CUSTOM FORCING: TRAINING-FREE SUBJECT CUSTOMIZATION FOR AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

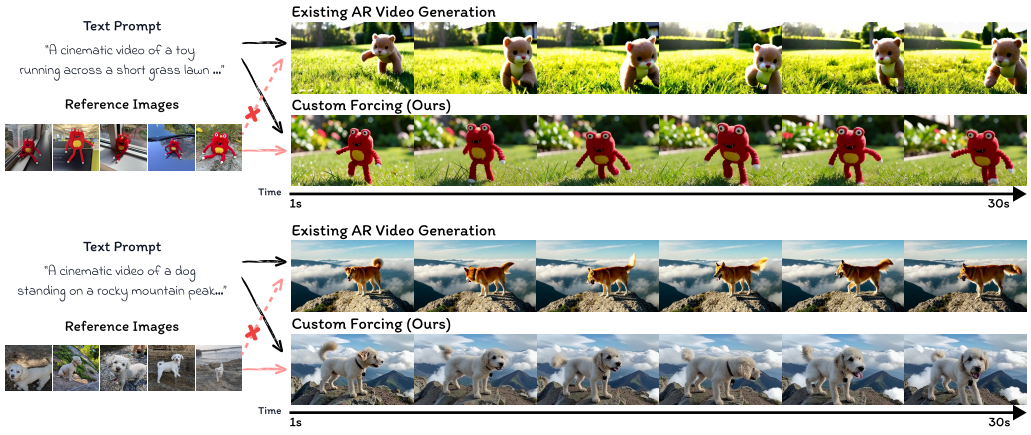


Figure 1: **Comparison with existing autoregressive video generation.** For each text prompt, an existing autoregressive video model (top) generates a generic instance of the target category, whereas Custom Forcing (bottom) also conditions on a few reference images and preserves the identity of the specific subject throughout the video without any training.

ABSTRACT

Autoregressive video models can generate minute-long videos in real time, but they produce generic subjects from text rather than specific subjects from user-provided images. Existing customization methods either require costly per-subject optimization or use pretrained conditioning networks that jointly process all video frames with bidirectional attention. Neither approach is designed for causal streaming. We present **Custom Forcing**, a training-free method that stores reference-based anchor frames in the persistent KV cache of a frozen autoregressive video model. However, fixed anchors face two limitations: simple conditioning allows identity to drift, and the text prompt continues to favor a generic subject. To address these problems, *drift-adaptive value amplification* (DVA) scales reference influence with the degree of identity drift, while *anchor contrast guidance* (ACG) steers generation away from the generic class prior. Over two-minute rollouts, fixed anchors fall from 0.58 to 0.42 in DINO-I, while Custom Forcing keeps it between 0.58 and 0.62 without reducing motion. Custom Forcing also achieves higher subject similarity than bidirectional customization methods and better preserves identity over 30s than causal image-to-video and reference-to-video models, while generating each frame 9.5–$28.5\times$ faster than these long-video baselines.

1 INTRODUCTION

Video generation models can now turn a text prompt or an input image into a high-quality video, and they are increasingly used for world models, robotics, and content creation (Brooks et al., 2024; Wan et al., 2025; Xu et al., 2026b; Yang et al., 2026b; Pondaven et al., 2026). Bidirectional video generation models (Brooks et al., 2024; Wan et al., 2025) denoise all frames jointly at every sampling step,

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.

so every step processes the entire video and generation is slow. Autoregressive video models instead generate frames or chunks in temporal order and condition each step on previous outputs (Chen et al., 2024; Teng et al., 2025). With distillation or adversarial post-training, they need only a few denoising steps per chunk and can stream video in real time (Yin et al., 2025; Huang et al., 2026; Lin et al., 2026), and rolling contexts bound their memory as the video grows (Liu et al., 2026b; Yang et al., 2025; Cui et al., 2026; Zhang et al., 2026). These properties support real-time and long-horizon applications such as interactive video generation and world models (Yang et al., 2025; Xu et al., 2026b; Yang et al., 2026b; Pondaven et al., 2026). In such applications, users often want a specific subject, such as their own pet or toy, to appear throughout the video. Given only a text prompt, however, a conventional autoregressive video model generates a generic instance of the category.

Image customization methods either optimize the model for each subject (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023) or use pretrained subject encoders or adapters that generalize to unseen subjects (Wei et al., 2023; Ye et al., 2023; Li et al., 2024; Wang et al., 2024). Video customization follows the same two approaches (Ma et al., 2024; Wei et al., 2024; Wu et al., 2025; Jiang et al., 2024; Yuan et al., 2025; Xue et al., 2026), but optimizing a video model for each subject is costly, and training with still images makes it difficult to preserve motion (Chefer et al., 2024; Wu et al., 2025). These methods also denoise the frames of a bounded video jointly, which a causal generator cannot do: its finalized frames cannot be revised, and the reference must remain available throughout an open-ended rollout. A simpler alternative gives an image of the subject to an image-to-video model as its first frame. As Sec. 4.2 shows, however, causal image-to-video models lose the subject within 30 s, and a bidirectional one cannot generate long videos.

In streaming generation, once a chunk is finalized, appearance errors propagate through subsequent chunks and cause progressive divergence from the external subject reference. We call this *identity drift*. Existing long-video methods maintain consistency with generated history using attention sinks (Liu et al., 2026b; Li et al., 2026b; Yi et al., 2025; Ye et al., 2026; Zhao et al., 2026b), persistent memory (Yang et al., 2026a; Kim et al., 2026; Yesiltepe et al., 2026; Dalva & Yanardag, 2026), or longer contexts (Cui et al., 2026; Yang et al., 2025; Chen et al., 2026b; Mao et al., 2026). However, consistency with generated history does not ensure fidelity to the user-provided subject. We store the user's reference images and a scene-specific *custom anchor* as persistent anchor frames in the KV cache, making them available to every generated chunk without fine-tuning or an additional conditioning network. These anchors establish the subject initially but cannot prevent identity drift: as shown by the red curve in Fig. 2, identity similarity steadily declines over a two-minute rollout. Weak anchor influence loses subject-specific details, and strong influence suppresses motion. The text prompt also biases generation toward a generic instance of the subject category.

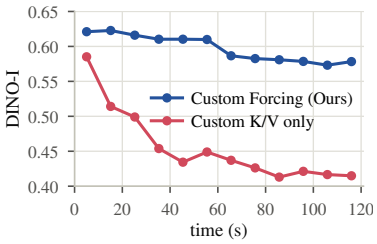


Figure 2: **DINO-I over two minutes**, $n = 100$. Custom K/V only uses fixed anchor frames without DVA or ACG.

We present **Custom Forcing**, which addresses these limitations by modifying only the backbone's self-attention. *Drift-adaptive value amplification* (DVA) scales reference influence with the measured drift of each finalized chunk. *Anchor contrast guidance* (ACG) contrasts attention outputs computed with and without the anchor frames and extrapolates along their difference, counteracting the generic class prior induced by the prompt.

Over two-minute rollouts, Custom Forcing maintains DINO-I between 0.58 and 0.62 (blue curve), while fixed anchors decline from 0.58 to 0.42 (red curve, Fig. 2). The improvement grows over time, reaching +0.163 DINO-I in the final window without reducing motion. Custom Forcing also better preserves subject identity than causal image-to-video baselines over 30 s, with only 11–13% overhead over fixed anchors.

Our contributions are as follows:

- We present Custom Forcing, a training-free method that customizes a frozen autoregressive video model by keeping the reference images and a custom anchor in its KV cache as anchor frames, without fine-tuning or an additional conditioning network.
- We introduce DVA and ACG, two self-attention operations that keep the subject over long rollouts. DVA scales the reference influence with the measured drift of finalized chunks,

and ACG extrapolates away from the attention output computed without the anchors to counter the generic class prior.

- We show that Custom Forcing keeps DINO-I between 0.58 and 0.62 over two-minute videos, where fixed anchors fall to 0.42, with comparable motion. It exceeds bidirectional customization methods and causal image-to-video models, and matches a 14B reference-to-video model with a 1.3B backbone while generating each frame 9.5–$28.5\times$ faster.

We review related work in Appendix A.

2 Preliminaries

2.1 Streaming Autoregressive Video Diffusion

Let a video be represented in the latent space of a causal VAE and divided into chunks $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \dots)$, where each chunk $\mathbf{x}^t$ contains F latent frames of $L = h \times w$ tokens. An autoregressive video diffusion model factorizes the distribution of the video over chunks,

$$p_\theta(\mathbf{x} \mid y) \;=\; \prod_{t \geq 1} p_\theta\big(\mathbf{x}^t \mid \mathbf{x}^{<t}, y\big)\,, \tag{1}$$

where y is the text prompt, and samples each factor with a diffusion model. Because real-time generation leaves room for only a few network evaluations per chunk, recent models distill the denoiser into a few-step generator G_θ (Yin et al., 2025; Huang et al., 2026). Starting from $\mathbf{x}^t_K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, chunk t is obtained by

$$\mathbf{x}^t_{k-1} \;=\; \Psi\big(G_\theta\big(\mathbf{x}^t_k, \sigma_k;\, y, \mathcal{C}_{t-1}\big),\, \sigma_{k-1}\big)\,, \qquad k = K, \dots, 1, \tag{2}$$

where G_θ predicts the clean chunk, $\Psi(\cdot, \sigma)$ re-noises its input to level σ with $\sigma_0 = 0$, and $\mathbf{x}^t := \mathbf{x}^t_0$. The past enters only through $\mathcal{C}_{t-1}$, the KV cache that stores, for every transformer block, the keys and values of the frames finalized so far. Once $\mathbf{x}^t$ is complete, a forward pass on the clean chunk writes its keys and values into the cache, and $\mathbf{x}^t$ is never revised afterwards.

Inside each DiT block, a token of the chunk being denoised forms a query $\mathbf{q} \in \mathbb{R}^d$ and attends to the cached tokens together with the tokens of its own chunk. For a set $\mathcal{J}$ of token indices, we write

$$\mathrm{Attn}(\mathbf{q}; \mathcal{J}) \;=\; \sum_{j \in \mathcal{J}} \frac{\exp\!\left(\mathbf{q}^\top \mathbf{k}_j / \sqrt{d}\right)}{\sum_{j' \in \mathcal{J}} \exp\!\left(\mathbf{q}^\top \mathbf{k}_{j'} / \sqrt{d}\right)}\, \mathbf{v}_j, \tag{3}$$

so that standard self-attention is $\mathrm{Attn}(\mathbf{q}; \mathcal{J}_\mathcal{C} \cup \mathcal{J}_{\mathrm{cur}})$, with $\mathcal{J}_\mathcal{C}$ indexing the cached tokens and $\mathcal{J}_{\mathrm{cur}}$ the tokens of the current chunk.

The cache cannot grow with the video, so streaming models bound it. A common design (Liu et al., 2026b; Li et al., 2026b; Yi et al., 2025) splits it as $\mathcal{C}_t = \mathcal{S} \cup \mathcal{W}_t$: an attention sink $\mathcal{S}$ that holds the first few frames and is kept for the entire rollout, and a window $\mathcal{W}_t$ of recent frames from which older frames are evicted. The sink is therefore the only part of the context that every chunk can see, and in the standard setting it holds the first frames that the model generates.

2.2 Subject Customization in Autoregressive Diffusion

We are given a small set of reference images $\{c_i\}_{i \in \mathcal{R}}$ of a particular subject and a prompt y that describes a scene and an action but names the subject only by its category (e.g., "a dog trotting down a forest trail"). The goal is to generate a video that follows y and in which the subject matches the reference images throughout the rollout. The generator G_θ stays frozen, and nothing is optimized for the subject.

Let $s(t)$ denote the similarity between the subject in chunk $\mathbf{x}^t$ and the reference images, which we define concretely in Eq. (4). Identity drift (Sec. 1) is the decline of $s(t)$ over the rollout. It differs from temporal consistency, which compares $\mathbf{x}^t$ with the preceding chunks $\mathbf{x}^{<t}$, because a video can remain smooth and self-consistent while its subject slowly turns into a generic member of the category. Any correction must therefore be made online, before the next chunk is generated.

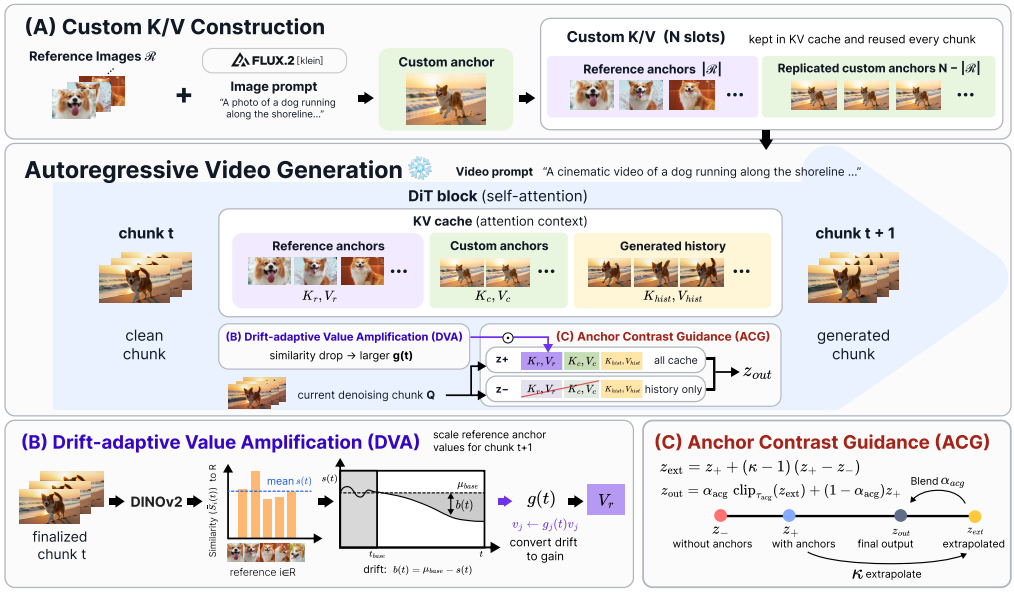


Figure 3: **Overview of CUSTOM FORCING. Top:** the reference images and copies of a custom anchor generated by FLUX.2 [klein] (Black Forest Labs, 2025) are written into the KV cache once as the custom K/V. **Middle:** in every self-attention layer of the frozen backbone, the current chunk reads the cache with and without the anchor frames, giving z_+ and z_-, after DVA scales the reference values by $g(t)$. **Bottom left:** DVA sets $g(t)$ from how far the similarity $s(t)$ of each finalized chunk falls below its initial level μ_{base}. **Bottom right:** ACG extrapolates away from z_- by κ and blends the result with z_+.

3 CUSTOM FORCING: SUBJECT CUSTOMIZATION IN THE KV CACHE

3.1 OVERVIEW

Existing video customization methods introduce a subject through per-subject optimization or an additional conditioning network (see Appendix A). A autoregressive video model may need neither, because every chunk already reads the KV cache and part of the cache persists for the entire rollout. We hypothesize that images of the subject written into this persistent part can serve as its reference throughout generation if the model reads them with the right strength and in the right direction.

CUSTOM FORCING has three components (Fig. 3). Custom K/V construction (Sec. 3.2) writes the reference images and one custom anchor of the subject into the cache before generation. Drift-adaptive value amplification (DVA, Sec. 3.3) sets how strongly the reference images are read, and anchor contrast guidance (ACG, Sec. 3.4) sets the direction away from the generic subject. DVA rescales the values before attention is computed, and ACG modifies the attention output.

3.2 CUSTOM K/V CONSTRUCTION

Before the first chunk is generated, we encode N images of the subject with the VAE, one latent frame per image, and pass them through G_θ as clean frames, which writes their keys and values into the cache of every block. We call these frames the *custom K/V* and index them by $\mathcal{A} = \{1, \ldots, N\}$. The custom K/V occupies the attention sink of Sec. 2.1 in place of the first generated frames. It precedes all generated frames and is never evicted, so every chunk reads it through the ordinary attention of Eq. (3), without a retrieval step or an additional conditioning branch.

Reference anchors specify the subject, and the custom anchor specifies the opening scene. The custom K/V contains two kinds of frames. The *reference anchors*, at positions $\mathcal{R} \subset \mathcal{A}$, are the user's reference images used as provided; they determine what the subject looks like. The remaining positions $\mathcal{A} \setminus \mathcal{R}$ hold the *custom anchor*, a single image of the subject placed in the scene described

by y. We generate it once with FLUX.2 [klein] (Black Forest Labs, 2025) from the reference images and a short image prompt derived from y (Appendix B). Because the video begins from this image, we replicate it in all $N - |\mathcal{R}|$ remaining positions and place it last, directly before the first chunk.

More copies of the custom anchor keep the opening scene stable but fix the subject to its pose, whereas more reference anchors restore motion at the cost of scene stability (Appendix F.2).

The custom K/V alone customizes the video but cannot prevent drift. Writing the custom K/V into the cache and changing nothing else already produces a customized video, which we call *Custom K/V only*. It cannot prevent identity drift, because it reads the anchors with a fixed strength and does nothing against the bias of the prompt toward a generic subject. Sec. 3.3 and 3.4 address these two limitations.

3.3 Drift-Adaptive Value Amplification

A fixed reference strength cannot serve the whole rollout. Reading the anchor frames more strongly reduces drift, but doing so throughout the video enlarges the subject and reduces motion (Appendix F.4). We therefore scale the reference anchors with the measured drift.

Drift can be measured online from finalized chunks. We decode each finalized chunk into F' pixel frames $\{x_f(t)\}_{f=1}^{F'}$, where $F' > F$ because the VAE compresses time, and embed each frame with a frozen DINOv2-B encoder ϕ (Oquab et al., 2024), using the ℓ_2-normalized CLS token. The similarity of chunk t to reference anchor c_i and its average over the reference anchors are

$$\bar{S}_i(t) \;=\; \frac{1}{F'} \sum_{f=1}^{F'} \big\langle \phi(x_f(t)),\, \phi(c_i) \big\rangle, \quad i \in \mathcal{R}, \qquad s(t) \;=\; \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \bar{S}_i(t). \tag{4}$$

We exclude the custom anchor, because the video starts from it.

The absolute value of $s(t)$ varies across subjects and prompts, so we compare it with a level estimated from the same video. The first t_{base} chunks keep the values unchanged, and their mean similarity defines

$$\mu_{\text{base}} \;=\; \frac{1}{t_{\text{base}}} \sum_{t=1}^{t_{\text{base}}} s(t). \tag{5}$$

From then on, drift is the amount by which $s(t)$ falls below this level, clipped to $[0, b_{\max}]$:

$$b(t) \;=\; \min\Big(\max\big(0,\; \mu_{\text{base}} - s(t)\big),\; b_{\max} \Big). \tag{6}$$

Amplification is restricted to the subject in each reference image. The drift is converted into an amplification of the values of the reference anchors, while the custom anchor, whose background is the opening scene, is left unchanged. Only the subject in a reference image carries identity, so we compute a subject mask for each reference anchor with SAM (Kirillov et al., 2023) and average it within each latent token. Each token j then carries a weight $w_j \in [0, 1]$, the fraction of its pixels that belong to the subject. Tokens with $w_j \le \tau_{\text{subj}}$ are treated as background and removed from attention, and the remaining tokens of reference anchor i form the set $\mathcal{T}(i)$ with $n_i = |\mathcal{T}(i)|$. For $j \in \mathcal{T}(i)$ and $i \in \mathcal{R}$, the target amplification is

$$\widetilde{g}_j(t) \;=\; \frac{\alpha_{\text{inj}}\, b(t)}{|\mathcal{R}|} \cdot \frac{n_i\, w_j}{\sum_{j' \in \mathcal{T}(i)} w_{j'}}. \tag{7}$$

The first factor divides a total budget $\alpha_{\text{inj}}\, b(t)$ evenly among the reference anchors, and the second has mean one over $\mathcal{T}(i)$, so it moves the share of each anchor toward the tokens that are mostly subject (Appendix B).

Since $b(t)$ can change abruptly between chunks, we smooth the amplification with an exponential moving average,

$$g_j(t) \;=\; \lambda\, g_j(t-1) + (1-\lambda)\, \widetilde{g}_j(t), \tag{8}$$

and read the values of these tokens as $g_j(t)\, \mathbf{v}_j$ for $j \in \bigcup_{i \in \mathcal{R}} \mathcal{T}(i)$, leaving the cached $\mathbf{v}_j$ unchanged. The amplification computed after chunk t is finalized takes effect from chunk $t+1$ onward. We set $g_j(t) = 1$ for $t < t_{\text{base}}$, as in Custom K/V only, after which g_j falls below 1 for small drift and rises above 1 as it grows.

3.4 Anchor Contrast Guidance

The prompt biases the subject toward a generic member of its category. DVA determines how strongly the anchor frames are read, but not what the subject is drifting toward. The prompt names only the category of the subject, so the backbone tends to produce a typical member of that category, which we call the *class prior*. The subject of Custom K/V only drifts toward such generic instances over time (Fig. 6). Counteracting the class prior requires a direction in feature space, and a second read of the same cache provides one.

Masking the anchors yields a negative branch from the same cache. Let $\mathcal{J}_{\mathcal{A}}$ index the anchor tokens and $\mathcal{J}_{\text{gen}}$ all other tokens a query can attend to, namely the retained generated frames and the chunk being denoised. Since the two sets are disjoint, the same query can be evaluated with and without the anchors:

$$z_+ \;=\; \mathrm{Attn}(\mathbf{q};\, \mathcal{J}_{\mathcal{A}} \cup \mathcal{J}_{\text{gen}}), \qquad z_- \;=\; \mathrm{Attn}(\mathbf{q};\, \mathcal{J}_{\text{gen}}). \tag{9}$$

The output z_- still reflects the anchor frames indirectly, because the generated frames in $\mathcal{J}_{\text{gen}}$ were produced under their influence. The two outputs differ only in whether this query attends to the anchor frames at the current step.

Since the softmax in Eq. (3) factorizes over disjoint index sets (Fan et al., 2024), z_+ is a convex combination of the two parts:

$$z_+ \;=\; m\, z_{\mathcal{A}} + (1-m)\, z_-, \qquad m \;=\; \frac{M_{\mathcal{A}}}{M_{\mathcal{A}} + M_{\text{gen}}} \in [0,1], \tag{10}$$

where $z_{\mathcal{A}} = \mathrm{Attn}(\mathbf{q};\, \mathcal{J}_{\mathcal{A}})$ and $M_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \exp(\mathbf{q}^\top \mathbf{k}_j / \sqrt{d})$. Hence $z_+ - z_- = m\,(z_{\mathcal{A}} - z_-)$. This difference points from the anchor-free output toward what the anchors contribute and is scaled by m, the share of attention that the query places on the anchors.

Extrapolating along this difference strengthens the anchors against the class prior. Following NAG (Chen et al., 2026a), we extrapolate with strength κ,

$$z_{\text{ext}} \;=\; z_+ + (\kappa - 1)\,(z_+ - z_-), \tag{11}$$

which has no effect for $\kappa = 1$. To keep the extrapolated feature within a reasonable range, we bound its ℓ_1 norm relative to z_+ and blend the result with z_+:

$$z_{\text{out}} \;=\; \alpha_{\text{acg}}\, \mathrm{clip}_{\tau_{\text{acg}}}(z_{\text{ext}}) + (1 - \alpha_{\text{acg}})\, z_+, \qquad \mathrm{clip}_\tau(z) \;=\; \min\!\left(1,\, \tau\, \frac{\|z_+\|_1}{\|z\|_1}\right) z. \tag{12}$$

When the clip is inactive, substituting Eq. (10) into Eqs. (11) and (12) gives

$$z_{\text{out}} \;=\; m'\, z_{\mathcal{A}} + (1 - m')\, z_-, \qquad m' \;=\; \big(1 + \alpha_{\text{acg}}(\kappa - 1)\big)\, m. \tag{13}$$

ACG therefore multiplies the attention share that each query already places on the anchor frames, while the clip bounds the norm of z_{ext}. Unlike RefDrop (Fan et al., 2024), which replaces this query-dependent share with a constant, ACG scales it, so queries that already read the subject from the anchors read it more strongly. Unlike NAG (Chen et al., 2026a), which contrasts a positive and a negative prompt in cross-attention, ACG obtains its negative branch by masking the anchor frames in self-attention. Neither branch requires an additional forward pass of the network (Appendix A.3).

ACG is applied to the self-attention of every DiT block at every denoising step, and cross-attention to the text is left unchanged. Like DVA, it starts after the first t_{base} chunks, with a linear warm-up (Appendix B). DVA acts on the next chunk, whereas ACG acts on the current one.

4 Experiments

4.1 Experimental Setup

Implementation details. We apply Custom Forcing to Rolling Forcing (Liu et al., 2026b), a streaming autoregressive video diffusion model built on Wan2.1-T2V-1.3B (Wan et al., 2025), without fine-tuning. We use 10 DreamBooth subjects (Ruiz et al., 2023), six animals and four

Method	Length	Long video	Training free	DINO-I over time ↑ all	first	middle	last	Alignment ↑ CLIP-I	CLIP-T	Video quality Subj. Cons. ↑	Bg. Cons. ↑	Dyn. Degree
Bidirectional												
VACE-1.3B	5.1 s	✗	✓	0.333	–	–	–	0.703	0.296	0.921	0.951	0.46
SMRABooth	3.3 s	✗	✗	0.407	–	–	–	0.664	0.309	**0.978**	0.969	0.21
CustomCrafter	2.0 s	✗	✗	0.490	–	–	–	0.669	0.301	0.966	**0.971**	0.17
MotionBooth	3.0 s	✗	✗	0.435	–	–	–	0.642	0.293	0.907	0.928	0.76
Autoregressive												
Vanilla AR model	30 s	✓	–	0.163	0.163	0.160	0.155	0.681	**0.360**	0.924	0.943	0.36
Vanilla AR + LoRA	30 s	✓	✗	0.435	0.418	0.432	0.413	0.642	0.331	0.936	0.948	0.28
Custom K/V only	30 s	✓	✓	0.541	0.579	0.506	0.475	0.791	0.355	0.913	0.939	0.41
CUSTOM FORCING (Ours)	30 s	✓	✓	**0.635**	**0.619**	**0.626**	**0.616**	**0.803**	0.353	0.928	0.945	0.41
Custom K/V only	2 min	✓	✓	0.501	0.583	0.450	0.415	0.785	**0.357**	0.882	0.925	0.47
CUSTOM FORCING (Ours)	2 min	✓	✓	**0.623**	**0.622**	**0.602**	**0.578**	**0.798**	0.352	**0.909**	**0.936**	0.45

Table 1: **Main results.** *Long video* marks whether a method can extend generation beyond a fixed length, and *Training free* whether it customizes the subject without training any model on it. Bold marks the best value in each column among the bidirectional methods and the 30 s rows, and the better of the two 2 min rows. Dynamic Degree is not bolded. Subject and Background Consistency favor the short bidirectional videos, which change little.

objects, with 10 prompts each Appendix H that name only the category of the subject, and generate one 30 s and one 2 min video per prompt. The custom K/V holds five reference anchors and four copies of one custom anchor, which FLUX.2 [klein] 4B (Black Forest Labs, 2025) generates for each prompt. We use $\alpha_{\text{inj}} = 300$ and $\kappa = 4$ for every subject and prompt (more details are in Appendix B).

Baselines. We compare with the bidirectional customization methods SMRABooth (Xu et al., 2026c), CustomCrafter (Wu et al., 2025), MotionBooth (Wu et al., 2024a), and the training-free VACE-1.3B (Jiang et al., 2025) at their native lengths of 2–5.1 s. We also compare with the image-to-video models FramePack (Zhang et al., 2026), FramePack-F1, and SkyReels-V2 DF (Chen et al., 2025a), which start from the custom anchor, and the reference-to-video model SkyReels-V3 (Li et al., 2026a). Finally, we report the *vanilla AR model* with the prompt alone, *vanilla AR + LoRA* fine-tuned on each subject, and *Custom K/V only* without DVA or ACG (Appendix B).

Evaluation metrics. DINO-I and CLIP-I measure similarity to the reference images, and CLIP-T measures agreement with the prompt. To expose drift, we also compute DINO-I on the first, middle, and last thirds of each 30 s video and on windows of the same duration at the start, middle, and end of each 2 min video. From VBench (Huang et al., 2024), we report Subject and Background Consistency and Dynamic Degree. All scores are means over the 100 prompts (see Appendix C).

4.2 MAIN RESULTS

Comparison with customization methods. Tab. 1 gives the main results. The bidirectional methods, including the training-free VACE-1.3B, stay below Custom K/V only in DINO-I (0.33–0.49), and fine-tuning a LoRA of the backbone raises DINO-I from 0.163 only to 0.435, below the 0.541 of Custom K/V only. Custom K/V only drifts from 0.579 to 0.475 within 30 s and from 0.583 to 0.415 over 2 min. CUSTOM FORCING stays nearly flat over 30 s (0.619 to 0.616) and declines only to 0.578 over 2 min, with Dynamic Degree at the level of Custom K/V only (Fig. 4 and Appendix D.1).

Comparison with image- and reference-to-video models. Tab. 2 compares CUSTOM FORCING with models that start from the custom anchor or the reference images, a two-stage route that prior customization work also uses as a baseline (Zhou et al., 2024a; Yuan et al., 2025). With a 1.3B backbone, CUSTOM FORCING matches the 14B SkyReels-V3 in whole-video DINO-I (0.635 against 0.641) and exceeds it in the last window (0.616 against 0.564), while generating each frame 9.4–28.5× faster than all four models. FramePack does not decline but scores lower overall and must fix the video length before generation, and the causal FramePack-F1 and SkyReels-V2 DF

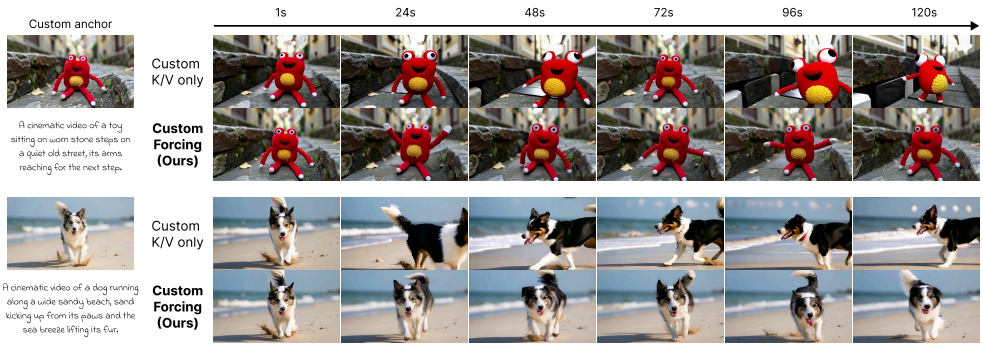


Figure 4: **Two-minute generation.** Custom K/V only drifts within the first minute, while CUSTOM FORCING keeps both subjects for two minutes.

				DINO-I over time ↑				Alignment ↑	
Method	Backbone	Time	Causal	all	first	middle	last	CLIP-I	CLIP-T
FramePack	HunyuanVideo 13B	9.4×	✗	0.625	0.616	0.620	**0.623**	0.795	0.357
FramePack-F1	HunyuanVideo 13B	9.5×	✓	0.477	0.534	0.418	0.392	0.773	0.355
SkyReels-V2 DF	Wan2.1 1.3B	11.5×	✓	0.449	0.552	0.393	0.307	0.767	**0.361**
SkyReels-V3	Wan2.1 14B	28.5×	✓	**0.641**	**0.693**	0.620	0.564	0.794	0.347
CUSTOM FORCING (Ours)	Wan2.1 1.3B	1.0×	✓	0.635	0.619	**0.626**	0.616	**0.803**	0.353

Table 2: **Comparison with image- and reference-to-video models**, 30 s, $n = 100$, identity and prompt alignment only. *Backbone* gives the architecture and size of each video diffusion transformer, *Time* is the generation time per frame relative to CUSTOM FORCING on one H100 GPU, and *Causal* marks whether each part of the video is final before the next is generated. Bold marks the best score and underline the second best.

drift within 30 s (Fig. 5). A user study with 30 participants prefers CUSTOM FORCING over every other method it includes except SkyReels-V3, against which the difference is not significant (see Appendix D.7).

Inference cost. On one H100 GPU, CUSTOM FORCING takes 1.13× the generation time of Custom K/V only at 30 s and 1.11× at 2 min. Most of the overhead comes from measuring drift, and ACG alone adds 4%.

4.3 ABLATION STUDY

Effect of DVA and ACG. Tab. 3 and Fig. 6 add the components one at a time. Adding DVA stops the drift, keeping DINO-I near 0.60 from the first window to the last, and adding ACG raises every window by a further 0.02–0.03. In Fig. 6, Custom K/V only turns both subjects toward generic instances of their categories by 21 s. DVA delays this change but still alters the face and shoulders of the toy, and adding ACG keeps both subjects as in the reference images (see Appendix E).

			DINO-I ↑		
	DVA	ACG	first	mid.	last
Custom K/V only	✗	✗	0.579	0.506	0.475
+ DVA	✓	✗	0.600	0.598	0.598
+ DVA + ACG (Ours)	✓	✓	**0.619**	**0.626**	**0.616**

Table 3: **Component ablation**, 30 s.

Further analyses. Replacing the DINOv2-B drift encoder of DVA with ConvNeXt-B (Liu et al., 2022) changes the scores by at most 0.003, so the improvement does not come from sharing the encoder with the metric (Appendix D.4). ACG lowers the similarity to the vanilla video of the same prompt, a measure of the class prior, by 0.015–0.040 in every third of the video (Appendix D.6). DVA stops the decline of DINO-I after about 10 s, while ACG alone raises the starting identity but still declines, so the two address different failures (Appendix D.5). Participants show no significant preference in quality between CUSTOM FORCING and Custom K/V only (45.8%, Appendix D.7). Appendix F ablates the remaining design choices.



Figure 5: **Qualitative comparison with image- and reference-to-video models.** The image-to-video models start from the custom anchor, and SkyReels-V3 starts from the reference images, so its scenes differ. CUSTOM FORCING keeps both subjects to the end.

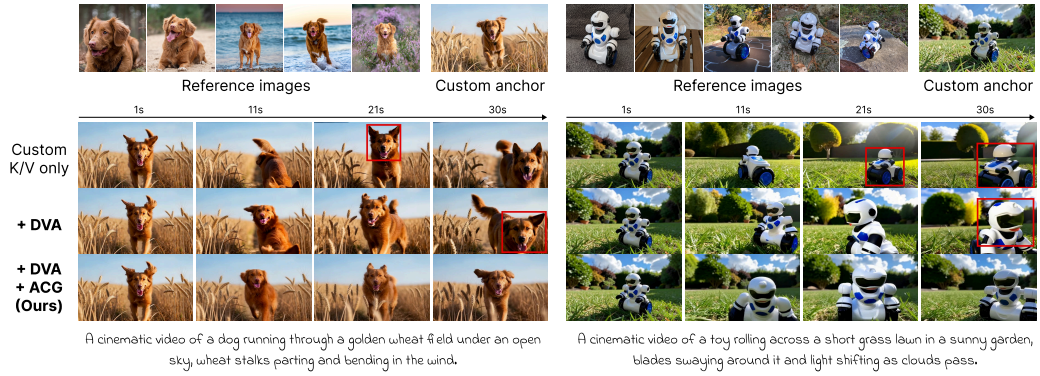


Figure 6: **Adding the components one at a time.** Red boxes mark where the subject departs from the reference images.

5 CONCLUSION

In this paper, we propose CUSTOM FORCING, a training-free method that customizes autoregressive video generation with images of a user-provided subject through the KV cache. It needs neither per-subject optimization nor an additional conditioning network, only five reference images and one generated image of the subject in the scene, and it adds 11–13% to the generation time of fixed anchors. The custom K/V, written into the cache once, establishes the identity from the first chunk. DVA reads the reference images more strongly the further the subject drifts, and ACG extrapolates away from the attention output computed without the anchor frames, steering generation away from the generic subject favored by the prompt. Without modifying the backbone, CUSTOM FORCING keeps the subject close to the reference images for two minutes without reducing motion, holding DINO-I between 0.58 and 0.62 while fixed anchors fall to 0.42. Because it changes only what the KV cache holds and how it is read, it also applies to Deep Forcing, another autoregressive model with an attention sink (Appendix D.2). Appendix G discusses its limitations. A direction for future work is to keep a user's subject in interactive world models, which also generate video causally over long horizons.

AI USE STATEMENT

In this work, we used generative AI tools from Anthropic and OpenAI solely for language-related assistance, including proofreading, rephrasing, and checking grammar, readability, and formatting. Generative AI was not used to formulate the research hypothesis, design or implement the method or experiments, analyze or interpret the results, select qualitative examples, create figures, or make research decisions. All AI-assisted edits were reviewed and approved by the authors. The authors take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The user study in Appendix D.7 asked 30 participants to compare pairs of generated videos of animals and objects. CUSTOM FORCING places a specific subject from a few images into generated videos without training. Applied to images of real people without their consent, this ability could produce misleading or harmful videos. Our experiments use only the animals and objects of the public DreamBooth dataset (Ruiz et al., 2023), and deployments of CUSTOM FORCING should obtain consent for the depicted subject and label generated videos as synthetic.

REPRODUCIBILITY STATEMENT

Sec. 3 specifies the full procedure of CUSTOM FORCING, including custom K/V construction, drift-adaptive value amplification (DVA), and anchor contrast guidance (ACG), and Table 4 reports the hyperparameters used across all subjects and prompts. Appendix B provides the backbone and input settings, custom-anchor generation procedure, implementation details, and the training and sampling configurations of the baselines. Appendix C describes the evaluation metrics and their computation, while Appendix D.7 specifies the protocols for the human study and VLM-based evaluation. The backbone models, image generator used for the custom anchor, evaluation encoders, and DreamBooth subjects used in our experiments are publicly available. We will release the implementation of CUSTOM FORCING, the full set of 100 evaluation prompts, and the custom anchors used in our experiments to facilitate reproduction of the reported results.

REFERENCES

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

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22503–22513. IEEE, 2023.

Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. Still-moving: Customized video generation without customized video data. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. *Advances in Neural Information Processing Systems*, 38:137412–137446, 2026a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6099–6110. IEEE, 2025b.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Yusuf Dalva and Pinar Yanardag. Adastate: Self-evolving anchors for streaming video generation. *arXiv preprint arXiv:2605.30349*, 2026.

Jiaojiao Fan, Haotian Xue, Qinsheng Zhang, and Yongxin Chen. Refdrop: Controllable consistency in image or video generation via reference feature guidance. 2024.

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.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. In *European Conference on Computer Vision*, pp. 324–344. Springer, 2026.

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.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 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.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid kv cache compression for efficient autoregressive video diffusion models. *arXiv preprint arXiv:2605.09681*, 2026.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700. IEEE, 2024.

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

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941. IEEE, 2023.

Hyunkoo Lee, Wooseok Jang, Jini Yang, Taehwan Kim, Sangoh Kim, Sangwon Jung, and Seungryong Kim. V-warper: Appearance-consistent video diffusion personalization via value warping. *arXiv preprint arXiv:2512.12375*, 2025.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, Mingshan Chang, Wenjing Cai, 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.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8640–8650. IEEE, 2024.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13146–13156. IEEE, 2025.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *Advances in Neural Information Processing Systems*, 38:41061–41086, 2026.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing narrative long video generation via training-free identity-aware memory. *arXiv preprint arXiv:2605.18733*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026b.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961. IEEE, 2025.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, June 2022.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

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

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Xiaofeng Mao, Shaohao Rui, Kaining Ying, Bo Zheng, Chuanhao Li, Mingmin Chi, and Kaipeng Zhang. Packforcing: Short video training suffices for long video sampling and long context inference. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2026.

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. *Transactions on Machine Learning Research (TMLR)*, 2024.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43727–43739, 2026.

Alexander Pondaven, Ziyi Wu, Igor Gilitschenski, Philip Torr, Sergey Tulyakov, Fabio Pizzati, and Aliaksandr Siarohin. Actionparty: Multi-subject action binding in generative video games. In *European Conference on Computer Vision*, pp. 74–92. Springer, 2026.

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 (ICML)*, 2021.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

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

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *arXiv preprint arXiv:2402.03286*, 2024.

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.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

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*, 2026.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dream video: Composing your dream videos with customized subject and motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549. IEEE, 2024.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15897–15907. IEEE, 2023.

Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Kai Chen, and Yunhai Tong. Motionbooth: Motion-aware customized text-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Tao Wu, Yong Zhang, Xiaodong Cun, Zhongang Qi, Junfu Pu, Huanzhang Dou, Guangcong Zheng, Ying Shan, and Xi Li. Videomaker: Zero-shot customized video generation with the inherent force of video diffusion models. *arXiv preprint arXiv:2412.19645*, 2024b.

Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8469–8477, 2025.

Xunzhi Xiang, Zixuan Duan, Guiyu Zhang, Haiyu Zhang, Zhe Gao, Junta Wu, Shaofeng Zhang, Tengfei Wang, Qi Fan, and Chunchao Guo. Pathwise test-time correction for autoregressive long video generation. *arXiv preprint arXiv:2602.05871*, 2026.

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

Gangwei Xu, Qihang Zhang, Jiaming Zhou, Xing Zhu, Yujun Shen, Xin Yang, and Yinghao Xu. Next forcing: Causal world modeling with multi-chunk prediction. *arXiv preprint arXiv:2606.11187*, 2026b.

Xuancheng Xu, Yaning Li, Sisi You, and Bing-Kun Bao. Smrabooth: Subject and motion representation alignment for customized video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026c.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-in: A lightweight and plug-and-play identity control for video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23314–23324, 2026.

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

Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026a.

Yifei Yang, Zehua Fan, Huan Li, Aoqi Wang, Lida Huang, Haibao Yu, Haiyan Liu, Xuanyao Mao, Jason Bao, Liang Xu, et al. Ar forcing: Towards long-horizon robot navigation world model. *arXiv preprint arXiv:2605.31314*, 2026b.

Bo Ye, Xinyu Cui, Jian Zhao, Tong Wei, and Min-Ling Zhang. Dysink: Dynamic frame sinks for autoregressive long video generation. *arXiv preprint arXiv:2605.21028*, 2026.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

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

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988. IEEE, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026.

Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14464–14474. IEEE, 2025.

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

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

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Nanxuan Zhao, Jing Shi, and Tong Sun. Sugar: Subject-driven video customization in a zero-shot manner. *arXiv preprint arXiv:2412.10533*, 2024a.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024b.

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

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

The Appendix is organized as follows:

- Sec. A: related work.
- Sec. B: the backbone and inputs, hyperparameters, the custom anchor, and the settings of the baselines.
- Sec. C: how the metrics are computed.
- Sec. D: details of the main results, results on Deep Forcing, results per subject, the drift sensor, identity over time, and a user study with VLM judges.
- Sec. E: qualitative observations on the two-minute videos, the image- and reference-to-video models, and the component ablation.
- Sec. F: injection and guidance strength, the composition of the custom K/V, generated images in place of the reference images, and the adaptive amplification and subject mask.
- Sec. G: limitations.

A Related Work

A.1 Autoregressive video generation

Bidirectional video generation models (Brooks et al., 2024; Wan et al., 2025) generate fixed-length videos by denoising the sequence jointly. Autoregressive alternatives produce frames or chunks causally, making it possible to stream outputs and extend generation beyond a fixed video length. Diffusion Forcing introduced per-frame noise levels for causal sequence generation (Chen et al., 2024); CausVid and Self Forcing distilled or post-trained causal few-step generators from pretrained video diffusion models (Yin et al., 2025; Huang et al., 2026); and MAGI-1 scaled chunk-wise autoregressive generation to long contexts (Teng et al., 2025). Later work improves distillation and rollout training (Lin et al., 2026; Zhu et al., 2026a; Zhao et al., 2026a; Lu et al., 2026; Zhu et al., 2026b; Guo et al., 2026), reduces attention and KV-cache cost (Zhang et al., 2026; Lv et al., 2026; Xu et al., 2026a; Ji et al., 2026), or trains causal world models for interactive control (Xu et al., 2026b; Yang et al., 2026b; Pondaven et al., 2026).

Long rollouts accumulate errors beyond the training horizon. Recent methods preserve selected parts of the generated history through rolling attention sinks (Liu et al., 2026b; Li et al., 2026b; Yi et al., 2025; Ye et al., 2026; Zhao et al., 2026b), positional re-indexing, anchor tokens, and evolving memory states (Yang et al., 2026a; Kim et al., 2026; Yesiltepe et al., 2026; Dalva & Yanardag, 2026), longer context windows (Cui et al., 2026; Yang et al., 2025; Chen et al., 2026b; Mao et al., 2026), or explicit spatial memory (Huang et al., 2025a). Bagger and Pathwise Test-Time Correction address rollout error through backward aggregation or test-time correction from an initial frame (Po et al., 2026; Xiang et al., 2026), while IAMFlow tracks identities that have already appeared in an evolving generated narrative (Liu et al., 2026a). These methods improve consistency with generated context. Custom Forcing instead maintains fidelity to user-provided reference images that are external to the rollout.

A.2 Subject-driven video customization

Subject customization has been extensively studied for images. Textual Inversion, DreamBooth, and Custom Diffusion optimize embeddings or model parameters for each subject (Gal et al., 2022; Ruiz et al., 2023; Kumari et al., 2023), while ELITE, IP-Adapter, PhotoMaker, and InstantID learn reusable encoders or adapters for zero-shot conditioning (Wei et al., 2023; Ye et al., 2023; Li et al., 2024; Wang et al., 2024). These approaches establish the two main strategies used by later video methods: subject-specific optimization and pretrained reference conditioning.

Video customization methods based on subject-specific optimization include Magic-Me, DreamVideo, and CustomCrafter (Ma et al., 2024; Wei et al., 2024; Wu et al., 2025). CustomVideo instead combines attention control with subject masks for multi-subject customization (Wang et al., 2026). Still-Moving transfers an image-customized model to video without requiring customized video data (Chefer et al., 2024). A second group uses pretrained or reusable video-conditioning systems for zero-shot or feed-forward customization, including VideoBooth, ID-Animator, ConsisID,

VideoMaker, SUGAR, MagicMirror, and Stand-In (Jiang et al., 2024; He et al., 2024; Yuan et al., 2025; Wu et al., 2024b; Zhou et al., 2024a; Zhang et al., 2025; Xue et al., 2026). Recent systems extend this setting to multiple subjects, broader control signals, or unified creation interfaces (Huang et al., 2025b; Liang et al., 2025; Liu et al., 2025; Fei et al., 2025; Hu et al., 2025; Jiang et al., 2025; Chen et al., 2025b). These methods primarily target bidirectional video generation and typically denoise a bounded video jointly. They generally do not address open-ended causal rollouts, where completed frames cannot be revised and reference information must remain available throughout generation.

V-Warper is closely related because it injects reference appearance through value representations without additional video training (Lee et al., 2025). Its coarse stage still uses image-domain subject adaptation, and its fine stage operates during bidirectional video denoising. Custom Forcing instead stores reference images in the persistent cache of a streaming autoregressive model and adapts their contribution from the observed rollout.

A.3 Inference-time attention control

Classifier-free guidance combines conditional and unconditional model predictions at inference time (Ho & Salimans, 2022). Normalized Attention Guidance (NAG) instead extrapolates between the attention outputs for a positive and a negative prompt and normalizes the result, which keeps negative guidance stable in few-step sampling (Chen et al., 2026a). Prompt-to-Prompt and MasaCtrl manipulate cross-attention or self-attention to control image editing and consistent synthesis (Hertz et al., 2022; Cao et al., 2023). For consistent subjects without training, ConsiStory and StoryDiffusion share self-attention across the images of a batch (Tewel et al., 2024; Zhou et al., 2024b), and RefDrop shows that such sharing interpolates self-attention and reference attention, replacing the interpolation weight with a user-set constant (Fan et al., 2024). Custom Forcing builds on these ideas for a causal stream. ACG adopts the normalization of NAG but derives its negative branch by masking reference frames in the cache rather than from a negative prompt, and, unlike RefDrop, it scales the attention share that each query already places on the references instead of mixing in a constant amount, so queries that place little attention on the references remain nearly unchanged. Unlike classifier-free guidance, neither branch of ACG requires an additional forward pass, only a second attention read in each block. DVA further adjusts the reference values during sampling from an online measurement of subject similarity.

B Implementation Details

Backbone and inputs. Rolling Forcing generates 832×480 video at 16 fps in chunks of $F = 3$ latent frames, and each query attends to 27 frames, namely the $N = 9$ anchor frames, $H = 3$ recent frames, and the 15 frames being denoised. The custom K/V holds the first five DreamBooth images of each subject as $|\mathcal{R}| = 5$ reference anchors, followed by 4 copies of the custom anchor. Rolling Forcing keeps several chunks in flight at staggered noise levels, which changes when a chunk is finalized but not the role of the cache. In Eq. (7), the second factor keeps the average amplification per token of a reference anchor independent of how much of the image the subject covers. We set $\alpha_{\text{inj}} = 300$ and $\kappa = 4$, which give the best trade-off between identity and motion (Appendix F.1).

Hyperparameters. Tab. 4 lists every constant in Custom Forcing. After μ_{base} is set, the ACG blend weight increases linearly to its final value α_{acg} over W_{warm} chunks:

$$\alpha_{\text{acg}}(t) \;=\; \alpha_{\text{acg}} \cdot \min\Big(1,\; \max\Big(0,\; \frac{t - t_{\text{base}}}{W_{\text{warm}}}\Big)\Big).$$

We generate every video with seed 0. The nine images of the custom K/V take consecutive temporal positions before the first chunk, with the copies of the custom anchor last. The subject masks of the reference images come from SAM (Kirillov et al., 2023).

ACG warm-up. ACG is inactive during the first t_{base} chunks. After μ_{base} is established, we linearly increase the blend weight α_{acg} to its final value over $W_{\text{warm}} = 5$ chunks, while keeping κ fixed. Fig. 8 shows the qualitative effect of this warm-up on Deep Forcing.

	Symbol	Value
Anchor slots	N	9
Reference anchors	$\lvert\mathcal{R}\rvert$	5
Retained history frames	H	3
Chunk size / tokens per frame	F / L	3 / 1560
Injection budget	α_{inj}	300
Drift cap	$b_{\max}$	0.2
Subject-mask threshold	τ_{subj}	0.5
Baseline chunks	t_{base}	3
Amplification EMA	λ	0.95
ACG extrapolation	κ	4
ACG blend	α_{acg}	0.25
ACG clip	τ_{acg}	2.5
ACG warm-up	W_{warm}	5 chunks

Table 4: **Hyperparameters of CUSTOM FORCING.** No value is tuned per subject or per prompt, and only $b(t)$ adapts to each video.

Custom anchor. FLUX.2 [klein] 4B (Black Forest Labs, 2025) receives the five reference images and an image prompt, which is the video prompt with its opening "A cinematic video of" replaced by "A photo of". It generates one 832×480 image with seed 0, which we use without editing.

Bidirectional customization baselines. VACE-1.3B uses the released Wan2.1-VACE-1.3B checkpoint, which adds 15 VACE context blocks to the same Wan2.1 1.3B transformer as our backbone, and the official Wan2.1 inference script. CustomCrafter is trained for 200 epochs (10k optimizer steps at an effective batch size of 4 and a learning rate of 3×10^{-5}) on the five reference images and 200 class regularization images, following the official recipe. It samples 16 frames at 320×512 and 8 fps with 50 DDIM steps, a guidance scale of 12, and dynamic weighted sampling ($K = 5$, $\lambda_s = 0.4$, $\lambda_l = 0.8$). Training for 50 epochs instead lowers DINO-I from 0.490 to 0.458 but raises Dynamic Degree from 0.17 to 0.32, so longer training gains identity at the cost of motion. We report the 200-epoch model, which gives CustomCrafter its highest identity.

SMRABooth customizes Wan2.1-T2V-1.3B with a spatial subject LoRA of rank 32 on the query, key, and first feed-forward projections, regularized by representation alignment to DINOv2-ViT-g features with weight 0.05. It is trained on the five reference images, their SAM masks, and the caption "A picture of sks [class]" at 512×512 for 2,000 steps with batch size 1, using AdamW with a constant learning rate of 10^{-4} in bf16. The official recipe uses two GPUs for 1,000 steps at batch size 2, so both see the same number of samples. SMRABooth samples 49 frames at 480×832 and 15 fps with 50 steps and the official negative prompt, and raises the LoRA scale from 0.6 to 0.7 at step 25.

MotionBooth fine-tunes the full UNet of Zeroscope v2 576w with its official configuration: a masked subject loss, a subject-token cross-attention loss with weight 0.01, and a video loss with weight 1. It is trained for 300 steps at an effective batch size of 20 with AdamW (learning rate 5×10^{-6}, weight decay 0.01) and gradient clipping at 1 in bf16. The Panda-70M clips that the official recipe uses for the video loss are no longer available, so we use 500 clips of 24 frames from UCF-101 instead. MotionBooth samples 24 frames at 320×576 and 8 fps with a guidance scale of 7.5, without subject motion or camera control. For three subjects, it uses its own class names, "toy car", "toy robot", and "toy monster".

Training-based autoregressive baseline. Vanilla AR + LoRA fine-tunes Rolling Forcing itself. Following DreamBooth (Ruiz et al., 2023), we add a new token <sks> to the text encoder and train it together with LoRA adapters of rank 16 ($\alpha = 16$) on the query, key, value, and output projections of the self- and cross-attention layers. Each reference image is repeated into a static clip of 9 frames (3 latent frames) and flipped horizontally with probability 0.5. The loss is the mean squared error of the predicted clean latent at one of the five denoising steps of the distilled generator, sampled uniformly and computed through the KV cache, plus a prior-preservation loss with weight 1 on three clips from each class video of the vanilla AR model. We train for 1,000 steps with AdamW, using a learning rate of 3×10^{-4} and weight decay 0.01 for the LoRA weights and a learning rate

of 10^{-2} without weight decay for the token embedding, with 20 warm-up steps, cosine decay, and gradient clipping at 1. Training takes about 12 minutes per subject. At inference, the LoRA scale is 0.5.

Image- and reference-to-video baselines. All of them also receive the video prompt. FramePack and FramePack-F1 use their released checkpoints, which are built on HunyuanVideo, with 25 sampling steps, distilled guidance 10, and TeaCache disabled, and generate 832×480 video at 30 fps. SkyReels-V2 DF uses the 1.3B 540P checkpoint in synchronous mode with 30 sampling steps and guidance scale 6, and generates 960×544 video at 24 fps. FramePack-F1 is the forward-only variant of FramePack. SkyReels-V3 uses the R2V-14B checkpoint with the first four reference images, its maximum, and its official sampling settings, and generates 5 s of 1280×720 video at 24 fps, which its official single-shot extension (V2V-14B) continues to 32.7 s. Its pipeline sets the output shape from the aspect ratio of the first reference image, so we pad that image to 16:9 with the pipeline's own resizing. CUSTOM FORCING receives the same custom anchor and all five reference images. We cut every output to the length of our videos. The *Time* column of Tab. 2 averages the generation time per frame over the 100 videos on one H100 GPU, covering every generation stage and VAE decoding but not model loading. Most of the overhead of CUSTOM FORCING over Custom K/V only comes from measuring drift, which decodes and embeds each finalized chunk.

C EVALUATION DETAILS

Identity and prompt alignment. We use every eighth frame of a video. DINO-I averages the DINOv2-B/14 (Oquab et al., 2024) features of these frames, normalizes the mean, and takes its cosine similarity with each reference image, averaged over the five reference images. CLIP-I repeats this with CLIP ViT-B/32 (Radford et al., 2021) image features, and CLIP-T compares the same normalized mean with the CLIP text feature of the prompt. DINO-I over time applies the same computation to three 10.5 s windows. Our short videos have 501 frames at 16 fps (31.3 s), which we round to 30 s in the text, so these windows are their three thirds.

VBench. Subject Consistency, Background Consistency, and Dynamic Degree come from the official VBench implementation (Huang et al., 2024), run on each video at its native frame rate. Subject and Background Consistency measure temporal stability and Dynamic Degree measures motion, so together they show whether stability comes at the cost of motion.

D ADDITIONAL QUANTITATIVE RESULTS

D.1 MAIN RESULTS IN DETAIL

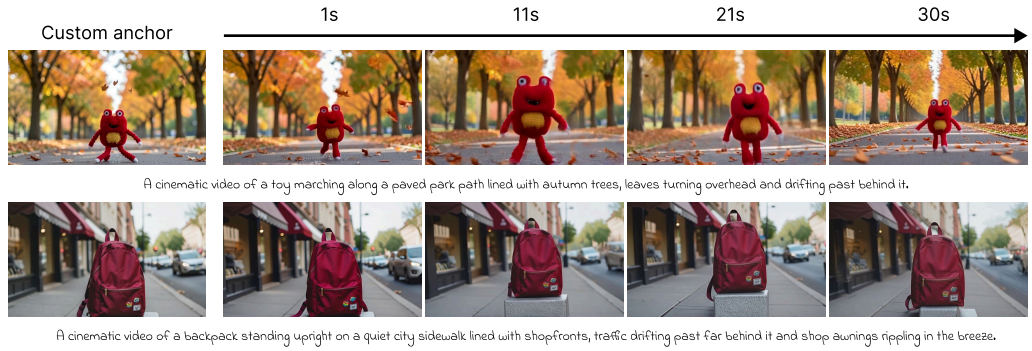
The bidirectional methods also stay below Custom K/V only in CLIP-I and CLIP-T in Tab. 1, including VACE-1.3B, which adds context blocks to our Wan2.1 1.3B backbone. Their short videos change little, which gives SMRABooth and CustomCrafter the highest Subject and Background Consistency. Vanilla AR + LoRA lowers CLIP-I, CLIP-T, and Dynamic Degree below those of the vanilla AR model. CUSTOM FORCING raises both consistency scores over Custom K/V only, and its CLIP-T stays within 0.007 of the highest value at 30 s, which the vanilla AR model reaches by following the prompt with no reference images. In Fig. 4, both subjects drift within the first minute with Custom K/V only, while CUSTOM FORCING keeps them for two minutes (Appendix E).

In Tab. 2, the CLIP-T of CUSTOM FORCING stays within 0.01 of the four image- and reference-to-video models. FramePack generates in inverted temporal order starting from the last window, so its DINO-I does not decline toward the end (0.616 to 0.623), but its whole-video DINO-I is lower than ours (0.625 against 0.635). Within 30 s, the DINO-I of FramePack-F1 falls from 0.534 to 0.392 and that of SkyReels-V2 DF from 0.552 to 0.307.

D.2 DEEP FORCING

We repeat the main experiments on Deep Forcing (Yi et al., 2025), a training-free method that extends Self Forcing (Huang et al., 2026) to long videos. We keep its default participative compression (budget 16, recent 4), write the same custom K/V into its cache, and use the constants of Tab. 4

(a) 30 s



(b) Two minutes

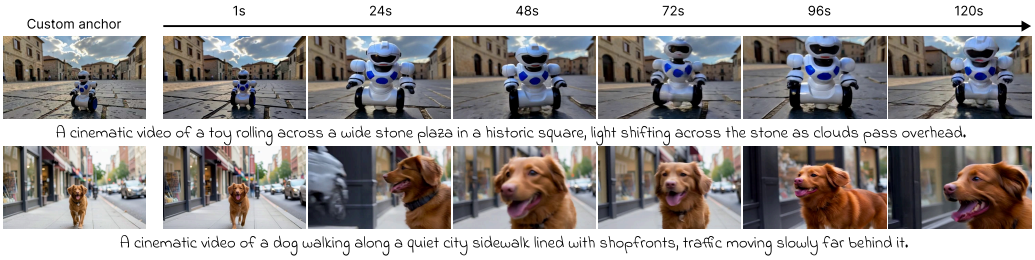


Figure 7: **CUSTOM FORCING on Deep Forcing.** The custom anchor is on the left of each row. (a) 30 s. Top: "A toy marching along a paved park path lined with autumn trees, …". Bottom: "A backpack standing upright on a quiet city sidewalk lined with shopfronts, …". (b) Two minutes. Top: "A toy rolling across a wide stone plaza in a historic square, …". Bottom: "A dog walking along a quiet city sidewalk lined with shopfronts, …". CUSTOM FORCING keeps the red body, yellow belly, and large eyes of the toy and the red color and shape of the backpack to the end of the 30 s videos, and the white body, black visor, and blue circles of the robot and the red-brown coat and drooping ears of the dog for two minutes.

without change. Tab. 5 shows the same pattern as Tab. 1. Custom K/V only again does not prevent identity drift, and CUSTOM FORCING raises DINO-I from 0.556 to 0.641 at 30 s and from 0.518 to 0.620 at 2 min, with the largest improvement in the last window (+0.110 and +0.125). Adding DVA raises DINO-I in every window, and adding ACG on top gives the highest value in every window. On this backbone, however, DVA alone lowers Dynamic Degree from 0.81 to 0.62, and adding ACG restores it. At both lengths, Subject and Background Consistency stay at the level of Custom K/V only, and Dynamic Degree is at least as high (0.87 against 0.81 at 30 s and 0.86 against 0.84 at 2 min). CUSTOM FORCING takes 1.30× the time of Custom K/V only at 30 s and 1.26× at 2 min.

Figure 7 shows CUSTOM FORCING on Deep Forcing in 30 s and two-minute videos.

				DINO-I over time ↑				Alignment ↑		Video quality		
Method	Length	DVA	ACG	all	first	middle	last	CLIP-I	CLIP-T	Subj. Cons. ↑	Bg. Cons. ↑	Dyn. Degree
Vanilla AR model	30 s	–	–	0.167	0.169	0.158	0.149	0.692	**0.364**	0.873	0.916	0.79
Custom K/V only	30 s	✗	✗	0.556	0.592	0.516	0.505	0.789	0.355	**0.890**	**0.923**	0.81
+ DVA	30 s	✓	✗	0.614	0.614	0.584	0.562	0.787	0.347	0.878	0.921	0.62
+ DVA + ACG (Ours)	30 s	✓	✓	**0.641**	**0.634**	**0.623**	**0.615**	**0.798**	0.348	0.889	**0.923**	0.87
Custom K/V only	2 min	✗	✗	0.518	0.586	0.466	0.447	0.780	**0.354**	0.859	0.911	0.84
+ DVA + ACG (Ours)	2 min	✓	✓	**0.620**	**0.632**	**0.583**	**0.571**	**0.793**	0.342	**0.866**	**0.914**	0.86

Table 5: **Results on Deep Forcing**, $n = 100$. Columns and bold follow Tab. 1, and rows follow Tab. 3.

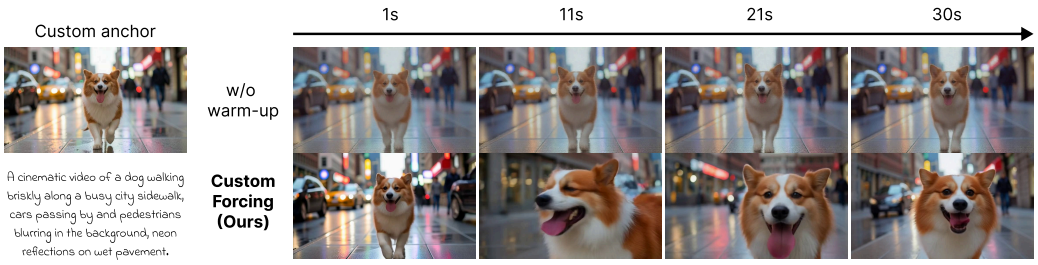


Figure 8: **Effect of the ACG warm-up on Deep Forcing**, Without warm-up, the generation remains close to the custom-anchor pose and exhibits limited motion. With warm-up, the subject shows greater motion while preserving its identity.

D.3 Results per Subject

Tab. 6 breaks the 30 s results down by subject. Custom Forcing raises the last window over Custom K/V only for every subject, by 0.142 on average (subject-level bootstrap 95% confidence interval 0.093 to 0.189) and in 88 of the 100 prompts. The gain is largest for the animals, which show the most identity drift under Custom K/V only, and smallest for the backpack and the RC car, which show little. The Custom anchor column is the DINO-I of the custom anchor image itself, 0.679 on average. Custom K/V only and Custom Forcing both start from this image and differ in how much of its identity survives to the last window.

		DINO-I, last window		
Subject	Custom anchor	Custom K/V only	Custom Forcing	Gain
backpack	0.669	0.647	0.662	+0.015
cat	0.695	0.425	0.636	+0.211
cat2	0.788	0.481	0.702	+0.221
dog	0.656	0.509	0.617	+0.108
dog3	0.460	0.055	0.316	+0.261
dog7	0.706	0.452	0.604	+0.151
dog8	0.678	0.407	0.589	+0.182
monster_toy	0.611	0.494	0.554	+0.060
rc_car	0.715	0.687	0.723	+0.036
robot_toy	0.810	0.589	0.761	+0.172
All	0.679	0.475	0.616	+0.142

Table 6: **Results per subject**, 30 s, 10 prompts each.

D.4 Drift Sensor

DVA senses drift with DINOv2-B, the same encoder behind DINO-I. Tab. 7 swaps the sensor for ConvNeXt-B (Liu et al., 2022) and scores identity with both encoders. The scores change by at most 0.003.

	Scored with	
Drift sensor	DINO-I	ConvNeXt-I
DINOv2-B (Ours)	0.635	0.626
ConvNeXt-B	0.634	0.623

Table 7: **Drift sensor and evaluator**, 30 s, $n = 100$.

D.5 Identity over Time

Fig. 9 plots DINO-I in 12 equal bins over the video. With Custom K/V only, it declines steadily, from 0.61 to 0.44 over 30 s and from 0.59 to 0.41 over 2 min. DVA stops the decline after about 10 s,

as its amplification responds to the drift, and ACG alone starts highest but still declines. CUSTOM FORCING stays near 0.60 for 30 s and ends the two-minute video at 0.58.

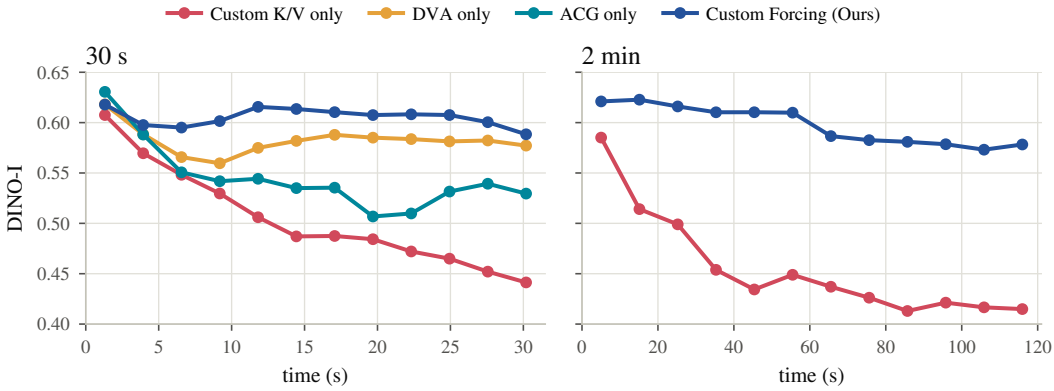


Figure 9: **DINO-I over time**, $n = 100$. Left: 30 s with each component. Right: 2 min.

D.6 CLASS PRIOR

Section 3.4 observes that the subject of Custom K/V only drifts toward generic instances of its category. To measure this, we compute s_{prior}, which is DINO-I with the frames of the vanilla AR video for the same prompt in place of the reference images, and report it on the same thirds as DINO-I (Tab. 8). With Custom K/V only, s_{prior} rises from 0.395 to 0.418 as DINO-I falls from 0.579 to 0.475. Adding ACG lowers s_{prior} in every third, by 0.015 to 0.040, and keeps it near 0.38, while DINO-I rises by 0.016 to 0.076.

	DINO-I ↑			s_{prior} ↓		
Method	first	middle	last	first	middle	last
Custom K/V only	0.579	0.506	0.475	0.395	0.414	0.418
+ ACG	**0.595**	**0.551**	**0.550**	**0.380**	**0.385**	**0.378**

Table 8: **Similarity to the reference images and to the class prior**, 30 s, $n = 100$.

D.7 USER STUDY AND VLM JUDGES

User study. We select 20 of the 100 prompts, two per subject, and pair the 30 s video of CUSTOM FORCING for each prompt with that of each of five baselines, Custom K/V only and the four image- and reference-to-video models, which gives 100 pairs. Thirty participants took part. Each participant rates 20 pairs, assigned by a Latin square so that every prompt appears once and every baseline four times, which gives 600 comparisons and 120 per baseline. Each screen shows the reference images, the prompt, and the two videos playing in sync, in random order and without method names, and participants can pause, seek, and speed up the videos. For each pair, they answer two questions, "Which video keeps the subject more similar to the reference photos, from start to end?" and "Which video has better overall quality, considering visual clarity, artifacts, and how natural the motion looks?" Fig. 10 shows one screen, on which participants compare the two videos and answer both questions by choosing A or B.

VLM judge. We ask Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025), Claude Sonnet 5, and Claude Opus 5.5 the same two questions on the same 100 pairs. Each pair is given as one image that contains the reference images, the prompt, and ten frames of each video at $t = 3, 6, \ldots, 30$ s in two rows, with CUSTOM FORCING in the top row for half of the pairs of each baseline. The model answers Top or Bottom for each question.

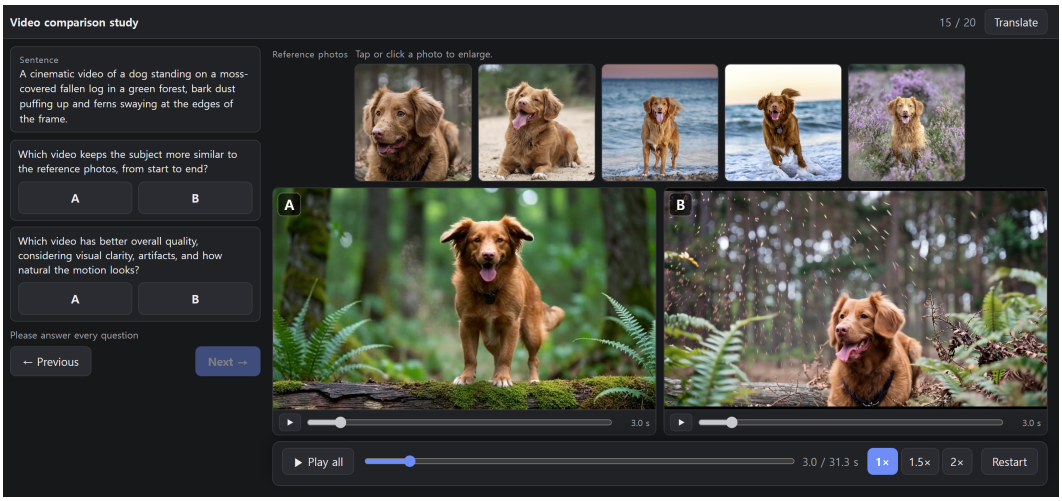


Figure 10: **User study interface.** Participants see the prompt, the reference images and choose A or B for each question before moving to the next pair.

Results. Tab. 9 gives the percentage of judgments that prefer CUSTOM FORCING. For the subject, participants prefer CUSTOM FORCING over Custom K/V only, FramePack, FramePack-F1, and SkyReels-V2 DF (60.0 to 64.2%) and in 58.5% of all 600 votes. Against SkyReels-V3, whose 14B backbone reaches a similar whole-video DINO-I, they show no significant preference (45.8%). For quality, they prefer CUSTOM FORCING over FramePack (70.0%) and overall (56.8%), and show no significant preference against Custom K/V only (45.8%), so DVA and ACG improve the identity without a measurable loss in quality.

Qwen3-VL, Claude Sonnet 5, and Claude Opus 5.5 also prefer CUSTOM FORCING overall, in 66, 63, and 76% of the pairs for the subject and in 67, 63, and 75% for quality. All three prefer it over FramePack-F1 and SkyReels-V2 DF in 65 to 95% of the pairs. All three VLM judges differ from the participants on SkyReels-V3, which they prefer less than CUSTOM FORCING (60 to 90%), and they rate the quality of FramePack closer to CUSTOM FORCING than the participants do (45 to 65% against 70.0%). They disagree with each other on Custom K/V only, which Claude Sonnet 5 prefers for both questions (40 and 35%), while Qwen3-VL and Claude Opus 5.5 prefer CUSTOM FORCING for the subject (60 and 80%). Unlike the participants, the VLM judges see ten frames rather than the videos.

	Human		Qwen3-VL-235B		Claude Sonnet 5		Claude Opus 5.5	
CUSTOM FORCING vs.	Subject	Quality	Subject	Quality	Subject	Quality	Subject	Quality
Custom K/V only	60.0	45.8	60.0	50.0	40.0	35.0	80.0	60.0
FramePack	60.0	70.0	50.0	45.0	45.0	50.0	55.0	65.0
FramePack-F1	62.5	57.5	65.0	70.0	80.0	75.0	95.0	85.0
SkyReels-V2 DF	64.2	59.2	75.0	80.0	80.0	75.0	90.0	85.0
SkyReels-V3	45.8	51.7	80.0	90.0	70.0	80.0	60.0	80.0
All	58.5	56.8	66.0	67.0	63.0	63.0	76.0	75.0

Table 9: **User study and VLM judges.** Percentage of judgments that prefer CUSTOM FORCING over each method; 50% means no preference. Human: 120 votes per cell from 30 participants. VLM judges: Qwen3-VL-235B, Claude Sonnet 5, and Claude Opus 5.5, each with one judgment on each of 20 pairs per cell and 100 in *All*.

E Qualitative Observations

Two-minute generation. In Fig. 4, the eyes of the toy under Custom K/V only grow from 24 s and rise on stalks by 48 s while its body swells, and the gray merle dog turns black tricolor by 48 s. Custom Forcing keeps the appearance of both subjects for the full two minutes.

Image- and reference-to-video models. In Fig. 5, FramePack and SkyReels-V3 keep the fur of both animals until 30 s as Custom Forcing does, but with ten times its parameters. The image-to-video models receive only the custom anchor, and SkyReels-V3 sees the reference images only in its first 5 s. Custom Forcing instead keeps them in its KV cache throughout generation, and its DINO-I stays between 0.616 and 0.626 from the first window to the last.

Adding the components. In Fig. 6, Custom K/V only raises the ears of the dog and turns the toy into a four-wheeled car by 21 s. Adding DVA, which corrects the strength of the reference conditioning, delays the ears to 30 s and keeps the toy a robot, but with a different face and shoulders. Adding ACG, which corrects its direction, keeps both subjects as in the reference images.

F Additional Ablations

F.1 Injection and Guidance Strength

Tab. 10 varies the injection budget α_{inj} at $\kappa = 4$. A smaller budget lowers identity (-0.009 DINO-I at 180). A larger one does not raise it, and Subject Consistency, Background Consistency, and Dynamic Degree fall ($0.41 \rightarrow 0.33$ at 420). Identity also drops toward the end of the video. At 420, DINO-I on the last third of the video is 0.041 below the first third, against 0.003 at 300. Tab. 11 varies the guidance strength κ at $\alpha_{\text{inj}} = 300$. DINO-I rises up to $\kappa = 4$ and saturates at 5, where Dynamic Degree already falls to 0.36. At $\kappa = 6$, CLIP-I, Subject Consistency, and Dynamic Degree fall ($0.41 \rightarrow 0.32$), and DINO-I drops toward the end of the video as it does at a large α_{inj}. On both axes, the operating point thus gives the best trade-off between identity and motion.

The blend weight needs no separate sweep. By Eq. (13), when the clip does not bind, α_{acg} and κ act only through $\alpha_{\text{acg}}(\kappa - 1)$, which the κ sweep already covers.

α_{inj}	DINO-I	CLIP-I	Subj. Cons.	Bg. Cons.	Dynamic
180	0.626	0.801	**0.930**	**0.947**	0.41
240	0.632	**0.803**	0.929	0.946	0.40
300*	**0.635**	**0.803**	0.928	0.945	0.41
360	0.632	**0.803**	0.923	0.942	0.35
420	0.623	0.801	0.918	0.941	0.33

Table 10: **Injection budget α_{inj}** at $\kappa = 4$, $n = 100$.

κ	DINO-I	CLIP-I	Subj. Cons.	Bg. Cons.	Dynamic
2	0.623	**0.805**	0.924	**0.945**	0.39
3	0.629	0.804	0.926	**0.945**	0.40
4*	<u>0.635</u>	0.803	**0.928**	**0.945**	0.41
5	**0.636**	0.802	0.926	0.944	0.36
6	0.626	0.797	0.920	0.943	0.32

Table 11: **Guidance strength κ** at $\alpha_{\text{inj}} = 300$, $n = 100$.

* The operating point of Custom Forcing. Bold marks the best value in each column, and an underline marks the operating point where it ranks second. Dynamic Degree is not bolded.

F.2 Composition of the Custom K/V

Fig. 11 varies $|\mathcal{R}|$ at fixed $N = 9$, so every reference image added replaces a copy of the custom anchor. DINO-I barely moves. It stays between 0.620 and 0.640 and is lowest at the two ends, $|\mathcal{R}| = 1$ and $|\mathcal{R}| = 8$. What $|\mathcal{R}|$ changes is the balance between the two kinds of anchor. With few reference images, the copies of the custom anchor dominate and the video barely moves, with a Dynamic Degree of 0.14 at $|\mathcal{R}| = 1$. With many, the one remaining copy cannot hold the scene. At $|\mathcal{R}| = 8$, Subject and Background Consistency fall below those of Custom K/V only (0.892 and 0.933 against 0.913 and 0.939). The higher Dynamic Degree there (0.74) reflects the scene drifting rather than the subject moving. $|\mathcal{R}| = 5$ is the smallest value at which Dynamic Degree reaches that of Custom K/V only (0.41), and the largest that uses each reference image once. Both consistency scores remain above those of Custom K/V only at this value.

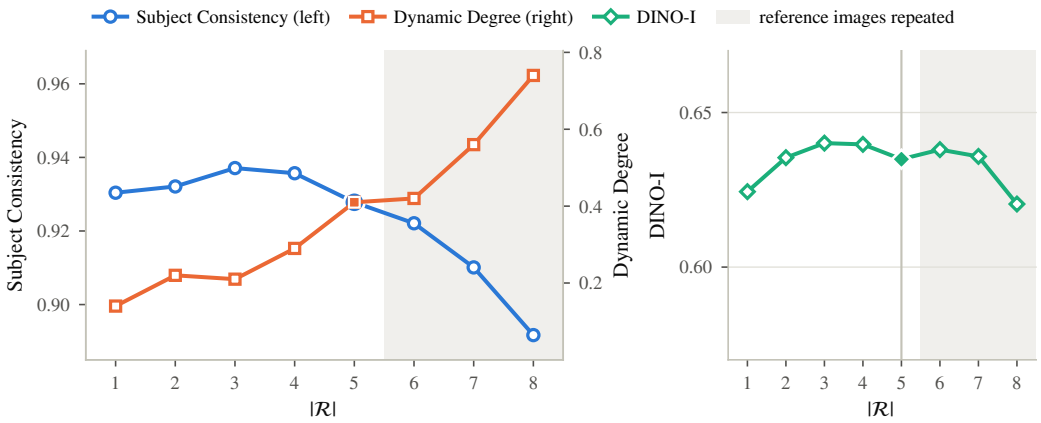


Figure 11: **Number of reference anchors at** $N = 9$, 30 s, $n = 100$. Reference images are repeated in the shaded region. Left: Subject Consistency and Dynamic Degree, with the two axes aligned at the default $|\mathcal{R}| = 5$. Right: DINO-I.

F.3 Generated Images in Place of the Reference Images

Tab. 12 replaces the reference images in the custom K/V with images that FLUX.2 [klein] generates from the same reference images and prompt with other seeds, without selection. G5K4 fills the five reference slots with seeds 1 to 5 and keeps four copies of the custom anchor. G8K1 uses seeds 1 to 8 and keeps one copy. No reference image enters the cache, although the generated images carry the subject from them, and DINO-I is still measured against the reference images. With G5K4, Custom Forcing scores only 0.005 DINO-I below the default R5K4, but Dynamic Degree drops from 0.41 to 0.30. With G8K1, DINO-I is 0.020 lower. In every setting, DVA and ACG raise DINO-I most in the last window.

		DINO-I				
		all	first	middle	last	Dynamic
R5K4	Custom K/V only	0.541	0.579	0.506	0.475	0.41
	Custom Forcing (Ours)	**0.635**	0.619	0.626	**0.616**	0.41
G5K4	Custom K/V only	0.571	0.601	0.550	0.523	0.34
	Custom Forcing (Ours)	0.630	**0.628**	**0.628**	0.606	0.30
G8K1	Custom K/V only	0.556	0.577	0.538	0.512	0.40
	Custom Forcing (Ours)	0.615	0.611	0.595	0.583	0.56

Table 12: **What the custom K/V holds**, 30 s, $n = 100$. R, G, and K count reference images, generated images, and copies of the custom anchor. Bold marks the best value in each column, and Dynamic Degree is not bolded.

F.4 Constant Amplification and Subject Mask

Tab. 13 replaces the measured drift $b(t)$ with a constant and removes the subject mask. A constant $b \equiv 0.05$ raises DINO-I by 0.016, but the subject grows until it fills the frame (Fig. 12). CLIP-T, Subject Consistency, and Dynamic Degree all drop. Larger constants break the video. The adaptive amplification avoids this because it stays small. Measured $b(t)$ averages 0.022, which gives a target amplification of about 1.3 against 3 for $b \equiv 0.05$, and exceeds 0.05 in only 14% of chunks. Removing the subject mask lowers DINO-I, Subject Consistency, and Dynamic Degree. Background tokens carry no identity, and amplifying them ties the video to the backgrounds of the reference images.

	DINO-I	CLIP-I	CLIP-T	Subj. Cons.	Bg. Cons.	Dynamic
CUSTOM FORCING (Ours)	0.635	0.803	**0.353**	**0.928**	**0.945**	0.41
$b \equiv 0.05$	**0.651**	**0.804**	0.335	0.887	0.926	0.28
$b \equiv 0.1$	0.403	0.729	0.329	0.793	0.875	0.08
$b \equiv 0.2$	0.158	0.643	0.309	0.691	0.823	0.02
no subject mask	0.622	0.803	0.352	0.913	0.936	0.30

Table 13: **Adaptive amplification and subject mask**, 30 s, $n = 100$. The largest constant, 0.2, equals the cap $b_{\max}$. Bold marks the best value in each column, and Dynamic Degree is not bolded.



Figure 12: **A constant amplification raises DINO-I but breaks the scene.** Frames at 17.5 s of two prompts, whose DINO-I rises from 0.612 to 0.681 and from 0.482 to 0.530.

G LIMITATIONS

CUSTOM FORCING controls identity for the subject as a whole, not for its parts. The drift signal $s(t)$ compares the whole subject with the reference images, and DVA sets the amplification of every subject token of the reference anchors from this single signal. A change confined to one small part, such as a wheel of the RC car or a strap of the backpack, may therefore change $s(t)$ little, and CUSTOM FORCING cannot target that part on its own. The fidelity of CUSTOM FORCING is also bounded by the frozen backbone. CUSTOM FORCING changes only how the backbone attends to the anchors, so appearance that the backbone cannot render remains out of reach. Finally, we evaluate one subject per video for up to two minutes. Multiple subjects and longer horizons remain for future work.

H EVALUATION PROMPTS

Table 14 lists all evaluation prompts. Every prompt names the subject only by its category. Three pairs of subjects share the same ten prompts (cat and cat2, dog and dog3, dog7 and dog8), so the 100 prompts contain 70 distinct sentences. The ten prompts of robot_toy describe five scenes twice, once with the toy moving and once with it holding still.

Table 14: **Evaluation prompts**, grouped by subject. The noun used in the prompts is in parentheses.

Slug	Prompt
backpack ("backpack")	
00_park_bench	A cinematic video of a backpack resting on a weathered wooden park bench in an autumn park, orange leaves turning overhead and drifting past behind it.
01_forest_rock	A cinematic video of a backpack sitting on a flat mossy rock beside a winding forest trail, dappled sunlight flickering through the canopy and ferns swaying at the edges.
02_city_sidewalk	A cinematic video of a backpack standing upright on a quiet city sidewalk lined with shopfronts, traffic drifting past far behind it and shop awnings rippling in the breeze.
03_beach	A cinematic video of a backpack resting on flat open sand at a wide beach, fine sand streaming past it and waves breaking far behind.
04_mountain_ridge	A cinematic video of a backpack sitting on a rocky mountain ridge above a deep valley, wind tugging at its straps and clouds drifting across the valley below.
05_wooden_floor	A cinematic video of a backpack standing upright on a polished wooden floor in a sunlit room, light shifting across the boards as clouds pass the window.
06_cherry_blossom	A cinematic video of a backpack resting on a stone path under cherry trees in full bloom, pink petals drifting down around it and branches swaying overhead.
07_library_desk	A cinematic video of a backpack resting on a long wooden desk in a bright library, dust motes floating in the window light and a page lifting in the draft.
08_cobblestone	A cinematic video of a backpack standing upright on a cobblestone street in an old town, light shifting across the stones as clouds pass overhead.
09_train_platform	A cinematic video of a backpack standing upright on an empty train platform under a long canopy, a train passing far down the line and litter stirring in the draft.
cat ("cat")	
00_windowsill	A cinematic video of a cat stretching on a wooden windowsill in a bright room, its back arching and a curtain stirring beside it.
01_garden_path	A cinematic video of a cat trotting along a stone garden path between low flower beds, its tail swaying and petals trembling in the breeze.
02_rooftop	A cinematic video of a cat walking across a flat sunlit rooftop above a quiet city, its fur stirring in the wind and clouds drifting behind it.
03_wooden_desk	A cinematic video of a cat sitting on a wooden desk beside a stack of books in a study, a page lifting in the draft as it lands.
04_cherry_blossom	A cinematic video of a cat running down a cherry blossom street in full spring bloom, pink petals scattering in its wake.
05_sofa	A cinematic video of a cat lying on a cozy sofa in a warmly lit living room, cushions dipping beneath it and dust motes floating in the light.
06_courtyard	A cinematic video of a cat trotting across a sunlit cobblestone courtyard in an old town, light shifting across the stones as clouds pass.
07_park_lawn	A cinematic video of a cat running across a wide grassy park lawn with trees behind it, grass parting around its paws and leaves turning overhead.
08_stone_wall	A cinematic video of a cat walking along a low stone wall beside a country lane, its tail counterbalancing each step and wildflowers nodding below.
09_terracotta_patio	A cinematic video of a cat sitting on a sunlit terracotta patio beside potted plants, potted leaves stirring as it passes.

Slug	Prompt (continued)
cat2 (“cat”)	
00_windowsill	A cinematic video of a cat stretching on a wooden windowsill in a bright room, its back arching and a curtain stirring beside it.
01_garden_path	A cinematic video of a cat trotting along a stone garden path between low flower beds, its tail swaying and petals trembling in the breeze.
02_rooftop	A cinematic video of a cat walking across a flat sunlit rooftop above a quiet city, its fur stirring in the wind and clouds drifting behind it.
03_wooden_desk	A cinematic video of a cat sitting on a wooden desk beside a stack of books in a study, a page lifting in the draft as it lands.
04_cherry_blossom	A cinematic video of a cat running down a cherry blossom street in full spring bloom, pink petals scattering in its wake.
05_sofa	A cinematic video of a cat lying on a cozy sofa in a warmly lit living room, cushions dipping beneath it and dust motes floating in the light.
06_courtyard	A cinematic video of a cat trotting across a sunlit cobblestone courtyard in an old town, light shifting across the stones as clouds pass.
07_park_lawn	A cinematic video of a cat running across a wide grassy park lawn with trees behind it, grass parting around its paws and leaves turning overhead.
08_stone_wall	A cinematic video of a cat walking along a low stone wall beside a country lane, its tail counterbalancing each step and wildflowers nodding below.
09_terracotta_patio	A cinematic video of a cat sitting on a sunlit terracotta patio beside potted plants, potted leaves stirring as it passes.
dog (“dog”)	
00_beach	A cinematic video of a dog running along the shoreline of a sandy beach at sunset, paws kicking up sand, waves rolling in and out, fur blowing in the sea breeze.
01_forest	A cinematic video of a dog trotting down a winding forest trail, dappled sunlight flickering through swaying leaves, dust and pollen drifting in the air.
02_city_street	A cinematic video of a dog walking briskly along a busy city sidewalk, cars passing by and pedestrians blurring in the background, neon reflections on wet pavement.
03_sunflower_field	A cinematic video of a dog bounding playfully through a vast field of blooming sunflowers, petals and leaves swaying in the wind, butterflies fluttering up.
04_mountain_top	A cinematic video of a dog standing on a rocky mountain peak, fur ruffled by strong gusts of wind, clouds drifting across the valley below.
05_snowy_clearing	A cinematic video of a dog trotting through a snowy clearing ringed by dark evergreen pines, snow puffing up from its paws, flakes drifting down from the branches.
06_wooden_bridge	A cinematic video of a dog trotting across an old wooden bridge, a river rushing and sparkling below, planks creaking, mist rising from the water.
07_eiffel_tower	A cinematic video of a dog walking in front of the Eiffel Tower, tail wagging, tourists strolling past and pigeons taking flight, soft golden hour light.
08_sofa	A cinematic video of a dog lying on a cozy sofa in a warmly lit living room, curtains gently swaying by an open window, dust motes floating in the light.
09_cherry_blossom	A cinematic video of a dog walking down a cherry blossom street in spring, pink petals falling and drifting on the breeze, branches swaying overhead.

Slug	Prompt (continued)
dog3 (“dog”)	
00_beach	A cinematic video of a dog running along the shoreline of a sandy beach at sunset, paws kicking up sand, waves rolling in and out, fur blowing in the sea breeze.
01_forest	A cinematic video of a dog trotting down a winding forest trail, dappled sunlight flickering through swaying leaves, dust and pollen drifting in the air.
02_city_street	A cinematic video of a dog walking briskly along a busy city sidewalk, cars passing by and pedestrians blurring in the background, neon reflections on wet pavement.
03_sunflower_field	A cinematic video of a dog bounding playfully through a vast field of blooming sunflowers, petals and leaves swaying in the wind, butterflies fluttering up.
04_mountain_top	A cinematic video of a dog standing on a rocky mountain peak, fur ruffled by strong gusts of wind, clouds drifting across the valley below.
05_snowy_clearing	A cinematic video of a dog trotting through a snowy clearing ringed by dark evergreen pines, snow puffing up from its paws, flakes drifting down from the branches.
06_wooden_bridge	A cinematic video of a dog trotting across an old wooden bridge, a river rushing and sparkling below, planks creaking, mist rising from the water.
07_eiffel_tower	A cinematic video of a dog walking in front of the Eiffel Tower, tail wagging, tourists strolling past and pigeons taking flight, soft golden hour light.
08_sofa	A cinematic video of a dog lying on a cozy sofa in a warmly lit living room, curtains gently swaying by an open window, dust motes floating in the light.
09_cherry_blossom	A cinematic video of a dog walking down a cherry blossom street in spring, pink petals falling and drifting on the breeze, branches swaying overhead.
dog7 (“dog”)	
00_beach	A cinematic video of a dog running along a wide sandy beach, sand kicking up from its paws and the sea breeze lifting its fur.
01_forest_trail	A cinematic video of a dog trotting down a winding forest trail, sunlight flickering through the canopy and pollen drifting in the air.
02_city_sidewalk	A cinematic video of a dog walking along a quiet city sidewalk lined with shopfronts, traffic moving slowly far behind it.
03_wheat_field	A cinematic video of a dog running through a golden wheat field under an open sky, wheat stalks parting and bending in the wind.
04_mountain_ridge	A cinematic video of a dog standing on a rocky mountain ridge above a deep valley, wind ruffling its coat and clouds drifting across the valley behind it.
05_fallen_log	A cinematic video of a dog standing on a moss-covered fallen log in a green forest, bark dust puffing up and ferns swaying at the edges of the frame.
06_autumn_park	A cinematic video of a dog running through a park path buried in orange autumn leaves, leaves scattering behind it and branches swaying overhead.
07_sofa	A cinematic video of a dog lying on a cozy sofa in a warmly lit living room, curtains swaying gently by an open window and dust motes floating in the light.
08_cherry_blossom	A cinematic video of a dog walking down a cherry blossom street in full spring bloom, pink petals drifting down and branches swaying overhead.
09_desert_dune	A cinematic video of a dog walking across a wind-sculpted sandy desert dune under a wide open sky, fine sand streaming off the crest behind it.

Slug	Prompt (continued)
dog8 (“dog”)	
00_beach	A cinematic video of a dog running along a wide sandy beach, sand kicking up from its paws and the sea breeze lifting its fur.
01_forest_trail	A cinematic video of a dog trotting down a winding forest trail, sunlight flickering through the canopy and pollen drifting in the air.
02_city_sidewalk	A cinematic video of a dog walking along a quiet city sidewalk lined with shopfronts, traffic moving slowly far behind it.
03_wheat_field	A cinematic video of a dog running through a golden wheat field under an open sky, wheat stalks parting and bending in the wind.
04_mountain_ridge	A cinematic video of a dog standing on a rocky mountain ridge above a deep valley, wind ruffling its coat and clouds drifting across the valley behind it.
05_fallen_log	A cinematic video of a dog standing on a moss-covered fallen log in a green forest, bark dust puffing up and ferns swaying at the edges of the frame.
06_autumn_park	A cinematic video of a dog running through a park path buried in orange autumn leaves, leaves scattering behind it and branches swaying overhead.
07_sofa	A cinematic video of a dog lying on a cozy sofa in a warmly lit living room, curtains swaying gently by an open window and dust motes floating in the light.
08_cherry_blossom	A cinematic video of a dog walking down a cherry blossom street in full spring bloom, pink petals drifting down and branches swaying overhead.
09_desert_dune	A cinematic video of a dog walking across a wind-sculpted sandy desert dune under a wide open sky, fine sand streaming off the crest behind it.
monster_toy (“toy”)	
00_wooden_floor	A cinematic video of a toy walking across a polished wooden floor in a sunlit room, its limbs swinging with each step and light shifting across the boards.
01_grass_lawn	A cinematic video of a toy running across a short grass lawn in a sunny garden, its arms swinging and grass rippling around its feet.
02_park_path	A cinematic video of a toy marching along a paved park path lined with autumn trees, leaves turning overhead and drifting past behind it.
03_stone_steps	A cinematic video of a toy sitting on worn stone steps on a quiet old street, its arms reaching for the next step.
04_beach	A cinematic video of a toy standing on flat open sand at a wide beach, sand puffing up beneath it and the sea breeze stirring its fur.
05_bed	A cinematic video of a toy sitting on a neatly made bed in a bright bedroom, the duvet dipping and springing back beneath it.
06_cherry_blossom	A cinematic video of a toy standing on a stone path under cherry trees in full bloom, its arm swinging in a wave and pink petals drifting down around it.
07_picnic_blanket	A cinematic video of a toy sitting on a checkered picnic blanket in a green meadow, its arms swinging side to side and grass rippling around it.
08_garden_path	A cinematic video of a toy walking along a stone garden path between low flower beds, its feet lifting in turn and petals trembling in the breeze.
09_woven_rug	A cinematic video of a toy sitting on a patterned woven rug in a warmly lit room, it rolling over once and warm light shifting across the rug.

Slug	Prompt (continued)
rc_car (“toy”)	
00_garage_floor	A cinematic video of a toy rolling across a smooth concrete floor in a bright garage, its wheels spinning and reflections sliding across the polished floor, camera tracking low alongside.
01_city_sidewalk	A cinematic video of a toy driving along a quiet city sidewalk lined with shopfronts, its wheels turning and shopfronts sliding past behind it, camera panning to keep pace.
02_dirt_track	A cinematic video of a toy driving along a flat dirt track across an open field, loose grit spraying from its wheels, camera tracking smoothly alongside.
03_beach	A cinematic video of a toy driving across flat firm sand at a wide beach, sand streaming out behind its wheels, camera gliding low alongside.
04_cobblestone	A cinematic video of a toy driving along a cobblestone street in an old town, the body bouncing over the stones, camera dollying alongside.
05_wooden_floor	A cinematic video of a toy rolling across a polished wooden floor in a sunlit room, its reflection sliding along the boards, camera tracking low from the side.
06_parking_lot	A cinematic video of a toy driving across a wide empty asphalt parking lot, painted lines sweeping past beneath it, camera craning up as it passes.
07_park_path	A cinematic video of a toy driving along a paved park path lined with trees, dappled light sweeping across its body, camera tracking low alongside.
08_stone_plaza	A cinematic video of a toy driving across a wide stone plaza in a historic square, pigeons scattering far ahead of it, camera slowly orbiting as it drives.
09_race_track	A cinematic video of a toy driving along a painted asphalt race track with curbing along its edge, curb stripes flashing past beneath it, camera tracking low alongside.
robot_toy (“toy”)	
00_wooden_floor__motion	A cinematic video of a toy rolling across a polished wooden floor in a sunlit room, light shifting across the boards as clouds pass the window.
01_wooden_floor__static	A cinematic video of a toy standing on a polished wooden floor in a sunlit room, light shifting across the boards as clouds pass the window.
02_park_path__motion	A cinematic video of a toy rolling across a paved park path lined with autumn trees, leaves turning overhead and dappled light sweeping over it.
03_park_path__static	A cinematic video of a toy standing on a paved park path lined with autumn trees, leaves turning overhead and dappled light sweeping over it.
04_stone_plaza__motion	A cinematic video of a toy rolling across a wide stone plaza in a historic square, light shifting across the stone as clouds pass overhead.
05_stone_plaza__static	A cinematic video of a toy standing on a wide stone plaza in a historic square, light shifting across the stone as clouds pass overhead.
06_grass_lawn__motion	A cinematic video of a toy rolling across a short grass lawn in a sunny garden, blades swaying around it and light shifting as clouds pass.
07_grass_lawn__static	A cinematic video of a toy standing on a short grass lawn in a sunny garden, blades swaying around it and light shifting as clouds pass.
08_race_track__motion	A cinematic video of a toy rolling across a painted asphalt race track with curbing along its edge, heat shimmering off the asphalt and banners rippling beyond the curbing.
09_race_track__static	A cinematic video of a toy standing on a painted asphalt race track with curbing along its edge, heat shimmering off the asphalt and banners rippling beyond the curbing.