Preserve reference features only insofar as they are compatible with the requested representation. Do not penalize instructed changes in style, material, lighting, or other appearance attributes, and do not reward copying attributes that contradict the caption. Distinguish identity loss from failure to realize the requested representation. |
| Scoring Rules:<br>1. Assign one overall integer score from 1 to 5 based on these separate checks. In multi-subject cases, consider every requested subject; success on one subject must not conceal a missing or incorrect subject.<br>2. Metric:<br>**5**: All requested subjects are recognizable, correctly represented, and placed in the requested visible relationships.<br>**4**: Identities, target representations, and principal relationships are preserved, with only minor local errors.<br>**3**: The result partially meets the request, but noticeable errors affect a subject's identity, representation, or a requested relationship.<br>**2**: Major errors affect a principal subject or relationship. Some identity cues remain recognizable, but the requested representation or scene is largely incorrect; unchanged reference appearance despite an explicitly requested transformation is one such failure.<br>**1**: The principal reference subject is absent or unrecognizable, or the output is unrelated to the requested subjects and scene.<br><br>3. Higher scores indicate better identity preservation under the subject-specific representations and relationships required by the caption. A subject that correctly remains in its original domain must not be penalized for lacking a transformation.<br>Strict Output Requirement: Return only the numerical score (no explanations, text descriptions, or additional comments). |

and finally across cases. Missing subjects therefore remain in the averaging denominators, including when they are absent throughout the video. We report the two scores in Nano Banana Pro / Qwen-Image-Edit order.

**CD-Score.** We use GPT-5.2 (OpenAI, 2025) and Qwen3-VL-8B-Instruct (Bai et al., 2025b) as alternative evaluators under the shared prompt and scoring rules in Tab. 7. Both assess subject identity, target representation, and visible relationships specified by the prompt. We apply the same averaging and normalization to each evaluator's scores and report the results separately in GPT-5.2




(a) Relation Alignment evaluation for fantasy-to-real transformations




(b) Relation Alignment evaluation for real–fantasy interactions

Figure 9: **Qualitative examples of Relation Alignment (RA).** Each case shows reference images, the generation prompt, the evaluation question, and representative frames with video-level VQAScores. (a) Fantasy-to-real transformation into a three-dimensional figurine. (b) Real–fantasy interaction in which a person watches a television displaying animated characters. Higher scores indicate greater model-assessed confidence that the evaluated condition is satisfied.

/ Qwen3-VL-8B-Instruct order. These are two evaluator variants of one metric, not separate metrics or an averaged ensemble score. The single-valued CD-CLIP results in the ablation tables use Nano Banana Pro and CD-Score results in the ablation tables use GPT-5.2. Across the 11 method variants in Tab. 1, the CD-Scores from GPT-5.2 and Qwen3-VL-8B-Instruct exhibit a strong method-level Pearson correlation ($r = 0.982$), indicating consistent scoring trends across evaluators.

For each evaluator, we average the frame-level scores within each case and then across cases:

$$\text{CD-Score} = \frac{1}{5N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{t=1}^{T_i} s_{i,t}, \tag{8}$$

where $s_{i,t} \in \{1, 2, 3, 4, 5\}$ and $T_i$ is the number of evaluated frames in case $i$. $N$=120. No additional normalization is applied.

To further assess agreement between CD-Score and human judgment, three domain experts independently evaluate *DomainShuttle* (Wan2.1), Kling1.6, RefAlign, and VACE-Wan2.2 on a 30-case subset containing 10 cases each of real-to-fantasy transformations, fantasy-to-real transformations, and real–fantasy interactions. They follow the rubric in Tab. 7 without access to automatic scores. The system-level Pearson correlations between mean expert ratings and CD-Score evaluated by GPT-5.2 and Qwen3-VL-8B-Instruct on the same subset are 0.837 and 0.812, respectively, supporting agreement between automatic and expert evaluations across the four methods on this subset.

**Relation Alignment (RA).** We use the Qwen3.5-9B (Qwen Team, 2026) version of VQAScore (Lin et al., 2024), following its standard computation protocol, to quantify model-assessed confidence in the fulfillment of fantasy-to-real transformations and real-fantasy interactions. Evaluation covers 80 cases: 40 cases of fantasy-to-real transformations and 40 cases of real-fantasy interactions. For transformations, we assess whether fantasy subjects take the intended real-world form, such as figurines or acrylic stands, or are correctly bound to a specified carrier, such as a character printed on a cup. For interactions, we assess whether real-world and fantasy subjects exhibit the requested interactions and spatial relationships. Tab. 8 summarizes the VQAScore configuration. Detailed scores for transformations and interactions are reported in Tab. 9.

Table 8: VQAScore configuration for Relation Alignment.

| Item | Configuration |
|---|---|
| Implementation | `t2v-metrics` 3.1 |
| Question | `Does this video show "{}"? Please answer Yes or No.` |
| Description | The case's `text` field describes the target form/carrier binding or actions/spatial relations. Neither the full generation prompt nor reference images are supplied. |
| Answer | `Yes` (case-sensitive, no punctuation), encoded without special tokens and verified as one token. |
| Chat template | The checkpoint's template with a generation prompt and thinking disabled (empty think block); one user message containing the video and question, without a system prompt. |
| Generation | One output token, greedy decoding (`do_sample=False`), with output scores and structured generation output enabled. |
| Scoring | Probability of `Yes` at the first answer position (`score_position="start"`), regardless of the token actually generated. |

Fig. 9 illustrates RA scores for two representative cases. In (a), the question assesses whether the fantasy character takes the form of a three-dimensional figurine. Higher-scoring outputs generally depict a physical figurine rather than a flat character image, with our Wan2.2 and Wan2.1 variants scoring $0.946$ and $0.931$, respectively. In (b), the question assesses whether a person is watching a television displaying animated characters, requiring both the viewing action and the characters' placement on the screen. Our variants score $0.962$ and $0.939$, respectively, while some baselines that do not clearly depict this combined relationship generally receive lower scores. These examples provide qualitative support for RA's sensitivity to the fantasy-to-real transformations and real–fantasy interactions.

Table 9: **Relation Alignment** Scores on 40 fantasy-to-real transformation cases and 40 real–fantasy interaction cases. Higher is better.

| Method | Transformation ↑ | Interaction ↑ | Overall ↑ |
|---|---|---|---|
| Kling 1.6 | 0.740 | <u>0.731</u> | 0.736 |
| MAGREF | 0.668 | 0.600 | 0.634 |
| SkyReels-V3 | 0.747 | 0.675 | 0.711 |
| Phantom | 0.519 | 0.656 | 0.587 |
| HuMo | 0.532 | 0.441 | 0.487 |
| BindWeave | 0.502 | 0.486 | 0.494 |
| RefAlign | 0.621 | 0.543 | 0.582 |
| FFGO-Wan2.2-14B | 0.713 | 0.697 | 0.705 |
| VACE-Wan2.2-14B | 0.730 | 0.615 | 0.672 |
| Ours (Wan2.1-14B) | <u>0.753</u> | 0.709 | 0.731 |
| Ours (Wan2.2-14B) | **0.857** | **0.797** | **0.827** |

**Interpreting Fidelity across Backbones.** In Tab. 1, the Wan2.2 variant improves CD-Score and Style Score over Wan2.1, while DINO-I decreases from $0.397$ to $0.381$ and CLIP-I from $0.694$ to

Table 10: Guidelines for human preference evaluation.

| |
|---|
| Guidelines: subject-driven video (S2V) personalization is an important downstream task of text-based video generation, aiming to generate videos based on user-provided images and prompts. S2V mainly involves two scenarios: in-domain (*e.g.*, multi-human, multi-object, and human-object interaction) scenarios and cross-domain (*e.g.*, real-world subject to fantasy domain, fantasy subject to real-world domain, and the interaction of the real-world and fantasy subjects) scenarios. Please watch the following randomly selected videos, reference images, and prompts. Compare their effects and evaluate the generated video based on three metrics: |
| 1. Overall Video Quality: Comprehensively evaluate the overall quality of the generated videos from three aspects: aesthetic quality, the smoothness of subject motions (avoiding static or frozen subjects and frame discontinuities), and the naturalness of color, texture, and saturation.<br><br>2. Evaluate text controllability based on the consistency between the generated video and the input text description (*e.g.*, corresponding real-world or fantasy domain descriptions, stylistic attributes, and subject interaction alignment).<br><br>3. Subject Consistency: Evaluate whether each generated subject is recognizable as the corresponding reference subject while following the prompt. The generated subject should preserve identifying features, such as facial structure, body shape, or distinctive markings, as long as they are compatible with the requested style or form. Changes in color, material, or texture requested by the prompt should not be penalized. For example, a silver sculpture does not need to retain the original skin or clothing colors. Each subject should be judged according to its intended representation, rather than only how closely it copies the reference appearance. |
| Please rank these methods across these three metrics. |

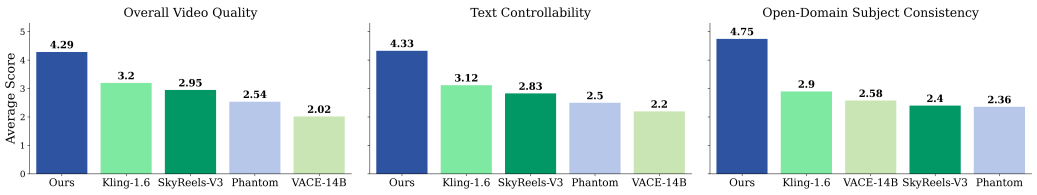



Figure 10: **Human preference evaluation**. *DomainShuttle* receives higher preference scores than baselines in video quality, text controllability, and subject consistency.

0.681. These metrics favor visual similarity to the reference image but do not account for prompt-requested changes. For example, turning a front-facing subject to a side view may lower the scores despite correctly following the prompt and preserving the subject's identity. This metric sensitivity may partly explain the lower in-domain scores alongside stronger cross-domain adaptation, where following the requested transformation is essential. Therefore, we should interpret reference similarity together with text controllability and cross-domain performance, rather than viewing it merely as an isolated metric for measuring personalization quality.

### C.3 Human Preference Evaluation

We invite 40 volunteers to compare *DomainShuttle* with four baselines. Each volunteer evaluates 20 cases randomly sampled from the 240-case test set, with five method outputs per case (100 videos in total). Method names are hidden throughout evaluation. The instructions provided to the participants are shown in Tab. 10. For the comparison in Fig. 10, we use the Wan2.1-based variant of *DomainShuttle*. VACE-14B refers to VACE-Wan2.2-14B. The participants rank the five outputs by video quality, text controllability, and open-domain subject consistency, assigning distinct scores from 5 (best) to 1 (worst). The reported means reflect human relative preference rather than absolute quality. Our method receives higher preference scores across all three criteria.

To further compare with the recent baseline RefAlign (Wang et al., 2026a), we recruit 40 volunteers for a pairwise preference evaluation. Each volunteer evaluates 20 cases independently sampled from the same 240-case test set, with method names hidden. We use the Wan2.1-based *DomainShuttle* and compare outputs generated from the same reference images and prompts. Participants indicate their preference or a tie for each of the three criteria. As shown in Fig. 11, our method receives higher win rates than RefAlign across criteria.

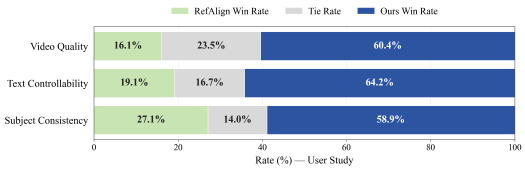



Figure 11: **Human preference against RefAlign**, using Wan2.1-based *DomainShuttle*.

## C.4 Discussion on Baselines

### C.4.1 Streaming Baseline Protocol and Ablation Controls

For the streaming comparison in Fig. 3, Qwen-Image + I2V uses a reference-conditioned initial frame, while RefAlign + Helios and FFGO + Helios use reference-conditioned initial clips. The three cascaded baselines follow Helios (Yuan et al., 2026) for history, first-frame sink, and context-window management. In inference, reference images are placed at the start of memory, so that video chunks can continually access them. They therefore directly retain reference features during streaming video generation. Ours-Bi + Helios-V2V (ID-5) uses the same conditioning settings through an initial clip in Tab. 2 (b). *DomainShuttle* can also directly interact with the original reference tokens at every chunk.

For the streaming ablations in Tab. 2(b), each variant is initialized from its corresponding bidirectionally trained Wan2.1 checkpoint and further trained for 5,000 steps. ID-6 inherits the bidirectional variant trained without Domain-MoT, ID-7 inherits the variant trained with Temporal-Offset RoPE instead of STRA-RoPE, and ID-8 inherits the full model. Each variant therefore retains its configuration across both training stages, rather than removing or replacing modules only when streaming fine-tuning begins. For ID-7 in Tab. 2(b), the $N$ reference images occupy fixed temporal indices $0, \ldots, N-1$. The video sliding window occupies a fixed temporal range immediately after this reference prefix, with video indices reassigned within that range as the window advances. Thus, reference-to-video temporal distances do not grow with the generated video length.

### C.4.2 Comparison with Closely Related Methods

We compare *DomainShuttle* with closely related methods in terms of modeling objectives, domain conditioning, reference positioning, and training supervision.

**Comparison with MoT and Video-As-Prompt.** MoT (Liang et al., 2025b; Deng et al., 2025a) uses modality-specific parameters with global attention for efficient multimodal modeling. Video-As-Prompt (Bian et al., 2026) couples a trainable reference expert with a frozen target backbone through joint attention, using independent QKV projections, feed-forward layers, and normalization to transfer semantics from reference videos. Domain-MoT differs in explicitly conditioning reference processing on the subject's desired output domain. Its reference AdaLN uses both the diffusion timestep and the target domain attribute, whereas its video AdaLN uses only the timestep. This separates the subject features supplied by reference images from the target domain control specified by the prompt, supporting subject preservation during domain transformation in both bidirectional and streaming autoregressive generation. The distinction lies in target domain modulation, beyond independent branches and joint attention. In our component ablation (Tab. 2(a)), Domain-MoT improves CD-Score from $0.712$ to $0.781$ over Dual Self-Attn.

**Comparison with MV-S2V and Temporal-Offset RoPE.** Both MV-S2V (Song et al., 2026) and Video-As-Prompt (Bian et al., 2026) use temporal offsets for reference positioning: MV-S2V distinguishes different subjects and their views along the temporal axis, while Video-As-Prompt separates reference and target videos. In contrast, STRA-RoPE anchors all reference tokens at temporal index $0$, using subject-aware spatial offsets to distinguish subjects and group references of the same

Table 11: Additional quantitative comparison with SDV-Gen. Higher values indicate better performance for all metrics. Paired CD-CLIP and CD-Score values follow the same conventions as in the main comparison table.

| Method | AES↑ | MS↑ | GME↑ | CD-CLIP↑ | Style Score↑ | Style Consist↑ | CD-Score↑ | RA↑ | DINO-I↑ | CLIP-I↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| SDV-Gen (Kim et al., 2025) | 0.413 | 0.923 | 0.645 | 0.563 / 0.587 | 0.687 | 0.873 | 0.641 / 0.682 | 0.536 | 0.291 | 0.618 |
| Ours (Wan2.1-14B) | **0.512** | **0.981** | **0.691** | **0.629 / 0.648** | **0.845** | **0.955** | **0.794 / 0.785** | **0.731** | **0.397** | **0.694** |

subject. Reference coordinates remain fixed as the video window slides. Additional reference images can be accommodated through spatial offsets **without expanding the temporal index range**, making this design well suited to fixed-window streaming autoregressive generation.

The default baseline previously described as *naive RoPE* in our ablations is Temporal-Offset RoPE. Thus, ID-2 versus ID-3 in Tab. 2 (a) directly compares Temporal-Offset RoPE with STRA-RoPE for bidirectional generation, and ID-7 versus ID-8 in Tab. 2 (b) provides the corresponding streaming autoregressive comparison. In particular, ID-7 replaces STRA-RoPE with Temporal-Offset RoPE. For bidirectional generation, STRA-RoPE improves CD-Score from $0.781$ to $0.808$ and GMEScore from $0.686$ to $0.693$, supporting its effectiveness over the evaluated temporal-offset baseline for cross-domain generation. These are controlled comparisons of positional encoding within our framework, rather than full-system comparisons with MV-S2V or Video-As-Prompt.

**Comparison with more personalization methods.** For academic baselines, our main comparisons prioritize the most closely related peer-reviewed methods with publicly available implementations compatible with our evaluation setting. We qualitatively discuss the following additional methods. Movie Weaver (Liang et al., 2025a) associates references with concepts using anchored prompts, and CustomVideo (Wang et al., 2026b) employs subject-specific fine-tuning with mask-guided attention. Although HunyuanCustom (Hu et al., 2025) reports multi-subject results, its public release supports only single-subject customization, preventing a direct comparison under our multi-subject protocol. SDV-Gen (Kim et al., 2025) disentangles identity and motion by stochastically alternating between image and video training objectives. DomainShuttle instead processes video and reference-image tokens through separate branches and introduces target domain conditioning into the reference branch. Its decoupling focuses on video and reference processing, rather than identity and motion. We additionally provide a quantitative comparison with SDV-Gen in Tab. 11, where DomainShuttle (Wan2.1-14B) outperforms SDV-Gen across all reported metrics.

**CCL versus Cross-Pair Data Supervision.** Different reference sets of the same subject may contain incidental variations in viewpoint, illumination, occlusion, and motion blur. Cross-pair data exposes the model to these variations, whereas CCL further explicitly penalizes their influence on generation. With the noisy video latent, timestep, and text condition fixed, CCL aligns predictions conditioned on two reference sets paired with the same target video, using a stop-gradient target branch with the current model parameters. This discourages reference-specific variations from altering the predicted video content, while allowing different prompts and target videos to specify different outputs.

Both ID-3 and ID-4 in Tab. 2(a) sample from the same reference pool (Sec. 3.3). ID-3 randomly selects one reference set per training step and applies standard flow matching (Eq. 1). ID-4 additionally samples a second reference set for the stop-gradient consistency target. Adding CCL improves CD-Score from $0.808$ to $0.854$ and GMEScore from $0.693$ to $0.706$ (Tab. 2(a), ID-3 versus ID-4), supporting the benefit of explicit consistency regularization beyond cross-pair data alone. Applying flow matching supervision to both reference sets requires two forward passes and gradient computation through both branches. In contrast, CCL uses two forward passes but backpropagates through only one branch, avoiding the additional backward computation associated with supervising both reference sets.

### C.5 More Quantitative Comparisons.

Tab. 12 extends the comparison in Fig. 3 with a broader set of metrics covering video quality, text controllability, and cross-domain performance. Every metric is computed using frames sampled at 10-frame intervals from the full 500-frame videos. *DomainShuttle* achieves the highest mean score

Table 12: Mean scores for the comparison in Fig. 3. Higher is better; the best scores are bold.

| Method | AES | MS | GME | CD-CLIP | CD-Score | Style Score | Style Consist. | RA | Imaging Quality |
|---|---|---|---|---|---|---|---|---|---|
| Qwen-Image + Helios I2V | 0.501 | 0.957 | 0.678 | 0.613 / 0.628 | 0.752 / 0.731 | **0.864** | 0.875 | 0.703 | 0.710 |
| RefAlign + Helios V2V | 0.404 | 0.954 | 0.659 | 0.587 / 0.610 | 0.713 / 0.641 | 0.708 | 0.804 | 0.564 | 0.714 |
| FFGO + Helios V2V | 0.348 | 0.951 | 0.643 | 0.573 / 0.593 | 0.616 / 0.625 | 0.687 | 0.794 | 0.697 | 0.647 |
| *DomainShuttle* | **0.511** | **0.961** | **0.681** | **0.621 / 0.631** | **0.773/ 0.753** | 0.837 | **0.933** | **0.727** | **0.731** |

Table 13: **Ablation Studies** of Ditto 1M on Wan 2.2.

| Method | CD-CLIP ↑ | CD-Score ↑ | DINO-I ↑ | CLIP-I ↑ | Style Score ↑ |
|---|---|---|---|---|---|
| w/o Ditto 1M | 0.632 | 0.820 | **0.389** | **0.687** | 0.867 |
| w Ditto 1M | **0.639** | **0.854** | 0.381 | 0.681 | **0.883** |

on four metrics in Fig. 3. Paired CD-CLIP values use edited references from Nano Banana Pro / Qwen-Image-Edit, and paired CD-Score values use GPT-5.2 / Qwen3-VL-8B-Instruct as evaluators, respectively. Qwen-Image + Helios I2V obtains the highest Style Score, potentially benefiting from Qwen-Image's image-level editing capabilities (Wu et al., 2025a), which establish the target style in the initial frame before Helios extends it into a video. However, strong initial-frame stylization alone does not ensure consistent subjects, styles, or interactions throughout the video. *DomainShuttle* outperforms this pipeline on all other reported metrics, including video quality and overall cross-domain performance.

### C.6 More Qualitative Comparisons.

We present additional bidirectional qualitative comparisons between *DomainShuttle* and the baselines in the Appendix, as shown in Fig. 13, Fig. 14, and Fig. 15. Our method outperforms existing methods across all three scenarios: mapping real-world subjects to the fantasy domain, mapping fantasy subjects to the real world, and interactions between fantasy and real-world subjects. We also present more streaming autoregressive comparisons in Fig. 16. These qualitative comparisons demonstrate that *DomainShuttle* can optimize subject consistency and text controllability across various complex applications.

### C.7 More Ablation Study

**Use of Ditto-1M.** In addition, we conducted an ablation study on whether to use Ditto-1M, as shown in Tab. 13. Without Ditto-1M, our method remains effective and still achieves cross-domain SOTA performance, improving the Style Score by 18.9% over baselines (0.729 of VACE Wan-2.2).




Figure 12: **Ablation Studies**. STRA-RoPE uses a subject-decoupled offset to better bind these references than a reference-decoupled offset.

**Decoupling Strategies of STRA-RoPE.** In some open-domain scenarios, multiple reference images may represent different attributes of the same subject. To address this, STRA-RoPE adopts a strategy in which multiple reference images of the same subject are offset only along the width dimension in RoPE, rather than applying offsets along both height and width for different reference images by default. As shown in Fig. 12, compared with the reference-decoupled strategy, the subject-decoupled strategy better binds multiple reference images of the same subject.

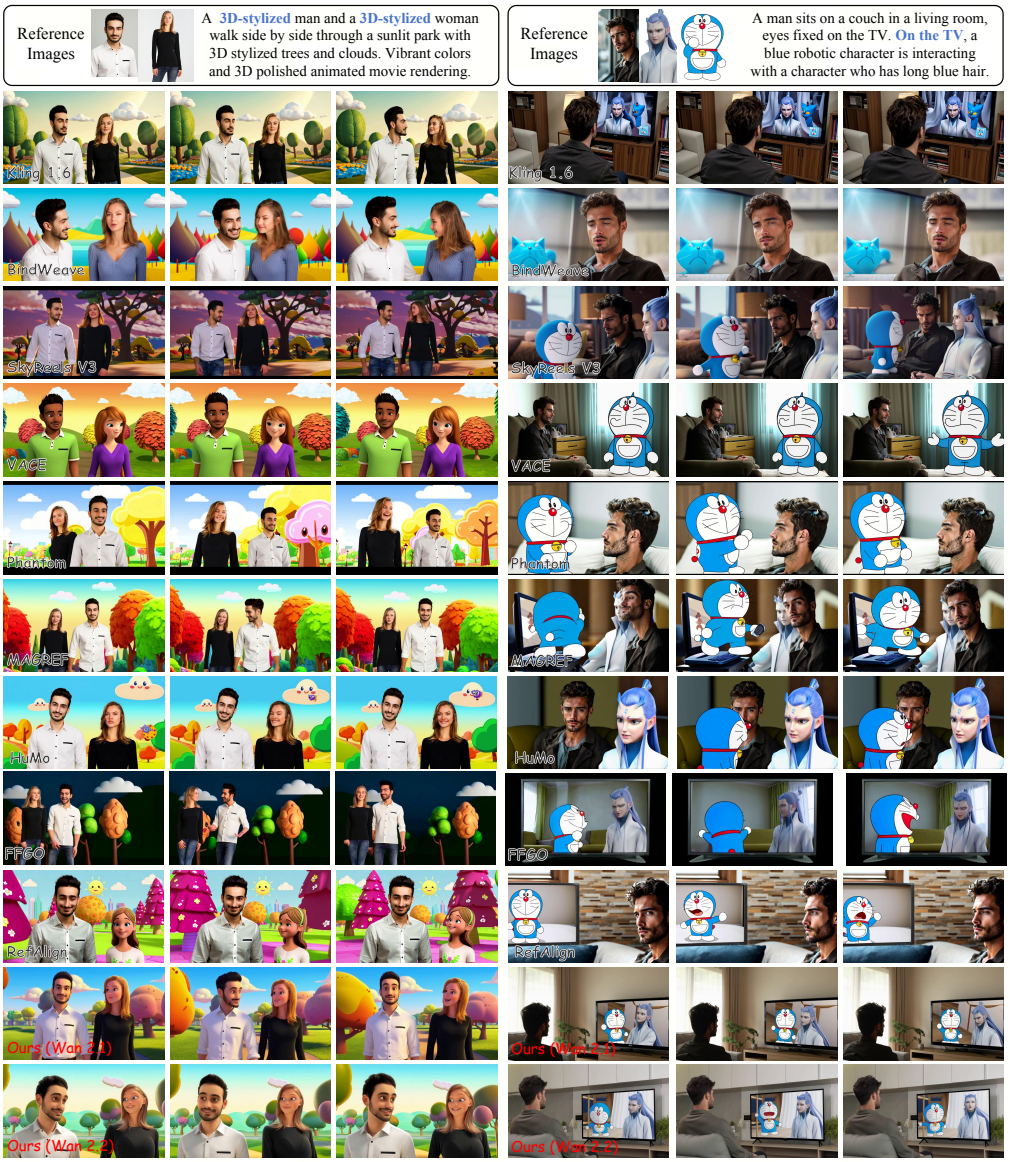



Figure 13: **Additional Qualitative comparison** with existing methods. *DomainShuttle* further demonstrates strong performance in more challenging cross-domain scenarios. Please watch the video via an anonymous website for the best experience.

### C.8 Limitations

Built on the Wan models, *DomainShuttle* inherits the limitations of its foundation models. Nevertheless, most baselines also use Wan backbones, providing a relatively fair comparison to evaluate personalization performance. In the streaming autoregressive setting, we focus on subject preservation and text controllability rather than real-time performance. We have not incorporated acceleration techniques such as few-step distillation. Future work will explore these techniques to reduce streaming latency while maintaining generation quality and subject consistency.

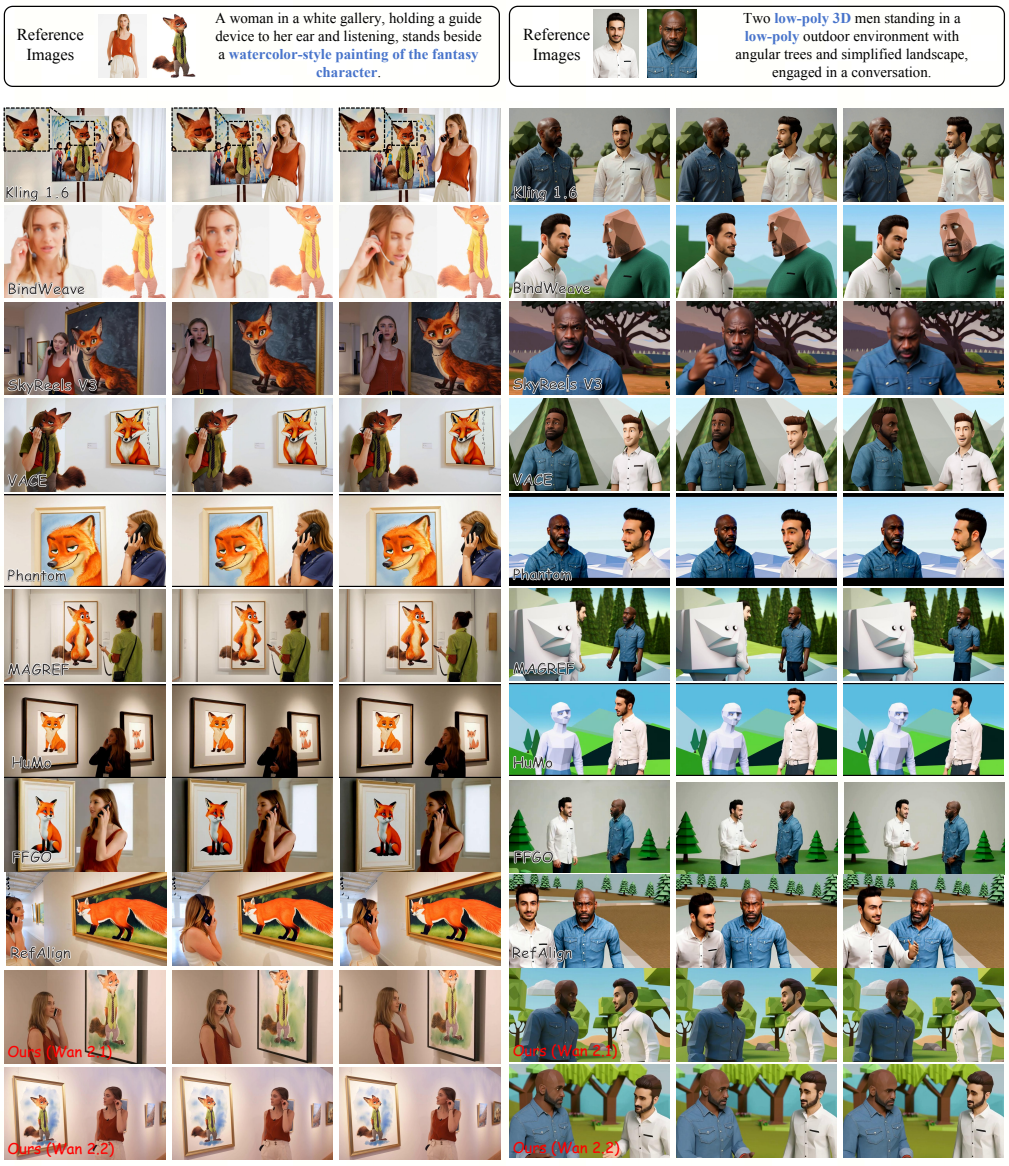



Figure 14: **Additional qualitative comparison** with existing methods. For real–fantasy subject interactions, *DomainShuttle* successfully generates interactions between the woman and the painting of the fantasy character. Among existing methods, the commercial model Kling-1.6 performs comparatively well, but the painted character blinks, which contradicts the static nature of the painting. For real-to-fantasy scenarios, our method accurately maps real-world subjects to the low-poly 3D domain. In contrast, existing methods either fail to transfer subjects or lose key features after transfer. Please watch the video via an anonymous website for the best experience.

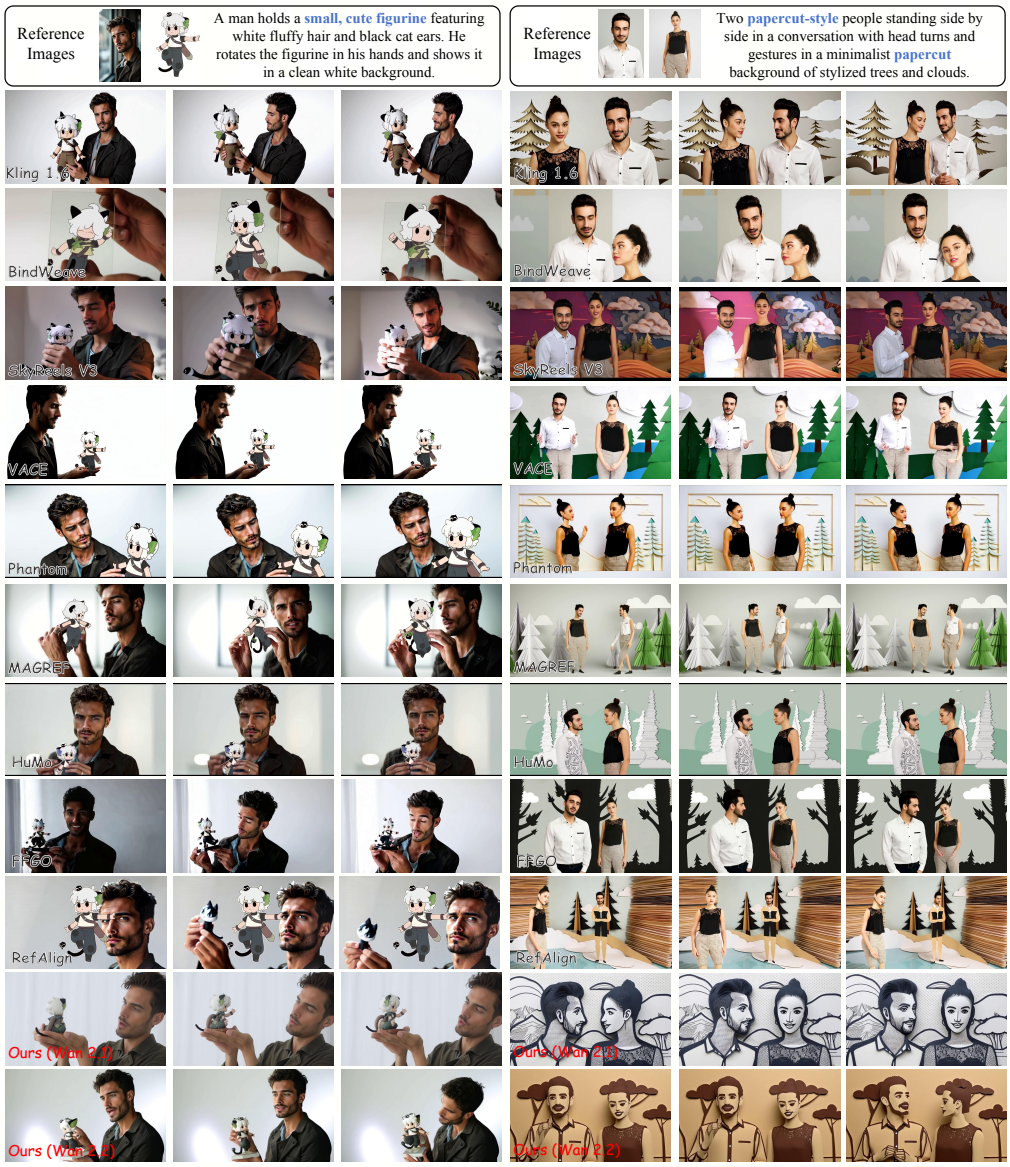



Figure 15: **Additional qualitative comparison** with existing methods. For mapping fantasy-domain subjects to real-world subjects, *DomainShuttle* successfully converts the fantasy character into a real-world small figurine, outperforming existing methods. For real-to-fantasy, *DomainShuttle* transforms real-world subjects into the corresponding paper-cut fantasy domain, whereas existing methods fail to achieve this conversion. Please watch the video via an anonymous website for the best experience.

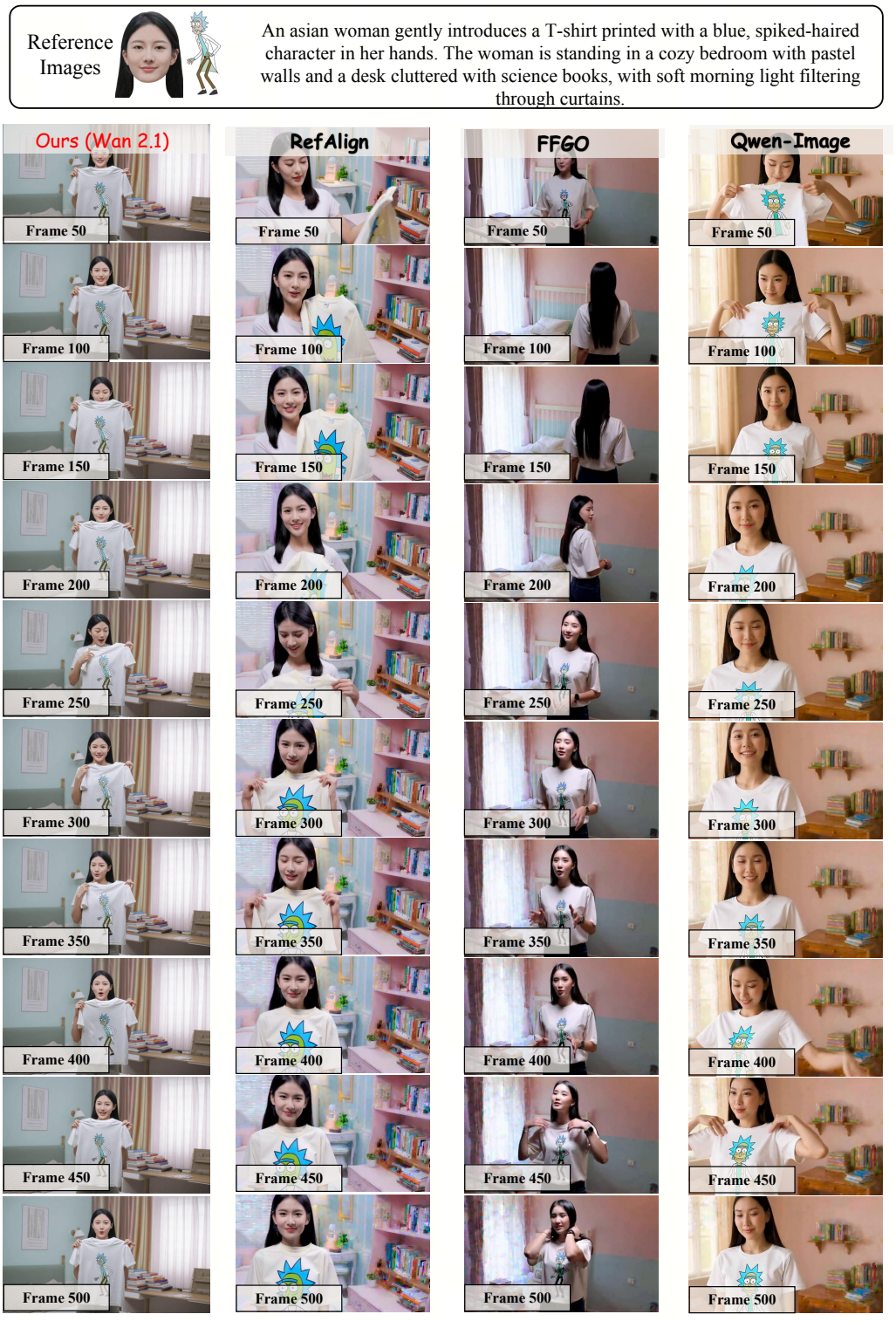



Figure 16: **Additional qualitative comparison** for streaming-based methods. *DomainShuttle* shows strong performance in streaming autoregression scenarios. Please watch the video via an anonymous website for the best experience.