TRAINING-FREE CONFLICT-AWARE ATTRIBUTE FORGETTING FOR CUSTOMIZED VIDEO GENERATION

Anonymous authors
Paper under double-blind review

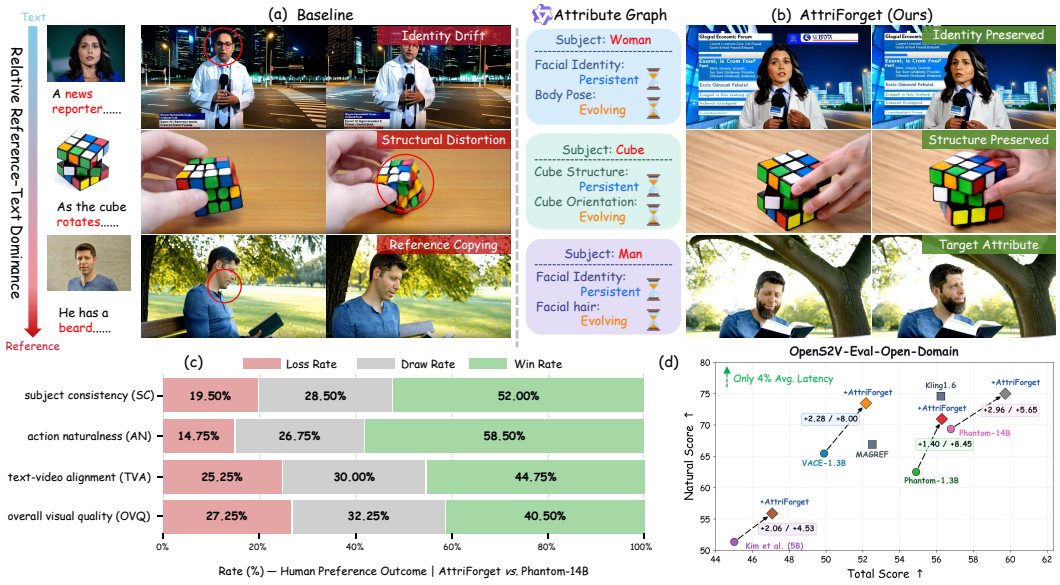


Figure 1: Motivation and Results. **(a)** Failure modes under varying relative reference-text dominance. **(b)** AttriForget models the reference validity of subject attributes to resolve these failures. **(c)** Human rating outcomes against Phantom-14B. **(d)** Performance across different S2V backbones.

ABSTRACT

Customized video generation aims to preserve personalized subject characteristics from reference images while generating dynamics aligned with text prompts. Existing methods typically apply reference conditioning uniformly throughout generation, implicitly assuming that all reference attributes remain valid for the target event. When a prompt requires an attribute to change, reference and textual conditions impose incompatible attribute states, resulting in *attribute conflict*. To address this conflict, we propose **AttriForget**, a training-free framework for conflict-aware attribute forgetting that constructs an event-conditioned attribute graph to model the validity of reference attributes under the target event. Guided by this graph, AttriForget probes orthogonal attribute-sensitive subspaces in reference features to preserve information associated with persistent attributes while "forgetting" reference constraints on evolving attributes, turning conflicting reference and textual conditions into complementary guidance. Furthermore, AttriForget can be integrated into various Diffusion Transformer architectures without additional training, extending global reference control to fine-grained attribute-level reference control. Experiments on OpenS2V-Eval across four subject-customized video generation backbones show that AttriForget consistently improves total scores and video naturalness while enhancing subject consistency on most evaluated settings, with an average inference latency increase of only 4% across multiple backbones. Code will be made publicly available.

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

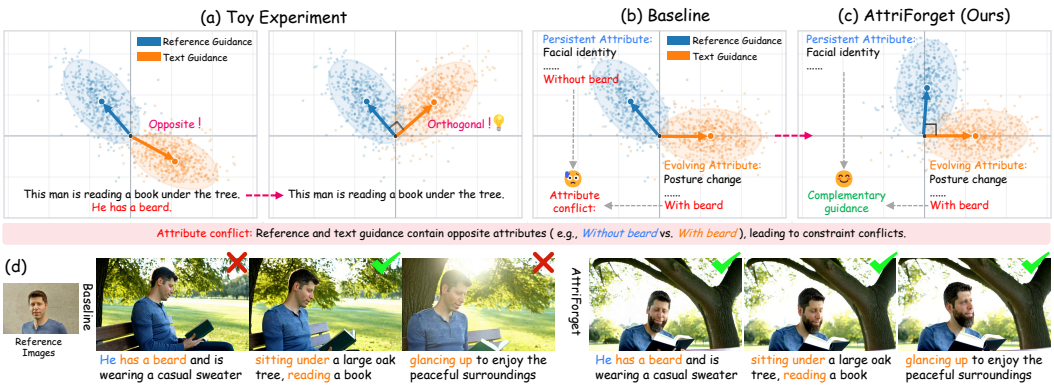


Figure 2: Analysis of attribute conflict. (a) Removing the conflicting attribute shifts reference and text guidance from opposition to near orthogonality. (b) Unified reference conditioning entangles persistent and evolving attributes, causing attribute conflict. (c) AttriForget forgets reference constraints on evolving attributes, enabling complementary guidance. (d) Improved Qualitative Results.

1 Introduction

Recent advances in video generation models, particularly Diffusion Transformers (DiT) (Peebles & Xie, 2023), have substantially improved subject consistency and visual quality, making customized video generation (Liu et al., 2025; Deng et al., 2025; Xue et al., 2026) an important direction for personalized content creation. Unlike customized image generation (Ruiz et al., 2023; Shi et al., 2024), customized video generation requires not only preserving the identity and personalized characteristics of the reference subject, but also allowing its attributes to evolve in accordance with the events specified by the text prompt. Balancing subject preservation with sufficient generative flexibility for event-driven evolution remains a central challenge in customized video generation.

To preserve personalized subject characteristics, existing customized video generation methods (Guo et al., 2026; Li et al., 2026a; Jiang et al., 2025) typically introduce reference images as additional visual conditions, reinforcing subject consistency through continuous attention interactions between reference information and video features. However, these methods (Wang et al., 2026a; Zhang et al., 2026b) generally apply all reference information as a unified and persistent constraint, although some constraints may compete with textual guidance when the target event requires subject attributes to change. As illustrated in Fig. 1(a), we observe three typical generation failures in existing methods: (1) Identity drift when textual guidance dominates. (2) Structural distortion when conflicting reference and textual demands disrupt coherent attribute evolution. (3) Reference copying when reference guidance dominates. These failure modes suggest that the key to resolving this competition lies not in simply strengthening or weakening overall reference constraints, but in whether all visual information from reference images should be treated as uniformly and persistently valid constraints throughout the generation process.

We further investigate why unified reference conditioning restricts event-driven attribute evolution through a controlled analysis of how reference and textual conditions guide the generation process. Our core hypothesis is that when the text requires a subject attribute to change while the reference image continues to impose its original visual constraints, the two conditions compete along the same attribute dimension. As shown in Fig. 2(a), we test this hypothesis by removing only the attribute description that contradicts the reference while preserving the reference image and the main textual semantics. The results reveal that the two guidance directions exhibit strong conflict when imposing opposing requirements on the same attribute, but become approximately orthogonal once the conflicting description is removed. This suggests that the interference between reference and textual conditions is primarily concentrated on specific attributes subject to opposing constraints, rather than arising from global semantic inconsistency. We term this phenomenon *attribute conflict*. More detailed analysis of attribute conflicts is provided in the supplementary material.

As illustrated in Fig. 2(b) and (d), such conflicts cause the reference condition to continuously interfere with the attribute changes required by the text, hindering the joint realization of personalization

preservation and event-driven evolution. ***This leads to a key insight: different attributes in reference images should not be assigned uniform reference validity. Persistent attributes should continue to provide personalization constraints, whereas the reference constraints on evolving attributes should be forgotten, allowing textual conditioning to govern their evolution.***

Building on this insight, we propose AttriForget, a training-free conflict-aware attribute forgetting framework. Given reference images and text prompt, AttriForget constructs an attribute graph that identifies target subjects and their attributes, distinguishing persistent from evolving attributes according to the target event. Based on this graph, we further introduce a selective attribute forgetting mechanism that probes orthogonal attribute-sensitive subspaces in the reference feature space, preserving effective representations of persistent attributes while forgetting the reference constraints on evolving attributes. This enables fine-grained attribute-level reference control, allowing reference conditioning to primarily preserve personalized characteristics while textual conditioning governs the dynamic evolution of evolving attributes, thereby forming approximately decoupled and complementary guidance. The entire framework requires no additional training and can be integrated as a plug-and-play module into different DiT-based video generation models.

In summary, the key contributions of our work are as follows:

- We identify an overlooked limitation of the prevailing reference-conditioning paradigm in customized video generation, revealing attribute conflict as a key obstacle to jointly achieving subject preservation and event-driven evolution.
- We formulate reference validity as a property of subject attributes conditioned on the target event and encode it in an attribute graph, supporting the transition from global reference conditioning to reference control at the attribute level.
- We propose a selective attribute forgetting mechanism that transforms reference and textual conditions into approximately decoupled and complementary guidance, reconciling subject preservation with the generative flexibility required for attribute evolution.

Extensive experiments demonstrate that AttriForget can be integrated into various DiT architectures, improving event-driven evolution and generation naturalness while enhancing subject consistency on most evaluated settings, with only 4% average inference latency overhead across backbones.

2 Related Work

Customized Video Generation. Customized video generation (Liang et al., 2025; Huang et al., 2025a) aims to synthesize videos that follow textual descriptions while preserving personalized subject characteristics from reference images. Early work, VideoBooth (Jiang et al., 2024), incorporates subject semantics and appearance details into video generation through coarse- and fine-grained visual embeddings and attention injection. Phantom (Liu et al., 2025) achieves consistent single- and multi-subject generation through joint text-image conditioning and cross-modal alignment, while Stand-In (Xue et al., 2026) enables lightweight identity control via restricted self-attention.

For more complex subject customization, ConceptMaster (Huang et al., 2025b) learns disentangled multi-concept embeddings to enable multi-concept video customization without test-time finetuning. BindWeave (Li et al., 2026b) leverages multimodal large language models to parse complex textual semantics and establish correspondences between textual instructions and visual subjects. RefAlign (Wang et al., 2026a) explicitly aligns reference-branch features with semantic representations from visual foundation models, enhancing subject consistency and semantic discriminability. Despite advances in reference feature modeling, cross-modal alignment, and subject disentanglement, these methods (Jiang et al., 2025; Chen et al., 2025; Kim et al., 2026) generally treat reference information as a unified whole and apply it persistently throughout generation, paying limited attention to the reference validity of different subject attributes under the target event.

Training-Free Video Control. Recent training-free video control methods (Kim et al., 2025; Xia et al., 2025) have improved generation quality and controllability by intervening in the inference process of pretrained models without additional training. For motion control, FlowMotion (Wang et al., 2026b) leverages latent predictions from flow-matching models to construct motion guidance for video motion transfer. MotiMotion (Hsin-Ying et al., 2026) combines vision-language reasoning with confidence-aware control to improve motion plausibility in complex interaction scenarios,

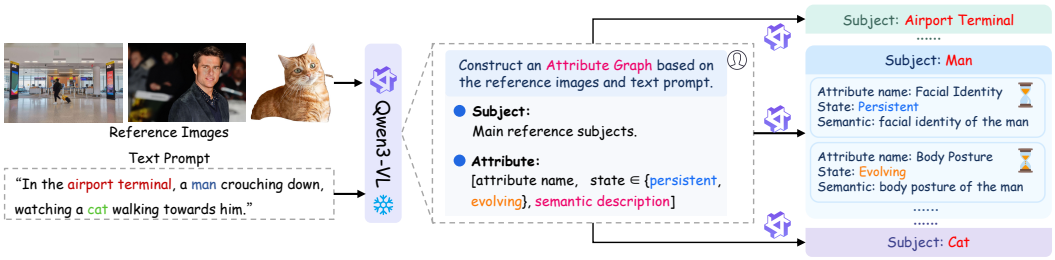


Figure 3: Attribute Graph. Given reference images and text prompt, a vision-language model decomposes each subject into attributes according to its characteristics and the target event, and determines the state of each attribute as persistent or evolving, together with its semantic description.

while ALG (Choi et al., 2026) applies adaptive low-pass filtering to conditioning images to alleviate motion suppression in image-to-video generation. For complex text-driven video generation, SwitchCraft (Xu et al., 2026) enables multi-event generation through event-aligned attention control. TS-Attn (Zhang et al., 2026a) improves semantic alignment and temporal coherence through temporal attention control, while NUMINA (Sun et al., 2026) enhances numerical alignment through attention analysis and layout guidance. In contrast, training-free customized video generation is still scarce. For example, TPIGE (Gao et al., 2025) alleviates the semantic gap between reference images and textual descriptions through input conditioning enhancement. Our analysis further characterizes this interference during generation, showing that it concentrates primarily on specific attributes subject to opposing constraints, rather than arising from global semantic inconsistency.

3 METHOD

Given reference images and a text prompt, we aim to resolve attribute conflict between reference and textual conditions, thereby reconciling subject preservation with the evolution prescribed by the target event. AttriForget comprises two components: an attribute graph and a selective attribute forgetting mechanism. The attribute graph models reference validity under the target event by distinguishing persistent from evolving attributes (Fig. 3). Guided by this graph, the forgetting mechanism maps attribute semantics to subspaces in the reference feature space, preserving representations of persistent attributes while selectively forgetting the reference constraints on evolving attributes (Fig. 4). Together, these components enable approximately decoupled and complementary guidance from reference and textual conditions without additional training.

3.1 ATTRIBUTE GRAPH

The validity of a reference attribute is inherently conditioned on the target event. For instance, the color of an object should remain unchanged under rigid motion, but becomes an evolving state when the prompt explicitly requests a color transition. We therefore avoid assigning each visual property a context-independent role. As illustrated in Fig. 3, given reference images $\mathcal{R} = \{I^{(m)}\}_{m=1}^{M}$ and a text prompt p, we employ a frozen vision–language model Φ_{VLM} to jointly interpret the visual evidence and textual intent, producing an event-conditioned attribute graph:

$$\mathcal{G}_p = \Phi_{\text{VLM}}(\mathcal{R}, p) = \big(\mathcal{V}_{\text{S}} \cup \mathcal{V}_{\text{A}}, \mathcal{E}_{\text{S}\to\text{A}}\big), \tag{1}$$

where $\mathcal{V}_{\text{S}}$ denotes the target-subject nodes, $\mathcal{V}_{\text{A}}$ denotes their visually grounded attribute nodes, and $\mathcal{E}_{\text{S}\to\text{A}}$ associates each attribute with its corresponding subject.

Each attribute node $a_i \in \mathcal{V}_{\text{A}}$ is represented by the structured tuple:

$$a_i = (n_i, \rho_i, r_i, d_i), \qquad \rho_i \in \{\text{P}, \text{E}\}, \quad r_i \in \{\texttt{ref_0}, \ldots, \texttt{ref_}(M-1), \texttt{all}\}, \tag{2}$$

where n_i is a canonical attribute name, ρ_i is its event-conditioned state, r_i identifies the reference image from which the attribute is grounded, and d_i is a concise semantic description of the attribute identity. The state ρ_i indicates whether the attribute is *persistent* (P) or *evolving* (E) under p.

Formally, the event partitions the subject attributes into two disjoint sets:

$$\mathcal{A}^{\text{P}}(p) = \{a_i \mid \rho_i = \text{P}\}, \qquad \mathcal{A}^{\text{E}}(p) = \{a_i \mid \rho_i = \text{E}\}, \qquad \mathcal{A}^{\text{P}}(p) \cap \mathcal{A}^{\text{E}}(p) = \emptyset. \tag{3}$$

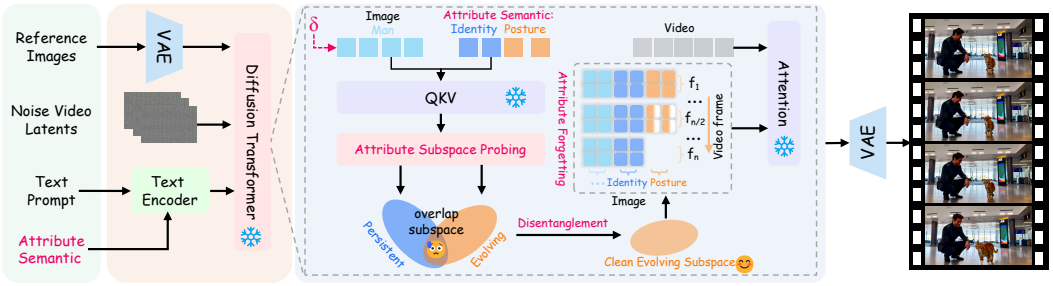


Figure 4: Overview of AttriForget. Three components enable attribute-level reference control within DiT attention: subspace probing identifies persistent and evolving attribute subspaces, subspace disentanglement removes their overlap, and attribute forgetting suppresses reference constraints on evolving attributes while preserving persistent information.

To avoid ambiguous semantic attribution, we require attribute scopes to be atomic and minimally overlapping. In particular, whenever a component contains both invariant appearance and a changing state, these factors are represented by separate nodes rather than absorbed into a coarse attribute. For a moving watch hand, for example, its color, material, and style constitute a persistent appearance node, whereas its orientation forms an evolving node. We exclude non-subject nuisance factors, including camera motion, lighting and sound, because they are not grounded in subject-reference features. The resulting graph is a compact representation of attribute-level reference validity and provides the semantic specification for the selective forgetting operation introduced next.

3.2 Selective Attribute Forgetting

Attribute Subspace Probing. While the attribute graph determines which attributes should be preserved or forgotten, it does not directly identify their representations within the entangled reference feature space. We therefore introduce an attribute subspace probing mechanism that maps attribute semantics to low-rank subspaces in the reference feature space. As illustrated in Fig. 4, the proposed mechanism operates on the reference features before attention key and value projection.

At layer ℓ, let $\boldsymbol{R}^\ell \in \mathbb{R}^{N_{\text{ref}} \times D}$ denote these reference features, where N_{ref} is the number of reference tokens and D is the feature dimension. For each attribute a_i, we encode its semantic description d_i using the frozen text encoder of the backbone, pool the token embeddings, and apply the existing text and query projections to obtain an attribute query $\boldsymbol{q}_i^{\ell,h}$ for attention head h. The reference identifier r_i selects the corresponding features $\boldsymbol{R}_i^\ell$. Thus, $\boldsymbol{q}_i^{\ell,h}$ conditions the attention response on the attribute semantics, while $\boldsymbol{R}_i^\ell$ provides the corresponding reference features.

To localize the feature subspace associated with each attribute, we introduce a perturbation $\boldsymbol{\delta} \in \mathbb{R}^D$ shared across the reference tokens, yielding $\boldsymbol{R}_i^\ell(\boldsymbol{\delta}) = \boldsymbol{R}_i^\ell + \mathbf{1}\boldsymbol{\delta}^\top$. Using the frozen key and value projections $\mathcal{K}_\ell^h$ and $\mathcal{V}_\ell^h$, the response to the attribute query is

$$\boldsymbol{u}_i^{\ell,h}(\boldsymbol{\delta}) = \text{softmax}\left(\frac{\boldsymbol{q}_i^{\ell,h}\mathcal{K}_\ell^h(\boldsymbol{R}_i^\ell(\boldsymbol{\delta}))^\top}{\sqrt{d_h}}\right)\mathcal{V}_\ell^h(\boldsymbol{R}_i^\ell(\boldsymbol{\delta})), \tag{4}$$

where d_h is the head dimension and $\boldsymbol{q}_i^{\ell,h}, \boldsymbol{u}_i^{\ell,h}(\boldsymbol{\delta}) \in \mathbb{R}^{d_h}$. The native query and key normalization is retained. To measure how strongly the selected reference features respond to attribute a_i, we compute the cosine similarity between the attention response $\boldsymbol{u}_i^{\ell,h}(\boldsymbol{\delta})$ and the attribute query $\boldsymbol{q}_i^{\ell,h}$ across attention heads. We then differentiate the score vector with respect to the perturbation to identify which directions in the reference feature space govern this response:

$$\boldsymbol{s}_i^\ell(\boldsymbol{\delta}) = \left[\cos\left(\boldsymbol{u}_i^{\ell,h}(\boldsymbol{\delta}), \boldsymbol{q}_i^{\ell,h}\right)\right]_h, \qquad \boldsymbol{J}_i^\ell = \left.\frac{\partial \boldsymbol{s}_i^\ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^\top}\right|_{\boldsymbol{\delta}=\mathbf{0}}. \tag{5}$$

For a small perturbation $\boldsymbol{\delta}$, the Jacobian $\boldsymbol{J}_i^\ell$ approximates the corresponding change in the attribute-response scores as $\boldsymbol{s}_i^\ell(\boldsymbol{\delta}) - \boldsymbol{s}_i^\ell(\mathbf{0}) \approx \boldsymbol{J}_i^\ell\boldsymbol{\delta}$. For any unit feature direction $\boldsymbol{v}$, the magnitude $\|\boldsymbol{J}_i^\ell\boldsymbol{v}\|_2$

measures how strongly perturbing the reference features along $\boldsymbol{v}$ changes the response to attribute a_i. The directions that maximize this response change are the leading right singular vectors of $\boldsymbol{J}_i^\ell$:

$$\boldsymbol{J}_i^\ell = \boldsymbol{U}_i^\ell \boldsymbol{\Sigma}_i^\ell (\boldsymbol{V}_i^\ell)^\top, \qquad \boldsymbol{B}_i^\ell = \boldsymbol{V}_i^\ell[:, 1{:}k], \tag{6}$$

where k is the rank budget. The columns of $\boldsymbol{B}_i^\ell$ therefore span the feature subspace associated with attribute a_i at layer ℓ. The basis is computed once per layer and cached across diffusion steps.

Attribute Subspace Disentanglement. Even with distinct semantic scopes, the probed subspaces of persistent and evolving attributes may contain shared feature directions. Directly suppressing an evolving subspace would also remove the persistent information carried by these directions. We therefore aggregate all persistent bases into a joint persistent basis. For each evolving attribute $a_i \in \mathcal{A}^{\mathrm{E}}(p)$, we then project $\boldsymbol{B}_i^\ell$ onto the orthogonal complement of this basis:

$$\boldsymbol{B}_{\mathrm{P}}^\ell = \mathrm{orth}\big([\boldsymbol{B}_j^\ell]_{a_j \in \mathcal{A}^{\mathrm{P}}(p)}\big), \qquad \widetilde{\boldsymbol{B}}_i^\ell = \mathrm{orth}\big(\big(\boldsymbol{I} - \boldsymbol{B}_{\mathrm{P}}^\ell (\boldsymbol{B}_{\mathrm{P}}^\ell)^\top\big)\boldsymbol{B}_i^\ell\big), \tag{7}$$

where $[\cdot]$ denotes column concatenation and $\mathrm{orth}(\boldsymbol{X})$ denotes an orthonormal basis for the column space of $\boldsymbol{X}$. The matrix $\boldsymbol{B}_{\mathrm{P}}^\ell (\boldsymbol{B}_{\mathrm{P}}^\ell)^\top$ projects onto the joint persistent subspace, so its complement removes all persistent components from $\boldsymbol{B}_i^\ell$. The columns of $\widetilde{\boldsymbol{B}}_i^\ell$ consequently span the clean evolving subspace and satisfy $(\widetilde{\boldsymbol{B}}_i^\ell)^\top \boldsymbol{B}_{\mathrm{P}}^\ell = \boldsymbol{0}$. Directions shared by persistent and evolving attributes remain in the persistent subspace and are preserved during forgetting.

Attribute Forgetting. We perform attribute forgetting by progressively removing the reference components associated with clean evolving subspaces. Since columns of $\widetilde{\boldsymbol{B}}_i^\ell$ are orthonormal, $\boldsymbol{R}^\ell \widetilde{\boldsymbol{B}}_i^\ell (\widetilde{\boldsymbol{B}}_i^\ell)^\top$ is the projection of the reference features onto the clean subspace of evolving attribute a_i. Let $\boldsymbol{\Gamma}_i \in \{0,1\}^{N_{\mathrm{ref}} \times N_{\mathrm{ref}}}$ be a diagonal token mask that selects the reference tokens specified by r_i, and let $\lambda_i(f) \in [0,1]$ denote the temporal forgetting strength. For a video query at frame f, we define the modulated reference features as

$$\widehat{\boldsymbol{R}}^\ell(f) = \boldsymbol{R}^\ell - \sum_{a_i \in \mathcal{A}^{\mathrm{E}}(p)} \lambda_i(f)\, \boldsymbol{\Gamma}_i \boldsymbol{R}^\ell \widetilde{\boldsymbol{B}}_i^\ell (\widetilde{\boldsymbol{B}}_i^\ell)^\top. \tag{8}$$

We set $\lambda_i(f) = \mathrm{clip}_{[0,1]}((f - f_{0,i})/(f_{1,i} - f_{0,i}))$, where $f_{0,i} = 0$ denotes the first frame and $f_{1,i} = 0.3(F-1)$ is chosen empirically, with F the total frame count, $\lambda_i = 0$ retains and $\lambda_i = 1$ removes the clean evolving component. Here, f denotes the temporal frame within the generated video, so the forgetting strength increases along the video timeline. To avoid duplicate attenuation, we sequentially orthogonalize the clean evolving bases in a fixed attribute order before applying Eq. 8.

The modulated reference features are then passed through the original key and value projections and used by the native attention operator. As $\lambda_i(f)$ increases, reference components in the clean evolving subspace of attribute a_i are progressively attenuated, reducing the corresponding information available to video queries through attention. Textual conditioning can then guide attribute evolution, while the persistent components from the reference remain available for subject personalization.

4 Experiments

4.1 Experimental settings

Datasets and Metrics. We evaluate AttriForget on the Open Domain and Human Domain subsets of OpenS2V-Eval (Yuan et al., 2025), comprising 180 and 60 subject and prompt pairs, respectively. The Open Domain subset covers diverse human and nonhuman subjects in scenarios with one or multiple subjects, whereas the Human Domain subset focuses on human appearance and identity fidelity. Following the official OpenS2V-Nexus evaluation protocol, we report Aesthetics, Motion Smoothness, Motion Amplitude, FaceSim, GmeScore, NexusScore, and NaturalScore. NexusScore is reported only for the Open Domain subset. All metrics are normalized and combined into TotalScore using the official normalization rules and the aggregation weights for each subset.

Implementation Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the vision-language model to construct attribute graphs from reference images and text prompts. To assess applicability across different architectures and parameter scales, we integrate AttriForget into Phantom-1.3B (Liu et al., 2025), VACE-1.3B (Jiang et al., 2025), the 5B model of Kim et al. (Kim et al., 2026), and

Table 1: Quantitative comparison across different methods on OpenS2V-Eval.

Method	Total↑	Aes↑	M. Smo.↑	M. Amp.↑	FaceSim↑	Gme↑	Nexus↑	Natural↑
OpenS2V-Eval: Open Domain								
Kling1.6	56.23%	44.59%	86.93%	41.60%	40.10%	66.20%	45.89%	74.59%
Saber	57.91%	42.42%	96.12%	21.12%	49.89%	67.50%	47.22%	72.55%
MAGREF	52.51%	45.02%	93.17%	21.81%	30.83%	70.47%	43.04%	66.90%
RefAlign	56.30%	42.96%	94.74%	20.74%	53.06%	66.85%	43.97%	66.25%
VACE-1.3B	49.89%	48.24%	97.20%	18.83%	20.57%	71.26%	37.91%	65.46%
+ AttriForget	**52.17%**	**48.49%**	**97.72%**	**22.96%**	**20.62%**	70.70%	**39.23%**	**73.46%**
Phantom-1.3B	54.89%	46.67%	93.30%	14.29%	48.56%	69.43%	42.48%	62.50%
+ AttriForget	**56.29%**	**47.28%**	**93.92%**	**20.51%**	45.30%	**70.51%**	40.67%	**70.95%**
Kim et al. (5B)	45.02%	34.95%	81.21%	20.38%	27.76%	64.08%	42.91%	51.37%
+ AttriForget	**47.08%**	**35.89%**	**82.90%**	**21.89%**	**28.55%**	**65.07%**	**44.97%**	**55.90%**
Phantom-14B	56.77%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	69.35%
+ AttriForget	**59.73%**	**47.41%**	**96.39%**	23.50%	**57.05%**	70.50%	**40.13%**	**75.00%**
OpenS2V-Eval: Human Domain								
VACE-1.3B	54.90%	53.18%	95.84%	16.87%	22.29%	73.61%	–	65.28%
+ AttriForget	**58.18%**	53.08%	**99.28%**	**17.53%**	**27.56%**	72.79%	–	**71.19%**
Phantom-1.3B	60.00%	50.80%	92.02%	14.09%	46.29%	72.17%	–	65.83%
+ AttriForget	**61.08%**	49.86%	92.00%	13.87%	**50.37%**	**72.67%**	–	**66.38%**

Phantom-14B (Liu et al., 2025). For every backbone, AttriForget is inserted into the DiT attention modules that carry reference conditioning and modulates the reference features before key and value projection. All other inference settings remain identical to those of the original model. AttriForget operates entirely at inference time, with all pretrained parameters frozen and no additional training or parameter updates. All experiments are conducted on 8 NVIDIA H20 GPUs.

Baseline Models. Beyond evaluating AttriForget on the aforementioned backbones, we compare it with representative recent subject-to-video (S2V) models, including the open-source methods MAGREF (Deng et al., 2025) and RefAlign (Wang et al., 2026a), and the closed-source model Kling1.6 (kling, 2024) and Saber (Zhou et al., 2026).

4.2 Quantitative Comparisons

As shown in Table 1, AttriForget improves NaturalScore and TotalScore in every evaluated backbone and subset setting, with NaturalScore gains of up to 8.45 percentage points on Open Domain. It improves FaceSim, NexusScore, and NaturalScore on VACE-1.3B, Kim et al. (5B), and Phantom-14B. On Phantom-14B, these metrics increase by 5.59, 2.70, and 5.65 percentage points, respectively, yielding the highest NaturalScore of 75.00% and TotalScore of 59.73% on Open Domain. FaceSim and NaturalScore improve together for both evaluated backbones on Human Domain. These gains support the design of AttriForget: reference images continue to constrain persistent attributes for identity preservation and subject consistency, while textual conditioning governs evolving attributes after their conflicting reference constraints are removed, improving video naturalness. Results across backbones with different architectures and parameter scales validate the plug-and-play capability of AttriForget, which incurs only 4% average additional inference latency as shown in Fig. 1(d).

4.3 Qualitative Comparisons

Across the diverse human and object scenarios in Fig. 5, AttriForget produces more natural videos with coherent structures, continuous motion, and faithful event execution while preserving persistent personalized attributes. By using reference images to anchor persistent attributes and text to govern evolving ones, AttriForget enables appearance preservation alongside changes in pose, clothing, and object state. For example, the second row depicts a book, a cup, and an ashtray containing a cigarette. The prompt requires preserving the book and cup appearances, increasing and then stabilizing the liquid level, and replacing the ashtray with a black plate holding a fork and cake. AttriForget pre-

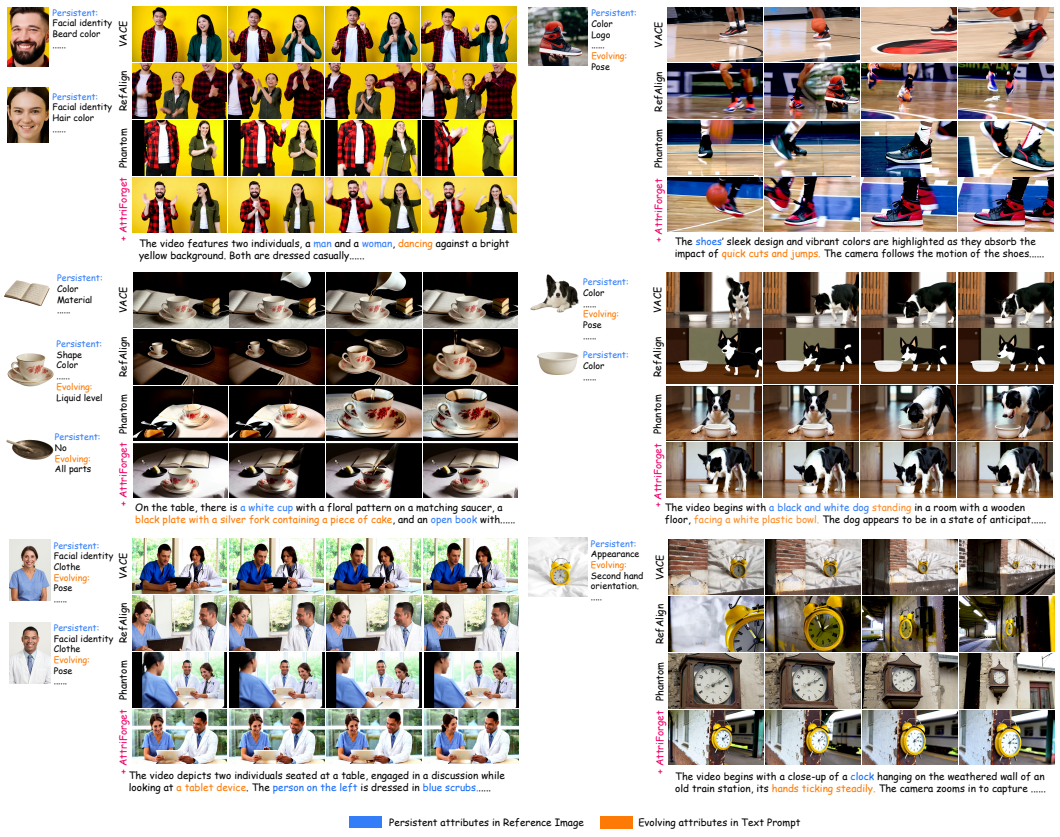


Figure 5: Qualitative comparison results on the OpenS2V-Eval dataset. The results labeled "+AttriForget" are obtained by applying AttriForget to the corresponding VACE or Phantom backbone.

serves the required appearances while discarding obsolete reference constraints on the liquid level and ashtray, allowing the prescribed evolution. Competing methods retain outdated reference content or fail to complete the specified changes. These results highlight the importance of separating attribute preservation from evolution rather than uniformly retaining reference information.

4.4 User Study

We further conducted a human evaluation comparing AttriForget with its unmodified Phantom-14B backbone on the Open Domain subset of OpenS2V-Eval. We recruited 20 evaluators (15 men and 5 women) of varied ages and educational backgrounds. Each evaluator independently received 20 randomly sampled cases from the Open Domain results, so the evaluated sets differed across participants. For each case, participants rated both videos on a five-point Likert scale for subject consistency (SC), action naturalness (AN), text-video alignment (TVA), and overall visual quality (OVQ). We compared the paired scores in each dimension and report the aggregated win, draw, and loss rates for AttriForget in Fig. 1(c). AttriForget wins more often than it loses in all four dimensions, particularly action naturalness, supporting its perceptual improvements. Results for additional backbones and further details are provided in the supplementary material.

4.5 Ablation Studies

All ablation studies are conducted using VACE-1.3B on the Human Domain.

Seed Search and Global Reference Strength. We examine whether the gains of AttriForget can be reproduced by searching random seeds in the baseline or uniformly adjusting its reference strength. Table 2 shows that seed search achieves a NaturalScore of 67.08% and a TotalScore of 56.68%. Global reference strengths of 0.5, 0.75, and 1.5 yield at most 69.16% in NaturalScore and 56.24%

in TotalScore. Neither control reproduces the joint improvements in naturalness and motion quality obtained by AttriForget in Table 1. These results suggest that favorable sampling or global strength adjustment alone cannot account for the observed gains. This further demonstrates that reference visual information should not be treated as a unified condition that persists throughout generation.

Table 2: Seed search and global reference strength ablations compared with AttriForget.

Method	Total↑	Aes↑	M. Smo.↑	M. Amp.↑	FaceSim↑	Gme↑	Nexus↑	Natural↑
Random seed search								
Seed search	56.68%	53.04%	99.22%	11.55%	27.16%	72.92%	–	67.08%
Global reference strength								
0.5	55.41%	51.75%	95.17%	13.37%	24.32%	73.11%	–	66.94%
0.75	56.15%	53.19%	96.21%	13.11%	23.40%	72.88%	–	69.16%
1.5	56.24%	53.44%	95.07%	11.06%	26.64%	73.16%	–	67.01%
AttriForget	**58.18%**	53.08%	**99.28%**	**17.53%**	**27.56%**	72.79%	–	**71.19%**

Subspace Probing. Table 3 evaluates the probing strategy and subspace rank. The counterfactual variant directly uses projected text differences between target events and their corresponding counterfactuals as basis directions for the evolving attribute subspace. Replacing probing based on attention responses with a counterfactual event formulation reduces TotalScore from 58.18% to 57.24%, supporting the benefit of deriving attribute subspaces from native attention responses. Among the tested ranks, rank 4 achieves the highest NaturalScore and TotalScore, while both lower and higher ranks yield smaller motion amplitude. This nonmonotonic trend highlights the need for appropriate subspace capacity rather than simply increasing rank.

Subspace Disentanglement. Removing attribute subspace disentanglement reduces FaceSim from 27.56% to 21.49%. This result supports removing persistent components from evolving subspaces before forgetting, so that persistent identity information is better preserved.

Attribute Forgetting. We further ablate the forgetting schedule. Immediate forgetting removes the evolving components outright instead of applying linear temporal forgetting. NaturalScore decreases from 71.19% to 67.91%, showing that progressive release better reflects the forgetting process, while immediate forgetting may discard relevant information at the beginning of the video.

Table 3: Design and subspace rank ablations of AttriForget.

Method	Total↑	Aes↑	M. Smo.↑	M. Amp.↑	FaceSim↑	Gme↑	Nexus↑	Natural↑
Design variants								
Counterfactual event	57.24%	53.27%	99.23%	15.79%	24.37%	72.83%	–	70.76%
Immediate forgetting	56.98%	53.12%	97.32%	16.35%	27.50%	72.77%	–	67.91%
w/o disentanglement	56.53%	52.89%	98.08%	17.15%	21.49%	73.63%	–	70.83%
Subspace rank								
1	56.96%	52.81%	99.25%	11.68%	27.09%	72.87%	–	68.22%
2	57.44%	53.04%	99.21%	15.70%	27.29%	72.80%	–	69.17%
8	57.31%	52.91%	99.05%	13.76%	27.77%	72.84%	–	68.64%
AttriForget (rank=4)	**58.18%**	53.08%	**99.28%**	**17.53%**	27.56%	72.79%	–	**71.19%**

5 Conclusion

In this paper, we uncover an overlooked attribute conflict in the prevailing paradigm of video generation conditioned on reference images and propose AttriForget to address it. AttriForget constructs an attribute graph that explicitly distinguishes persistent and evolving attributes in reference information under the target event. Through attribute subspace probing and selective attribute forgetting, it progressively relaxes reference constraints on evolving attributes, allowing text to guide their evolution while reference features preserve persistent attributes. Experiments across backbones with different architectures and parameter scales validate the effectiveness and broad applicability of AttriForget. This work offers a new perspective on reference conditioning in video generation.

AI USE STATEMENT

We used generative AI tools to improve the readability of the manuscript. The authors reviewed and verified the edited content.

ETHICS STATEMENT

AttriForget improves the fidelity of reference-conditioned video generation. This capability could be misused to create misleading videos of identifiable individuals without their consent. We encourage the use of reference images only with appropriate rights and consent. For the human evaluation, we recruited 20 evaluators through an open call. They were informed about the study, voluntarily consented to participate, and received no compensation. We report their ratings in aggregate.

REPRODUCIBILITY STATEMENT

The paper provides the details needed to reproduce AttriForget. Section 3 defines the event-conditioned attribute graph, attribute subspace probing, persistent-aware disentanglement, and selective reference forgetting. Section 4 specifies the datasets, evaluation metrics, vision-language model, video-generation backbones, and inference settings. All parameters remain frozen, and the method is applied only during inference. All experiments are conducted on NVIDIA H20 GPUs. Code, model configurations, evaluation scripts, and instructions for reproducing the reported results will be released.

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*, 2025a.

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

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

June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Improving motion in image-to-video models via adaptive low-pass guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40364–40375, 2026.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation with subject disentanglement. *arXiv preprint arXiv:2505.23742*, 2025.

Jiayi Gao, Changcheng Hua, Qingchao Chen, Yuxin Peng, and Yang Liu. Identity-preserving text-to-video generation via training-free prompt, image, and guidance enhancement. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13751–13757, 2025.

Qin Guo, Tianyu Yang, Xuanhua He, Fei Shen, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, and Dan Xu. Wildactor: Unconstrained identity-preserving video generation. *arXiv preprint arXiv:2603.00586*, 2026.

Lee Hsin-Ying, Hanwen Jiang, Yiqun Mei, Jing Shi, Ming-Hsuan Yang, and Zhixin Shu. MotiMotion: Motion-controlled video generation with visual reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

Yi Huang, Wei Xiong, He Zhang, Chaoqi Chen, Jianzhuang Liu, Mingfu Yan, and Shifeng Chen. Dive: Taming dino for subject-driven video editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16004–16014. IEEE, 2025a.

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

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.

Daneul Kim, Jingxu Zhang, Wonjoon Jin, Sunghyun Cho, Qi Dai, Jaesik Park, and Chong Luo. Learning zero-shot subject-driven video generation using 1% compute. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Jaemin Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Free2Guide: Training-free text-to-video alignment using image LVLM. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17920–17929. IEEE, 2025.

kling. Image to video elements feature. `https://klingai.com/image-to-video/multi-id/new/`, 2024.

Wei Li, Yiheng Zhang, Fuchen Long, Zhaofan Qiu, Ting Yao, Xiaoyan Sun, and Tao Mei. Reactid: Synchronizing realistic actions and identity in personalized video generation. In *International Conference on Learning Representations*, volume 2026, pp. 62463–62485, 2026a.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. BindWeave: Subject-consistent video generation via cross-modal integration. In *International Conference on Learning Representations*, volume 2026, pp. 107960–107979, 2026b.

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.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8543–8552. IEEE, 2024.

Zhengyang Sun, Yu Chen, Xin Zhou, Xiaofan Li, Xiwu Chen, Dingkang Liang, and Xiang Bai. When numbers speak: Aligning textual numerals and visual instances in text-to-video diffusion models. *arXiv preprint arXiv:2604.08546*, 2026.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhen Wang, Youcan Xu, Jun Xiao, and Long Chen. Flowmotion: training-free flow guidance for video motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38143–38153, 2026b.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15982–15993. IEEE, 2025.

Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. Switchcraft: Training-free multi-event video generation with attention controls. *arXiv preprint arXiv:2602.23956*, 2026.

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.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Chongyang Ma, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hongyu Zhang, Yufan Deng, Zilin Pan, Peng-Tao Jiang, Bo Li, Qibin Hou, Zhen Dong, Zhiyang Dou, and Daquan Zhou. Ts-attn: Temporal-wise separable attention for multi-event video generation. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Yiheng Zhang, Zhaofan Qiu, Zunxu Liu, Yingwei Pan, Ting Yao, and Tao Mei. Evoid: Reinforced evolution for identity-preserving video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41870–41880, 2026b.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam-Woh Ng, Tian Xie, et al. Scaling zero-shot reference-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9253–9262, 2026.

A Details of Attribute Conflicts

This section details the computation underlying the analysis of attribute conflict in Fig. 2(a), then examines an asymmetry between appearance and motion attributes. The additional examples show that conflicts are more pronounced for appearance changes, while a counterexample involving pose clarifies why attribute conflict is not restricted to appearance.

A.1 Conflict Measurement

We use *attribute conflict* to describe incompatible requirements on a subject attribute imposed by the reference condition and the target prompt. The guidance measure below provides an aggregate directional diagnostic that complements the controlled prompt comparison and qualitative generation results.

Computing Conditional Guidance. Given the reference images $\mathcal{R}$ and prompt p, let $\boldsymbol{r}$ denote the encoded reference latents and $\boldsymbol{c}_p$ the text encoding. At sampling iteration s, the latent of the generated video is $\boldsymbol{z}_s$ and the model timestep is t_s. We denote the model prediction for the generated video by $F_\theta(\boldsymbol{z}_s, t_s; \boldsymbol{r}, \boldsymbol{c}_p)$. To separate the influence of each condition, we evaluate three predictions at the same latent state and timestep:

$$\begin{aligned} \boldsymbol{f}_s^0 &= F_\theta(\boldsymbol{z}_s, t_s; \boldsymbol{0}, \boldsymbol{c}_-), \\ \boldsymbol{f}_s^{\mathrm{I}} &= F_\theta(\boldsymbol{z}_s, t_s; \boldsymbol{r}, \boldsymbol{c}_-), \\ \boldsymbol{f}_s^{\mathrm{T}} &= F_\theta(\boldsymbol{z}_s, t_s; \boldsymbol{0}, \boldsymbol{c}_p). \end{aligned} \tag{9}$$

Here, $\boldsymbol{0}$ is a zero reference latent with the same shape as $\boldsymbol{r}$, and $\boldsymbol{c}_-$ encodes the fixed negative prompt. The baseline $\boldsymbol{f}_s^0$ is shared by the prediction conditioned on reference images, $\boldsymbol{f}_s^{\mathrm{I}}$, and the prediction conditioned on text, $\boldsymbol{f}_s^{\mathrm{T}}$. Subtracting this baseline measures the marginal prediction change under each condition:

$$\boldsymbol{D}_s^{\mathrm{I}} = s_{\mathrm{I}}\big(\boldsymbol{f}_s^{\mathrm{I}} - \boldsymbol{f}_s^0\big), \qquad \boldsymbol{D}_s^{\mathrm{T}} = s_{\mathrm{T}}\big(\boldsymbol{f}_s^{\mathrm{T}} - \boldsymbol{f}_s^0\big), \tag{10}$$

where s_{I} and s_{T} are the image and text guidance scales. We collect these paired fields before latent updates at selected sampling iterations to examine the interaction between the two conditions over generation.

Aggregating Local Directions. To summarize guidance within local regions, we apply spatial adaptive average pooling to each field $\boldsymbol{D}_s^b$, where $b \in \{\mathrm{I}, \mathrm{T}\}$, retaining its temporal and channel dimensions. Each pooled region yields a vector $\boldsymbol{v}_n^b \in \mathbb{R}^C$, where C is the latent channel dimension and n indexes the recorded iteration, latent frame, and spatial region. We normalize each vector to compare its orientation with equal weight across locations, then average over all N recorded vectors to obtain the mean direction for each condition:

$$\widehat{\boldsymbol{v}}_n^b = \frac{\boldsymbol{v}_n^b}{\|\boldsymbol{v}_n^b\|_2 + \epsilon}, \qquad \boldsymbol{\mu}^b = \frac{1}{N}\sum_{n=1}^{N} \widehat{\boldsymbol{v}}_n^b, \qquad \epsilon > 0. \tag{11}$$

Comparing Mean Guidance Directions. To characterize the aggregate directional relationship between the two conditions, we compare their mean normalized guidance vectors through cosine similarity and the corresponding angle:

$$\begin{aligned} \gamma_{\mathrm{I,T}} &= \cos\big(\boldsymbol{\mu}^{\mathrm{I}}, \boldsymbol{\mu}^{\mathrm{T}}\big) = \frac{(\boldsymbol{\mu}^{\mathrm{I}})^\top \boldsymbol{\mu}^{\mathrm{T}}}{\|\boldsymbol{\mu}^{\mathrm{I}}\|_2 \|\boldsymbol{\mu}^{\mathrm{T}}\|_2}, \\ \alpha_{\mathrm{I,T}} &= \frac{180}{\pi} \arccos\big(\gamma_{\mathrm{I,T}}\big). \end{aligned} \tag{12}$$

Positive and negative values of $\gamma_{\mathrm{I,T}}$ indicate aligned and opposing components, respectively, in the mean guidance directions. A value near zero indicates approximate orthogonality of these means. We interpret this statistic together with the controlled comparison in Fig. 2(a), where the reference image and main event description are retained while the conflicting attribute description is removed. The resulting shift toward zero is consistent with competing attribute requirements affecting the aggregate guidance relationship. The attribute-level interpretation is supported by this targeted prompt edit and the qualitative failures, while the centroid statistic summarizes the associated directional change.

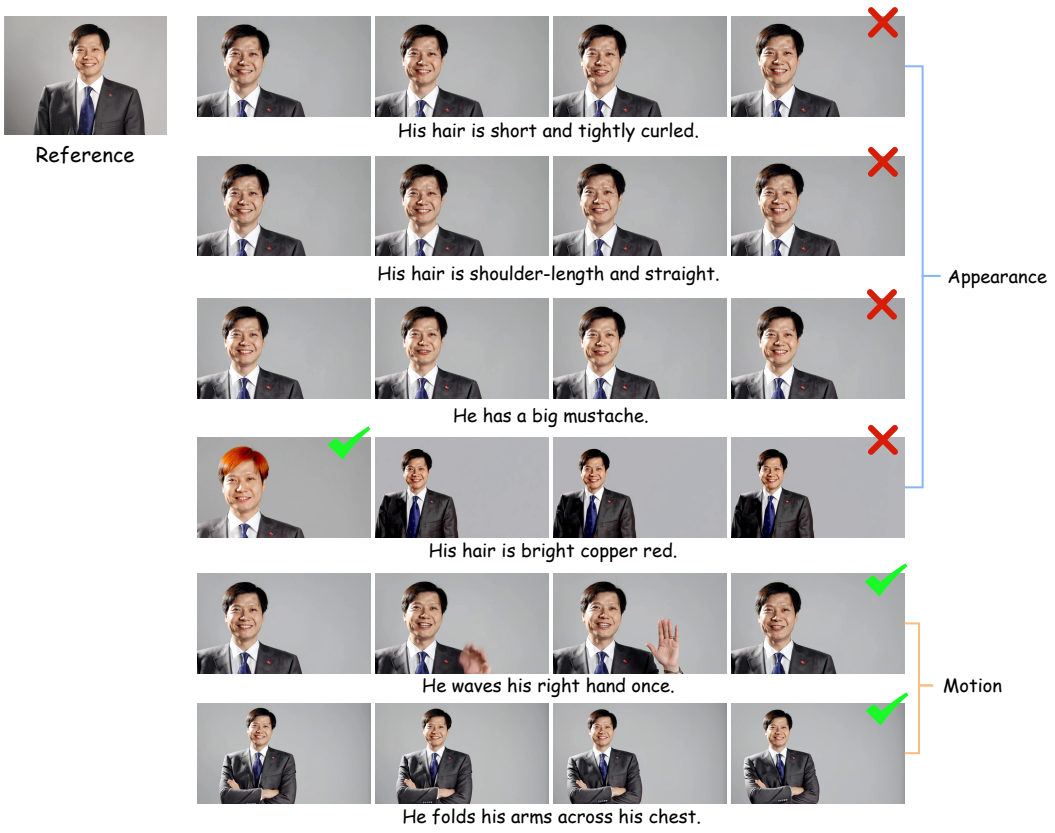


Figure 6: Separate prompts request either appearance changes or actions using the same reference image from OpenS2V-Eval (Yuan et al., 2025). Requests to change hair texture, hair length, facial hair, or hair color largely yield reference copying. In the copper red hair example, the requested color appears in the first displayed frame but reverts to the reference color in subsequent frames. In the action cases, waving and crossing the arms are successfully generated according to the respective prompts. These examples show that the model follows action prompts but fails to realize the requested appearance changes.

A.2 Interesting Observations

Qualitative Observation. Fig. 6 compares separate prompts requesting either appearance changes or actions using the same reference image. Requests for tightly curled hair, straight hair reaching the shoulders, or a large mustache result in the reference appearance being retained, producing a *copy and paste* effect. Even when copper red hair appears in the first displayed frame, subsequent frames revert to the original hair color, showing that the requested change is not sustained. In the action cases, the model successfully generates waving and arm crossing according to the respective prompts. Across these separate cases, the model succeeds on action prompts but fails on appearance prompts, where the identity constraints imposed by the reference image prevent the requested changes.

Quantitative Evidence. Fig. 7 supports this observation using the guidance measure in Sec. A.1. Across the tested prompts, appearance changes involving hair, facial hair, and clothing yield strongly negative cosine similarities between the mean reference and text guidance directions, approximately between -0.7 and -0.4. Motion instructions, including nodding, turning the head or body, crossing the arms, and waving, yield values closer to zero, approximately between -0.2 and 0. Within this prompt set, appearance-change prompts therefore exhibit stronger opposition between the mean guidance directions than action prompts, consistent with the qualitative failures observed for appearance changes.

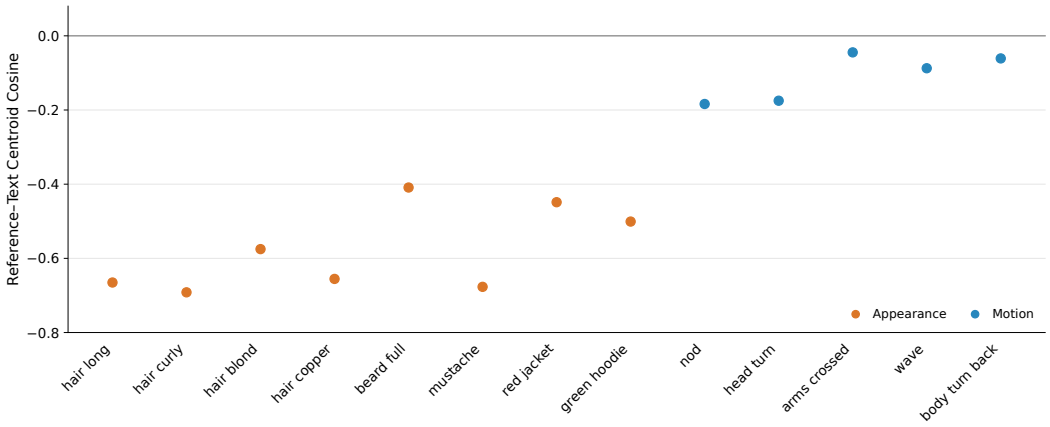


Figure 7: Cosine similarity between the mean reference and text guidance directions for appearance and motion prompts. The centroid cosine corresponds to $\gamma_{\mathrm{I,T}}$ in Eq. 12. Appearance prompts exhibit substantially more negative values than motion prompts, indicating stronger opposition between the mean guidance directions. Values closer to zero for motion prompts indicate weaker opposition between these means and do not establish the absence of local conflicts.

A.3 Discussion and Implications

A Counterexample Involving Pose. The dog example in Fig. 5 (second row, right) shows that conflict can also affect pose and action. The reference depicts a dog lying down, while the prompt specifies that the video begins with the dog standing. The Phantom baseline nevertheless reproduces the lying pose in its initial displayed frames. This failure is consistent with the model retaining a static reference pose as a constraint, even though the target event requires it to change.

Why Appearance Conflicts Are More Pronounced. We interpret this asymmetry as a consequence of which information the model extracts from the reference. When reference conditioning mainly retains appearance cues associated with identity, these cues can compete with appearance changes requested by the text. When it also retains a static pose, a requested action or pose change can face the same competition. A plausible explanation for the observed prevalence of appearance conflicts is that the S2V task and its training data encourage models to extract information that preserves identity from reference images. Meanwhile, S2V models fine-tuned from T2V backbones inherit strong capabilities for generating motion that follows text instructions. Together, these tendencies may make action conflicts less frequent than appearance conflicts.

Implication for Reference Conditioning. These observations reinforce our central premise: reference images should constrain only persistent attributes under the target event, while text should govern evolving attributes, including both actions and explicitly requested appearance changes. Appearance is therefore not inherently persistent, nor is motion inherently free of reference constraints. The video generation model should therefore avoid attending to reference attributes that conflict with the changes required by the text. AttriForget implements this principle by selectively and progressively suppressing the corresponding components in reference features while preserving information about persistent attributes. This selective forgetting allows reference and textual conditions to complement each other, preserving subject consistency while enabling the evolution specified by the prompt.

B Details of Attribute Graph

B.1 An Example of the Attribute Graph

Our attribute graph represents the validity of reference attributes under the target event. It links each subject to its visually grounded attributes and classifies them as persistent or evolving according to the text prompt. Persistent attributes provide reference constraints for subject preservation, while the constraints on evolving attributes are progressively relaxed to allow the requested changes.

If all subject attributes in the reference images are persistent under the target event ($\mathcal{A}^{\mathrm{E}}(p) = \emptyset$), there is no attribute conflict and AttriForget reduces to the original backbone with unchanged reference features (Eq. 8), avoiding unnecessary suppression of valid reference information.

Fig. 8 gives an example for a man reading a book and occasionally glancing up. The graph contains one subject grounded in `ref_0`, with five persistent and three evolving attributes.

Prompt. The video shows a man sitting on a park bench under a large oak tree, reading a book. He has a beard and is wearing a casual sweater and jeans. The park is quiet and green, with sunlight filtering through the tree branches. The man seems completely absorbed in his book, occasionally glancing up to enjoy the peaceful surroundings.



Reference

Shared fields: `subject`: `man`; `reference_id`: `ref_0`.

`name`	`semantic description`	`value`†
Persistent (`mode`: `persistent`)		
`hair color`	color of the man hair	`brown`
`skin tone`	skin tone of the man	`light`
`eye color`	color of the man eye	`light brown`
`upper body clothing color`	color of the man upper body clothing	`blue`
`upper body clothing style`	upper body clothing style of the man	`henley sweater`
Evolving (`mode`: `evolving`)		
`facial hair`	facial hair of the man	`none`
`expression`	expression of the man	`neutral`
`head pose`	pose of the man head	`looking forward`

`name`	`target behavior`† *(reference only)*
`facial hair`	add a beard from the beginning and maintain it throughout the video
`expression`	change subtly to reflect reading and glancing up
`head pose`	tilt slightly upward when glancing up

Figure 8: An attribute graph example with its reference image and text prompt. The group headings specify `mode` for each node. Fields marked with † are auxiliary annotations for reference only and are not used during video-generation inference.

Interpreting the Graph. Each node follows the tuple (n_i, ρ_i, r_i, d_i) in Eq. 2. Its `name` identifies the attribute, and `subject` links it to the subject node. The `mode` field specifies ρ_i and assigns the node to the persistent or evolving set. During inference, `semantic_description` (d_i) is encoded to construct the attribute query, while `reference_id` (r_i) selects its reference features. Here, facial hair evolves because the prompt requests a beard, while expression and head pose evolve with reading and glancing up. Hair color, skin tone, eye color, and clothing color and style remain persistent under this event.

The graph additionally records `value`, describing the observed reference state, and `target_behavior`, describing the intended evolution. Both fields serve only as additional reference annotations for the attribute graph and are unused throughout video-generation inference. Neither is encoded into attribute queries, supplied as additional conditioning, or used to determine the forgetting schedule. For example, the facial-hair query uses "facial hair of the man," without encoding either "none" or the instruction to add a beard. The original prompt provides the generation guidance for the requested beard and actions.

B.2 Attribute Graph Quality

We investigate the quality of the generated attribute graphs and its effect on video generation. For 60 cases from the Human Domain subset of OpenS2V-Eval, we construct attribute graphs from the reference images and text prompts using three generations of Qwen-VL: Qwen2-VL-7B-Instruct (Wang et al., 2024), Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), and Qwen3-VL-8B-Instruct (Bai et al., 2025a). We evaluate the resulting graphs with GPT-4o and apply each set of graphs to AttriForget with VACE-1.3B to assess how graph quality affects the generated videos.

GPT-4o assigns a score from 1 to 5 to each graph on four dimensions, with higher scores indicating better quality. Attribute Completeness (AC) measures coverage of key subject attributes relevant to subject preservation and the target event. Subject–Attribute Matching (SAM) evaluates whether each attribute is associated with the correct subject. Attribute State Correctness (ASC) assesses whether persistent and evolving states are assigned correctly under the target event. Semantic Description Correctness (SDC) evaluates whether each semantic description accurately and unambiguously identifies its attribute without conflating its identity with a specific value or evolution instruction. These dimensions assess attribute coverage, subject association, reference validity, and semantic specification, respectively. The auxiliary `value` and `target_behavior` fields are excluded from the quality criteria because they do not participate in video generation inference.

Table 4: GPT-4o evaluation of attribute graph quality on 60 Human Domain cases. Each dimension is scored from 1 to 5 and averaged across cases. Total denotes the mean of the four dimensions. All graph generators use their Instruct variants. Bold indicates the best score in each column.

Graph generator	Total↑	AC↑	SAM↑	ASC↑	SDC↑
Qwen2-VL (7B)	4.07	3.89	4.26	3.84	4.29
Qwen2.5-VL (7B)	4.17	3.84	4.41	3.98	4.46
Qwen3-VL (8B)	**4.43**	**4.12**	**4.80**	**4.25**	**4.57**

As shown in Table 4, the overall graph quality score increases from 4.07 to 4.17 and then to 4.43 across the three models. Qwen3-VL achieves the highest score in all four dimensions, including an attribute state correctness score of 4.25, compared with 3.84 and 3.98 for the earlier models. This result is consistent with the advances in multimodal reasoning, understanding of fine visual details, and complex instruction following reported for the Qwen3-VL family (Bai et al., 2025a). These capabilities support interpreting visual evidence together with textual intent and expressing the resulting decisions as structured attribute graphs.

Table 5: Effect of the graph generator on video generation with AttriForget and VACE-1.3B on Human Domain. All graph generators use their Instruct variants. Metric names and normalization follow Table 1. Bold indicates the best score in each column.

Graph generator	Total↑	Aes↑	M. Smo.↑	M. Amp.↑	FaceSim↑	Gme↑	Nexus↑	Natural↑
Qwen2-VL (7B)	57.21%	52.34%	98.89%	17.14%	25.52%	72.62%	–	70.33%
Qwen2.5-VL (7B)	57.85%	**53.84%**	99.03%	17.01%	26.79%	72.37%	–	70.60%
Qwen3-VL (8B)	**58.18%**	53.08%	**99.28%**	**17.53%**	**27.56%**	**72.79%**	–	**71.19%**

Table 5 shows that higher overall graph quality is accompanied by better video generation performance. As the graph quality score rises from 4.07 to 4.43, TotalScore increases from 57.21% to 58.18%, NaturalScore from 70.33% to 71.19%, and FaceSim from 25.52% to 27.56%. All three metrics improve across the three graph generators. Individual metrics need not improve monotonically. Nevertheless, the joint gains in overall quality, naturalness, and face similarity support the importance of accurately separating persistent and evolving attributes. A more accurate graph allows reference features to constrain persistent attributes while textual conditioning governs evolving ones, enabling complementary guidance that better reconciles subject preservation with the changes required by the target event.

C Analysis of Selective Attribute Forgetting

We examine the semantic relevance of the subspaces identified by the Jacobian probe through attention visualization and generation results. Fig. 9 presents examples with and without attribute conflict using Phantom-1.3B.

With Attribute Conflict. The prompt requires the woman's hair to change from red in the reference image to blonde. Phantom-1.3B retains the red hair, whereas AttriForget produces the requested blonde hair. The attention visualization shows suppression of reference attention in parts of the hair region, consistent with relaxing reference constraints associated with the evolving hair attribute.

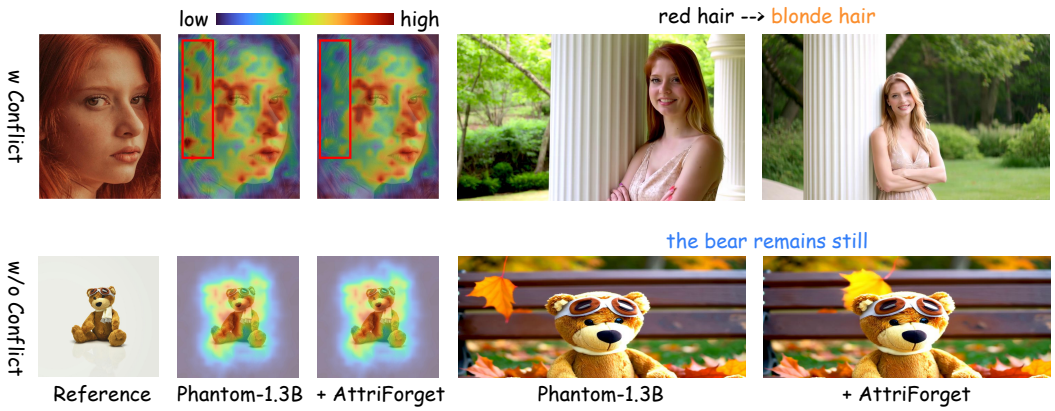


Figure 9: Attention visualization and generation results with and without attribute conflict. The heatmaps show reference attention with AttriForget, with warmer colors indicating higher attention. Top: the prompt requests changing the woman's hair from red to blonde. AttriForget suppresses attention to parts of the hair region and enables the requested color change. Bottom: the prompt asks the bear to remain still, and the attention remains unchanged after adding AttriForget.

The correspondence between the affected region and the requested appearance change provides qualitative support for the semantic relevance of the probed subspace.

Without Attribute Conflict. The prompt asks the bear to remain still, so its subject attributes remain persistent. In this case, $\mathcal{A}^{\mathrm{E}}(p) = \emptyset$ and no forgetting is applied: Eq. 8 leaves the reference features unchanged, reducing AttriForget to the original backbone. The attention visualization remains unchanged after adding AttriForget, consistent with preserving valid reference constraints when no attribute evolution is required.

D Inference Latency

We evaluate the inference latency of VACE-1.3B (Jiang et al., 2025), Kim et al. (5B) (Kim et al., 2026), Phantom-1.3B, and Phantom-14B (Liu et al., 2025) before and after adding AttriForget. Table 6 reports the inference times and the additional latency introduced by AttriForget. Attribute graphs are constructed and cached in advance. The reported latency covers the complete video generation process using these cached graphs and excludes attribute graph construction time.

Table 6: Inference latency across four backbones. Absolute and relative overheads are measured against the corresponding baseline. The final row reports their arithmetic means across backbones.

Backbone	Resolution	Frames	Steps	GPUs	Baseline (s)	+AttriForget (s)	Overhead (s)	Overhead
VACE-1.3B	832 × 480	81	50	1×H20	542.5	552.9	+10.4	+1.92%
Kim et al.(5B)	720 × 480	49	50	1×H20	421.5	440.9	+19.4	+4.60%
Phantom-1.3B	832 × 480	81	50	1×H20	541.8	558.4	+16.6	+3.06%
Phantom-14B	832 × 480	81	50	1×H20	2428.6	2602.1	+173.5	+7.14%
Average	–	–	–	–	–	–	**+55.0**	**+4.18%**

AttriForget modifies reference features during inference before the existing attention key and value projections. It introduces no additional trainable parameters, requires no training or parameter updates, and does not increase the number of sampling steps. The attribute subspace bases are computed once per layer and cached across diffusion steps, allowing the subsequent forgetting operations to reuse these bases when modulating reference features.

The relative latency increase ranges from 1.92% to 7.14% across the tested backbones. Averaged across the four backbones, the additional runtime is 55.0 seconds and the relative overhead is 4.18%,

or approximately 4%. This modest overhead demonstrates that AttriForget provides attribute control with little additional inference cost across different video generation architectures.

E Details of User Study

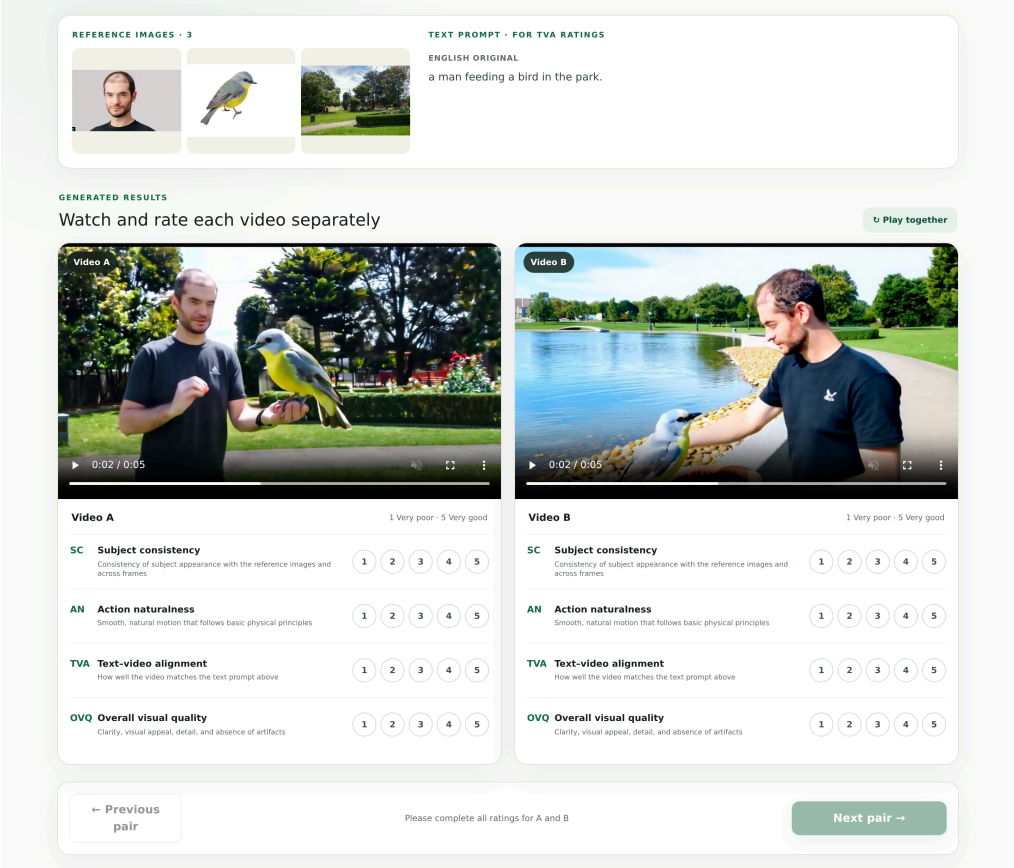


Figure 10: Example interface used in the user study. The interface presents the reference images and text prompt, then displays two generated videos as Video A and Video B. Evaluators rate each video separately using a five-point scale.

Fig. 10 illustrates the evaluation interface. We conduct a pairwise user study on the Open Domain subset of OpenS2V-Eval. We recruit 20 evaluators with varied ages and educational backgrounds. Each evaluator receives 20 randomly sampled cases and rates both generated videos for subject consistency (SC), action naturalness (AN), text-video alignment (TVA), and overall visual quality (OVQ). Every dimension is scored from 1 to 5, where a higher score indicates better perceived quality.

To keep the evaluation objective, we hide the method identities from the evaluators. The two videos are shown as Video A and Video B, and their order is randomized for each case. The interface does not display the backbone name, the method name, or any indication of which video uses AttriForget. Evaluators therefore judge the visual results from the same reference images and text prompt without being able to identify our method from the presentation. Each video is rated separately before the scores are compared.

Table 7: User study results on the Open Domain subset of OpenS2V-Eval. Scores use a five point scale. Higher values indicate better perceived quality. The method identities are hidden during evaluation.

Method	SC↑	AN↑	TVA↑	OVQ↑
VACE-1.3B	3.15	3.06	2.67	3.25
+ AttriForget	**3.49**	**3.23**	**3.01**	**3.39**
Phantom-14B	3.77	3.36	3.43	3.61
+ AttriForget	**4.06**	**3.74**	**3.74**	**3.86**

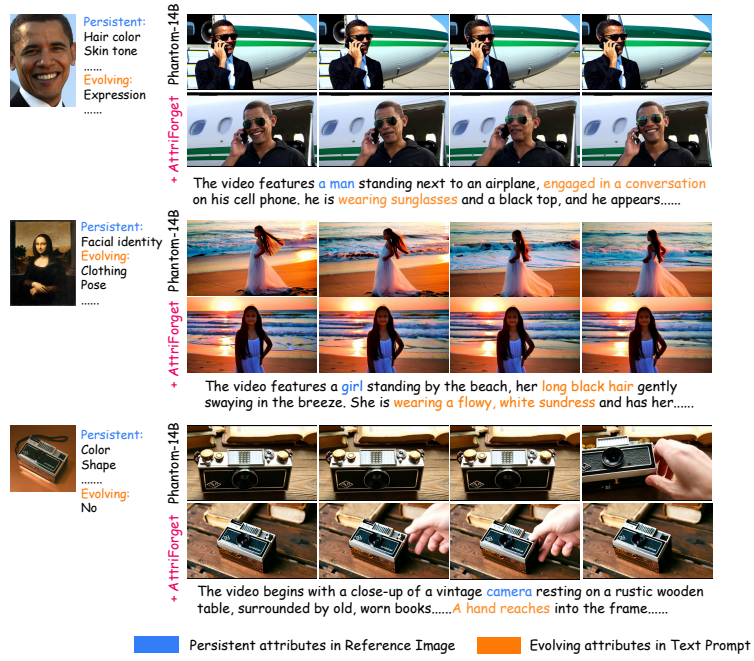
Table 7 shows that AttriForget improves all four user-rated dimensions on both backbones. With VACE-1.3B, the scores increase from 3.15 to 3.49 for subject consistency, from 3.06 to 3.23 for action naturalness, from 2.67 to 3.01 for text video alignment, and from 3.25 to 3.39 for overall visual quality. With Phantom-14B, the corresponding scores increase from 3.77 to 4.06, from 3.36 to 3.74, from 3.43 to 3.74, and from 3.61 to 3.86. The consistent gains across two backbones and four dimensions support the perceptual benefit of selectively relaxing reference constraints on evolving attributes while preserving persistent subject information.

F Additional Visualization

We provide additional qualitative results to further demonstrate the effectiveness of AttriForget in subject consistency and video naturalness. Figs. 11 and 12 show results from VACE-1.3B, Phantom-14B, and Phantom-1.3B across human and object scenarios. AttriForget preserves persistent subject attributes from the reference images while allowing necessary attributes such as pose, expression, clothing, and interaction state to evolve according to the text prompt. The generated videos therefore maintain the identity and appearance of the reference subjects while exhibiting coherent changes and natural temporal behavior. These examples further support our central insight that reference features should provide constraints for persistent attributes and textual conditioning should govern evolving attributes, allowing the two sources of guidance to complement each other in customized video generation.



(a) VACE-1.3B



(b) Phantom-14B

Figure 11: Additional qualitative results across two backbones. AttriForget preserves persistent subject attributes while enabling the evolving attributes specified by the text prompt.

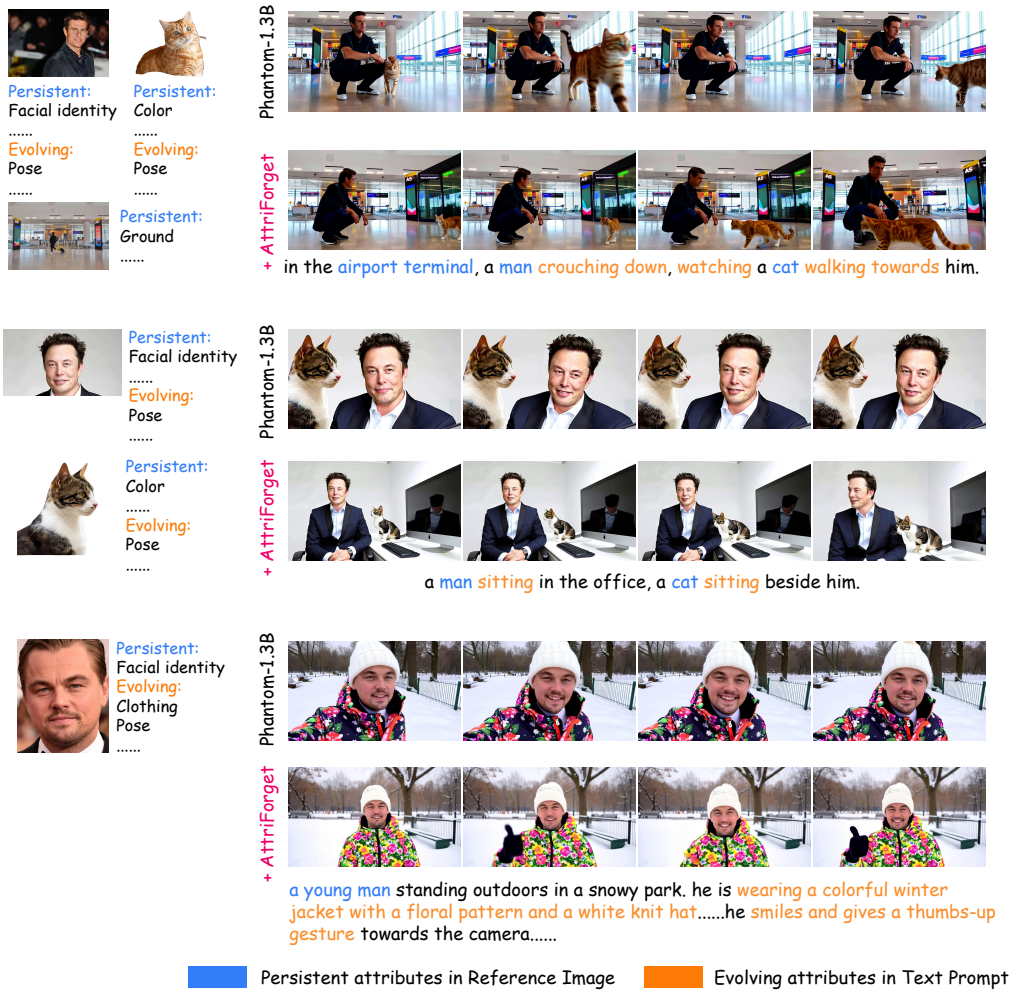


Figure 12: Additional qualitative results with Phantom-1.3B. AttriForget preserves persistent subject attributes while enabling the evolving attributes specified by the text prompt.

G LIMITATION AND FUTURE WORK

AttriForget relies on VLM-generated attribute graphs, whose coverage and state assignments may be imperfect under ambiguous prompts or subtle visual distinctions. Reference conflicts may not be fully resolved when evolving attributes are strongly entangled with persistent ones. It also operates on reference representations provided by the backbone: reducing coupling between persistent and evolving attributes does not explicitly strengthen reference information extraction. Consequently, improvements in subject consistency and naturalness may not recover reference details that the backbone inadequately represents or preserves. In Fig. 13, AttriForget improves naturalness and better preserves the book's appearance, but the reference bridge scene remains absent. Future work could refine uncertain graph assignments using confidence estimates and combine AttriForget with mechanisms that enhance reference information extraction to better preserve relevant reference details.

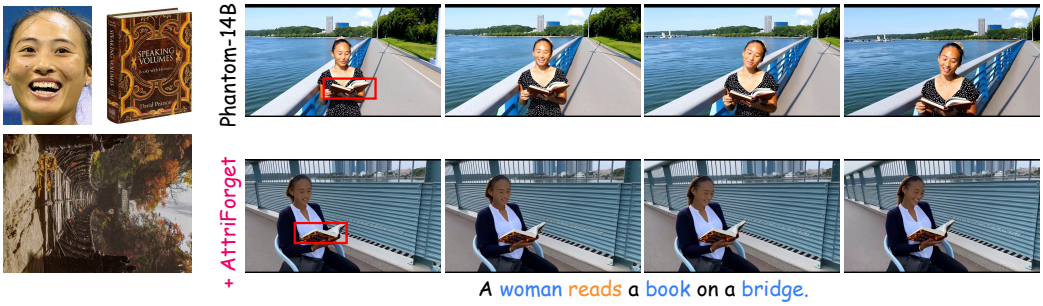


Figure 13: A failure case of background reference preservation.