# MVI2V: Human Centric Image to Video Generation with Multiview Consistent Appearance

Pengfei Liu[1*] Mingyi Xu[1,2*†] Wentao Jiang[1] Tiezheng Ge[1]
Ming Zeng[2‡]

[1] Taobao & Tmall Group of Alibaba, Beijing, China
{zhukong.lpf,xumingyi.xmy,winter.jwt,tiezheng.gtz}@taobao.com
[2] Xiamen University, Xiamen, China
zengming@xmu.edu.cn

**Abstract.** Current image-to-video models struggle to maintain appearance consistency such as the back of a garment. Lacking complete observations, they hallucinate missing views, leading to visual inconsistencies, limiting deployment in applications like e-commerce, where strict visual fidelity is required. In this paper, we propose the Multiview Enhanced Image-to-Video Generation Model (MVI2V), which introduces additional multi-view images of a person or garment as reference inputs. Concretely, MVI2V adds a structurally identical forward stream for reference images, upgrading single- or dual-stream baselines into dual- or triple-stream architectures, and performs cross-stream fusion via self-attention to enable bidirectional information flow among token types. To better exploit the references, we adopt an inpainting sub-task that randomly masks the person region in the conditioning image, forcing the model to rely more on the reference views. We further design a data curation pipeline that selects videos with large viewpoint changes and extracts diverse multi-view reference frames. Extensive experiments on single-stream Wan2.1 and our in-house dual-stream model demonstrate that MVI2V effectively leverages multi-view references while preserving the base models' single-image to video capability, validating our method's generalizability to different architectures.

**Keywords:** Multiview Image-to-Video Generation · Multiview Reference Integration · Appearance Consistency · E-commerce Applications

## 1 Introduction

Recently, significant progress in denoising diffusion models (DDM) has greatly enhanced the generation ability of videos from text descriptions [13, 22, 29].

[*] Equal contribution.
[†] Work done during the internship at Alibaba Group.
[‡] Corresponding author.

Among diffusion model applications, Image-to-Video (I2V) generation is a promising direction that animates a static image based on text prompts, balancing controllability with creative flexibility. As such, it has found widespread use in entertainment, social media, and personalized content creation.

In this paper, we focus on the human-centric application of I2V, which takes a human image as input and animates the person according to user-provided text prompts. Unfortunately, in demanding scenarios such as e-commerce, where visual fidelity and appearance consistency across different views and poses with actual subjects or garments are critical, it is insufficient for the model to generate videos with only one conditioning image. The model is forced to hallucinate the appearance of unseen regions of the subject and garment from the first image. A straightforward solution to this is to incorporate additional input views into the model. Therefore, in this paper, to address the aforementioned issue of clothing inconsistency, we propose the multi-view image-to-video (MVI2V) generation task and present our methodology. As shown in Fig. 1, conventional image-to-video approaches (e.g., Wan2.1) generate wrong clothing patterns for unseen viewpoints. In contrast, our method effectively leverages additional reference images to maintain high fidelity with the ground-truth cloth.

Our MVI2V task can be distinguished with other three related human-centric video generation paradigms as illustrated in Fig. 2. **Image-to-video or human animating** [10, 31] takes a single reference image plus a text prompt (and optionally pose sequence) and outputs a video where the character moves; unseen regions such as the back of a garment are thus hallucinated. **Video virtual try-on** [5, 15] conditions on a source video or pose sequence and a garment image to produce a try-on video, emphasizing garment transfer rather than animating one identity. In contrast, **our multiview image-to-video** takes one reference image, a prompt, and multiple views of the same person or garment, and generates an animated video that remains consistent across views. Our setting is more demanding: the model must fuse multi-view references of the *same* subject or garment and preserve that consistency throughout the generated video, which neither single-view I2V nor video try-on is designed to address.

However, it is non-trivial to achieve the desired multi-view reference functionality from scratch training. The primary challenge is the scarcity of human-centric videos featuring dynamic body reorientations, as such data occupies the long tail of the overall distribution. Moreover, due to the requirement that the reference clothing image's perspective (front or back) must differ from that of the video's first frame to supply complementary information, the pool of suitable training pairs becomes even more limited. The scarcity of such data makes training a generalizable multi-view I2V model from scratch challenging. Therefore, we opt to extend existing powerful I2V models by finetuning them on a targeted multi-view dataset.

Specifically, we extend the original pretrained model with an additional forward stream, structurally identical to the primary one for video tokens, to process multi-view reference images and enable bidirectional information flow between reference image tokens and video tokens via self-attention in each transformer

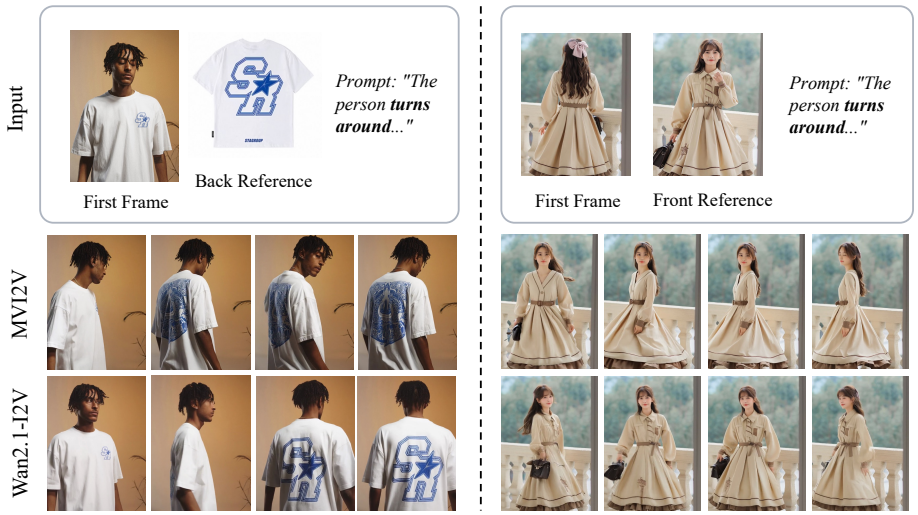


**Fig. 1:** Our MVI2V multi-view generation results compared with Wan2.1-I2V baseline.

block. In this way, video tokens can gather information about the unseen garment appearance from the multi-view image tokens. The newly introduced forward streams employs low-rank adaptation (LoRA) to further reduce the number of training parameters, thereby conserving training resources. We name this architecture Multiview Enhanced Image-to-Video Generation Model (MVI2V). MVI2V is general, effective, and robust, readily adaptable to existing popular video diffusion backbones, including single-stream based models [13, 29] and dual-stream based models [22, 26]. As detailed in Sec. 5, extensive experiments show that our MVI2V can generate high-quality videos that maintain appearance consistency across different views with the actual input garment.

We also constructed a specialized data pipeline to filter out multi-view targeted dataset to accelerate model training. Starting from a large video pool, this pipeline sequentially filters for single-person videos and then for those featuring large angles of subject rotation. Furthermore, we formulated a multi-view reference frame extraction strategy. It identifies five comprehensive views, including frontal and back perspectives. Concurrently, the top and bottom garments are segmented from the frontal and back views to augment the dataset of garment reference images. We found that this data curation pipeline substantially improves both the model's learning efficiency and its final reference capability.

Despite employing the MVI2V model and the meticulously curated dataset, we observe that the model can still "learn a shortcut" during training: it can generate a temporally consistent video while completely ignoring the reference frames, although the resulting appearance will not match them. This is because that for a pre-trained I2V model, the conditioning first frame already provides a strong prior for the subject's appearance. This tendency impedes the model's ability to learn from the new reference images. To counteract this, we introduce

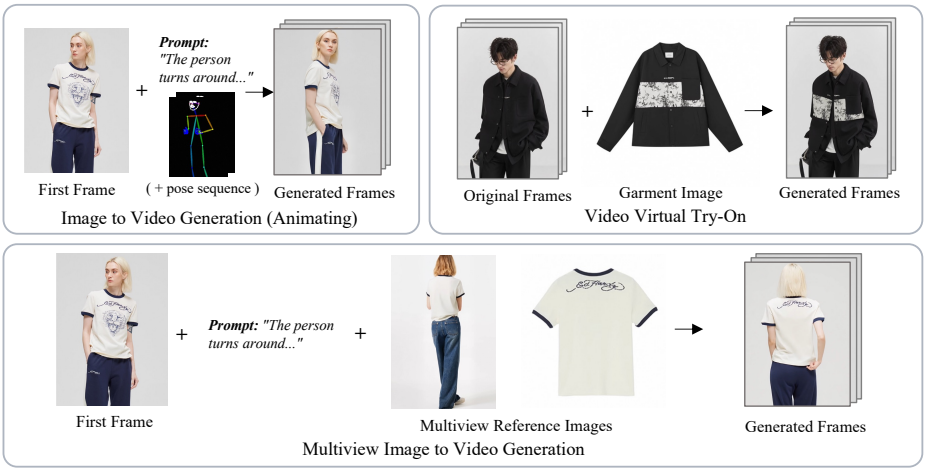


**Fig. 2:** Comparison of three human-centric video generation tasks.

an auxiliary inpainting subtask. By randomly masking out the person's region in the conditioning first frame, this strategy compels the model to reconstruct the subject's appearance by leveraging the information from solely reference images. Our experiments demonstrate that this simple approach further enhances the model's capability to adhere to the reference images.

In summary, our contributions are outlined below:

- We introduce a novel multi-view image-to-video task and propose the first training framework that integrates our MVI2V architecture to augment existing image-to-video generation models with multi-view reference support.
- MVI2V architecture is readily adaptable to both single-stream and dual-stream-based backbones, with only a slight overhead in trainable parameters. The proposed inpainting sub-task also compels the model to escape local optima, thereby further enhancing its multi-view reference capability.
- Our carefully designed dataset collection pipeline is effective in improving the training efficiency for this specialized task.
- Experiments demonstrate the effectiveness of our proposed MVI2V model, inpainting subtask and dataset curation pipeline.

## 2 Related Work

### 2.1 Video Generation Models

The success of Diffusion Models in image generation [8, 23] has spurred their exploration in the video domain [6, 13, 16, 26, 29]. Video Diffusion Models (VDMs) was the first work among them to achieve this by extending the 2D U-Net architectures of image generation models into 3D U-Net for video synthesis. Other works [7, 27, 35] introduce 1D temporal attention to reduce computation

overhead. The advent of the Diffusion Transformer (DiT, [21]) introduced the Transformer architecture [25] into diffusion models, leading to substantial improvements in generation quality. Consequently, modern video generation models now widely employ the DiT architecture for modeling both visual appearance and temporal dynamics.

Modern DiT-based video generation models typically process multi-modal inputs in two ways. The **single-stream** approach (e.g., MovieGen [22] and Wan2.1 [26]) models video tokens through a single set of parameters and interacts with text tokens through cross-attention mechanism. The second is a **dual-stream** approach based on the Multimodal Diffusion Transformer (MMDiT), which processes video and text with separate parameters, and merges them during self-attention mechanism (e.g., CogvideoX [29], HunyuanVideo [13] and SeeDance [6]). In practice, both methods achieve excellent results in video-text alignment and generation realism.

### 2.2 Conditional Video Generation

Conditional video generation aims to guide the synthesis process using various inputs, with Image-to-Video (I2V) being a prominent sub-task. Typical I2V methods adapt pre-trained T2V models, conditioning on a single initial frame by concatenating its latent features [6, 13, 26, 29] or using its CLIP embedding [26]. However, these single-frame approaches inevitably fail to maintain multi-view consistency, as they must hallucinate unseen parts. This specific challenge of fusing multiple reference images for consistency distinguishes our work. For example, Subject-to-Video (S2V) generation [19] focuses on identity preservation across scenes rather than the accurate rendering of unseen views. Similarly, other conditional methods [32] are typically designed to govern distinct conditions (e.g., pose or depth) and are not designed to fuse multiple viewpoints. Our MVI2V is built to overcome this limitation, introducing a dedicated stream to process multiple reference images and thereby ensure the final video is consistent with all provided views.

## 3 Preliminary

Flow matching models [18, 20] provide a theoretically grounded framework for learning continuous-time generative processes in diffusion models, and have demonstrated strong capabilities in visual generation. It circumvents iterative velocity prediction, enabling stable training via ordinary differential equations (ODEs) while maintaining equivalence to maximum likelihood objectives. In general practice, the probability path usually exists in a latent space, not in a pixel space, to reduce computation and speed up training. During training, the video tensor is first compressed by a video encoder $\mathcal{E}$ into latent space as $x_1$. Given a randomly sampled noise $x_0 \sim \mathcal{N}(0, \mathcal{I})$ and a timestep $t \in [0, 1]$, an intermediate latent $x_t = tx_1 + (1-t)x_0$ is obtained as the training input. The ground truth $v_t$ is $v_t = \frac{dx_t}{dt} = x_1 - x_0$.

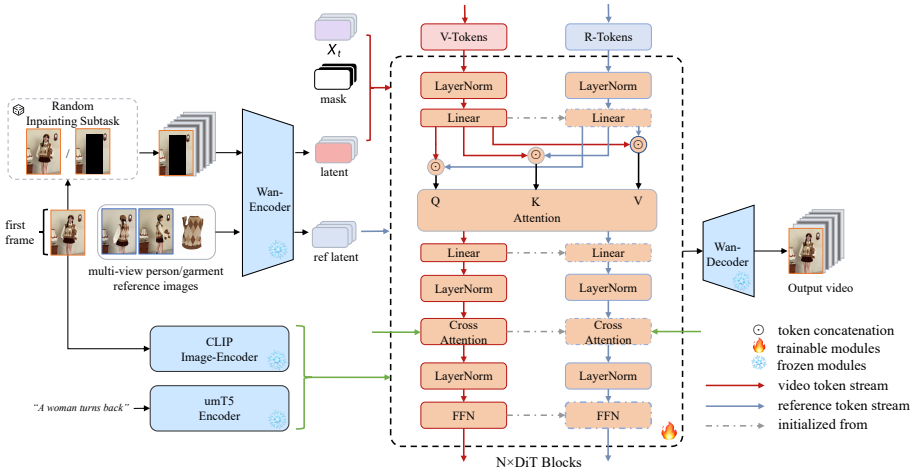


**Fig. 3:** Multiview enhanced image-to-video generation training framework based on Wan2.1 [26] I2V model.

For image-to-video generation, the model is trained to predict the velocity, given the input prompt $c_{txt}$ and the first frame $c_{img}$. Therefore, the loss function is formulated as the mean squared error between the model output and $v_t$,

$$L = E_{x_0,x_1,c_{txt},c_{img},t}||u(x_t, c_{txt}, c_{img}, t; \theta) - v_t||^2 \tag{1}$$

In this paper, we extend the task to multiview enhanced image-to-video generation. Therefore, the model is conditioned on extra reference images $c_{ref}$ and the training loss is reformulated as:

$$L = E_{x_0,x_1,c_{txt},c_{img},c_{ref},t}||u(x_t, c_{txt}, c_{img}, c_{ref}, t; \theta) - v_t||^2 \tag{2}$$

## 4 Methodology

### 4.1 Method Overview

MVI2V enables video generation from multi-view references by enhancing the powerful, pre-existing I2V models with an extra forward stream specifically for processing reference images, which is compatible with dominant video diffusion model backbones such as single-stream and dual-stream structures. Furthermore, we introduce an auxiliary inpainting task during training to facilitate model convergence. To support this specialized task, we also developed a data construction pipeline to curate a high-quality, training-efficient dataset of multi-view video-image pairs. In the following part of this section, we first introduce the training details including network architecture in Sec. 4.2 and inpainting subtask in Sec. 4.3. Then, we describe the dataset construction pipeline in Sec. 4.4.

### 4.2 Multiview Enhanced Image-to-Video Generation

To ensure multi-view consistent garment appearance, we introduce a new input type: clean multi-view reference images. Unlike semantic text prompts, these images provide direct visual information, making them fundamentally different from other inputs. For noisy video tokens, they are linear interpolations of noise and clean video latent, representing intermediate points along the flow-matching trajectory, and thus contain a mix of signal and noise. In contrast, clean reference images are more similar to the endpoint of this trajectory, embodying purely clear visual information. Therefore, we treat these reference images as a unique modality in this paper.

We seek a general way to incorporate the new modality and facilitate information exchange with the other two modalities. Following MMDiT, we introduce a dedicated forward stream to process the reference images and employ self-attention to fuse its features with those from other modalities. In the following, we illustrate the architectural modifications based on Wan2.1 I2V backbone as a concrete example, which is shown in Fig. 3. We leave the modification details on our in-house I2V backbone (dual stream) in supplementary material.

**Reference Feature Process.** We first resize all reference images to a uniform size, and then encode each image into a latent representation using a VAE encoder. Subsequently, these individual latents are stacked to form the reference feature $c_{ref}$. This feature is then patchified to produce the final reference tokens $y$, which are fed into the dedicated reference stream of our network.

**MVI2V Network Architecture.** A typical transformer block of Wan2.1 consists of 3 primary modulated components: self-attention, cross-attention, and feedforward network (FFN). The forward process can be formulated as:

$$\begin{aligned} x_1 &= \textbf{SelfAttention}(x; M_S, W_S) \\ x_2 &= \textbf{CrossAttention}(x_1, c_{txt} \oplus c_{img}; M_C, W_C) \\ x_3 &= \textbf{FFN}(x_2; M_F, W_F) \end{aligned} \tag{3}$$

where $x$ is the input video tokens, $x_1, x_2, x_3$ are the output video tokens of each layer, $M_S, M_C, M_F$ are the collection of modulation parameters, including scale, shift, gate for self-attention, cross-attention, and feedforward layers, respectively. $W_S$ and $W_C$ are the collections of linear projection parameters to project tokens into query, key, value, and output in self-attention and cross-attention operations. And $W_F$ is the parameters of FFN. $c_{txt}, c_{img}$ are the conditional text embeddings and CLIP embeddings of the first image, and $\oplus$ denotes the concatenation operation. In this formulation, we omit the internal residual connections for simplicity.

As shown in the DiT block part of Fig. 3, MVI2V extend each operation with an extra forward stream and parameters for the additional reference image tokens. These two modalities only interact at self-attention operation. This

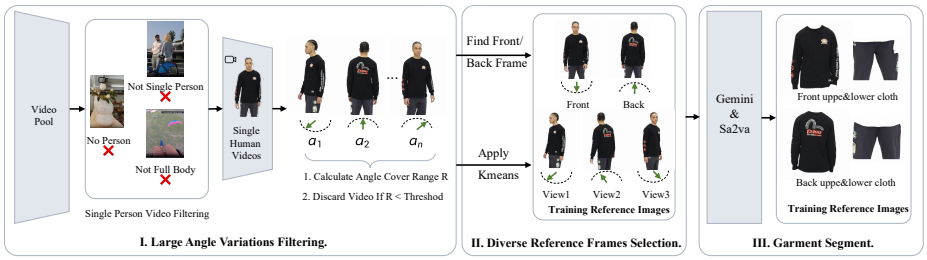


**Fig. 4:** Our targeted data collection pipeline.

modified architecture is formulated as follows:

$$
\begin{aligned}
x_1, y_1 &= \textbf{SelfAttention}(x \oplus y; M_S, M'_S, W_S, W'_S) \\
x_2 &= \textbf{CrossAttention}(x_1, c_{txt} \oplus c_{img}; M_C, W_C) \\
y_2 &= \textbf{CrossAttention}(y_1, c_{txt} \oplus c_{img}; M'_C, W'_C) \\
x_3 &= \textbf{FFN}(x_2; M_F, W_F) \\
y_3 &= \textbf{FFN}(y_2; M'_F, W'_F)
\end{aligned} \tag{4}
$$

Here, $y$ represents the input reference tokens derived from one or more reference images. The variables $y_1, y_2, y_3$ denote the output tokens for the reference image stream from each respective layer. The additional parameters, denoted by $M'_S, M'_C, M'_F, W'_S, W'_C, W'_F$, are specifically introduced to process the multi-view reference images.

To accelerate convergence, a straightforward approach is to initialize the additional parameters that process reference tokens by duplicating the weights of the pretrained one that process video tokens. However, this method doubles the number of trainable parameters, significantly increasing computational overhead and memory footprint. Therefore, we propose a more parameter-efficient strategy employing Low-Rank Adaptation (LoRA) [9]. Instead of full parameter replication, we augment each original weight matrix with low-rank matrices as follows:

$$
\begin{aligned}
W'_S &= W_S + \alpha A_S B_S \\
W'_C &= W_C + \alpha A_C B_C \\
W'_F &= W_F + \alpha A_F B_F
\end{aligned} \tag{5}
$$

where $\alpha$ is the scaling factor for the LoRA weights. The matrices $A_{(.)}$ and $B_{(.)}$ are the down-projection and up-projection matrices, respectively, for the LoRA layers applied to different components of the model. This parameter-efficient approach significantly reduces the number of trainable parameters. In our configuration, the added LoRA layers on the Wan2.1 model (I2V-14B-720P) account for only 0.3B (approximately 2.1%) extra parameters.

### 4.3 Inpainting Subtask

A key challenge in extending pretrained I2V models is that the conditioning first frame already provides a strong appearance prior. This enables the model during training to easily generate appearance-consistent video frames using only information from the first frame, particularly for viewpoints closely resembling the first frame (a common case for the early frames of generated video). This problem leads the model to easily get trapped in a local optimum and may cause the model to ignore the new multi-view reference images. This also presents a data construction challenge, as reference images must provide novel viewpoints distinct from the first frame to encourage the model to learn complementary features, limiting the range of applicable training data.

To address this and improve data utilization, we propose a simple yet effective strategy: randomly masking the human region in the conditioning frame during training. Specifically, we utilize human detection to derive the subject's bounding box and subsequently mask the entire rectangular region as shown in Fig. 3. This forces the model to learn the subject's appearance exclusively from the reference images, thereby decoupling it from the conditioning frame's identity. Experiments validate that this approach significantly enhances the model's ability to synthesize subjects based on reference images. Notice that this masking strategy is exclusively applied during the training phase; during inference, the full conditioning frame is used as input.

### 4.4 Dataset Construction

Fig. 4 shows the overall data collection pipeline. We begin by filtering a large internal video pool using a human detection model [28]. We discard videos that contain no people, multiple people, or partial views of one person, retaining only single-person videos with a sufficiently large subject.

**Large Angle Variations Filtering.** For each remaining video clip, we estimate the per-frame body orientation angles using a human reconstruction model [12]. Clips with an angular cover range greater than a predefined threshold are selected for training, while others are filtered out.

**Diverse Reference Frames Selection.** Our reference frame selection is a two-step procedure. First, recognizing that frontal and back views provide the most comprehensive information on clothing appearance, we identify the two frames closest to these canonical views as training reference images. Second, to ensure viewpoint diversity, we apply k-means clustering to all orientation angles in the clip and select frames closest to the resulting cluster centroids as additional training reference images.

**Garment Segmentation.** Finally, to address the data scarcity of the garment images, we employ the Gemini [24] and Sa2va [30] model to segment the upper and lower garments from the frontal and back reference frames, respectively. These segmented regions are then incorporated as additional reference frames for training. In the final dataset, each video is associated with nine reference images: two frontal/back views, three clustered multi-view images, and four segmented garment images.

# 5 Experiments

## 5.1 Baselines and Metrics

Our task formulation differs from character animation [10,31] and video virtual try-on [5,15]: we take multi-view reference images of the same subject or garment as additional input, whereas the former assumes a single reference image (plus optional poses) and the latter assumes one source video and one garment image. Because the input modalities and objectives are not compatible, those methods cannot be applied to our setting as baselines for a fair comparison. We therefore compare against the following I2V backbones extended with multi-view inputs under the same data and evaluation protocol.

**Baselines.** A naive approach to incorporate multi-view reference images is to directly concatenate reference tokens with the video tokens. The combined sequence is then processed through Transformer blocks, after which only the video tokens are isolated for decoding. We term this method Token-Concat. Recent studies have demonstrated the effectiveness of this method in some tasks of Conditional Video Generation [11,32]. Concat-ID [34] further enhances Token-Concat by adding an individual time projection layer for new reference tokens. We implement Token-Concat, Concat-ID on Wan2.1 (I2V-14B-720P) as the single-stream base model and compare them with our MVI2V architecture (MVI2V-Wan2.1). Furthermore, we comprehensively validate the effectiveness of our proposed architectural modifications across both single-stream and dual-stream base models. Specifically, we employ a 15B in-house image-to-video generation model as our baseline. This model features a hybrid architecture, combining dual-stream and single-stream components, similar to Flux [14]. We then integrate our MVI2V method by extending the dual-stream layers into triple-stream layers, resulting in MVI2V-In-House, and evaluate its performance against the original baseline.

**Benchmark.** To evaluate multi-view reference capability, we construct a new benchmark tailored for this task. Specifically, we manually collected multi-view images of the same subject from e-commerce websites. For each set of images, one view serves as the conditioning first frame, while the others serve as reference images. Our benchmark comprises a total of 300 test samples, encompassing both person and garment as reference images. The evaluation metrics are categorized into the following two types:

**Garment Consistency.** We assess garment consistency between the generated video frames and the reference images using GPT [1] and DINO [2]. For GPT-based metrics, inspired by VQAScore [17], we provide detailed instructions and rubrics to guide the GPT to grade the generated videos across garment consistency dimension. For the DINO-based metric, we compute the cosine similarity of DINO features between each generated frame and reference images as the garment consistency score. We apply the two scoring strategies to the multi-view person and garment evaluation sets, respectively, yielding four distinct metrics $GPT_{\text{person}}, GPT_{\text{garment}}, DINO_{\text{person}}, DINO_{\text{garment}}$.

**Table 1:** Quantitative comparison with other baselines on our multi-view benchmark.

| Model | Garment Consistency | | | | Video Quality | | |
|---|---|---|---|---|---|---|---|
| | $GPT_{\text{person}}$ | $GPT_{\text{garment}}$ | $DINO_{\text{person}}$ | $DINO_{\text{garment}}$ | Subject Consistency | Background Consistency | Motion Smoothness |
| Wan2.1 | 0.752 | 0.409 | 0.667 | 0.582 | 0.916 | 0.928 | 0.986 |
| Token-Concat | 0.797 | 0.512 | 0.715 | 0.620 | 0.902 | 0.922 | 0.982 |
| Concat-ID | 0.841 | 0.542 | 0.719 | 0.613 | 0.918 | 0.936 | 0.988 |
| MVI2V-Wan2.1 (Ours) | **0.862** | **0.585** | **0.728** | **0.656** | **0.927** | **0.938** | **0.988** |
| In-House | 0.692 | 0.416 | 0.700 | 0.601 | 0.926 | 0.934 | **0.993** |
| MVI2V-In-House (Ours) | **0.868** | **0.607** | **0.776** | **0.633** | **0.930** | **0.934** | **0.993** |

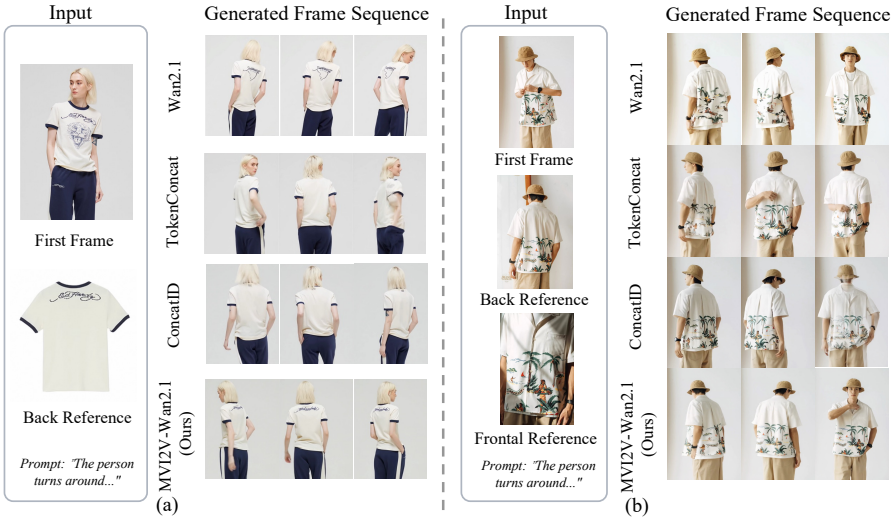


**Fig. 5:** Qualitative comparisons with other multi-view image-to-video baselines using garment (a) or person (b) images as reference.

**Video Quality.** Consistent with previous work [3, 4], we evaluate the "Subject Consistency", "Background Consistency", and "Motion Smoothness" metrics in VBench [33] to evaluate how the two major challenges including temporal consistency and smoothness are addressed.

### 5.2 Qualitative and Quantitative Evaluation

Fig. 5 presents a qualitative comparison between our MVI2V and other baselines. The original Wan2.1 model, lacking information from other perspectives, tends to hallucinate when encountering unseen views, generating detailed textures that are inconsistent with the actual garment. While Token-Concat and Concat-ID demonstrate a basic reference ability, they fail in hard cases with scarce training data, like garment reference images. Furthermore, their model architectures intermingle different types of tokens, which reduces the model's disentanglement capabilities. This, in turn, makes them prone to generating anatomical errors

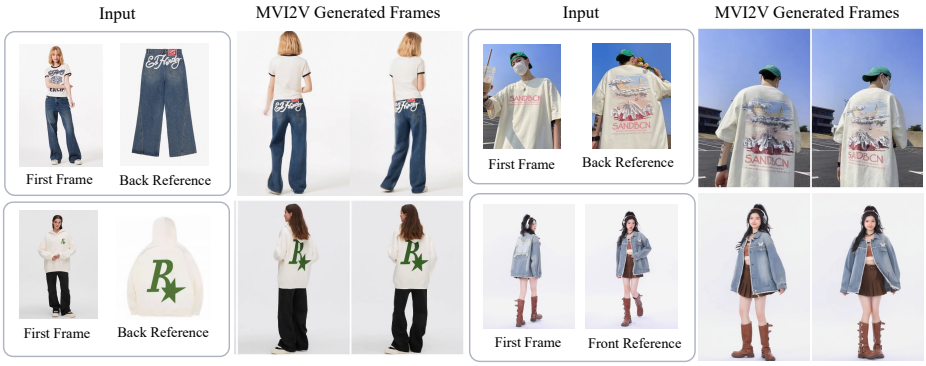


**Fig. 6:** More examples of our MVI2V multi-view generation results.

**Table 2:** Ablation results. The '+' symbol denotes the addition of a core component to the configuration of the preceding row.

| Model | Ablation Item | | | Garment Consistency | | | | Video Quality | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MVI2V Architecture | Reference Selection | Inpainting Subtask | $GPT_{\text{person}}$ | $GPT_{\text{garment}}$ | $DINO_{\text{person}}$ | $DINO_{\text{garment}}$ | Subject Consistency | Background Consistency | Motion Smoothness |
| Wan2.1 | ✗ | ✗ | ✗ | 0.752 | 0.409 | 0.667 | 0.582 | 0.916 | 0.928 | 0.986 |
| + MVI2V Architecture | ✓ | ✗ | ✗ | 0.784 | 0.470 | **0.744** | 0.631 | 0.905 | 0.923 | 0.982 |
| + Reference Selection | ✓ | ✓ | ✗ | 0.853 | 0.562 | 0.743 | 0.638 | 0.916 | 0.932 | 0.986 |
| + Inpainting Subtask (Ours) | ✓ | ✓ | ✓ | **0.862** | **0.585** | 0.728 | **0.656** | **0.927** | **0.938** | **0.988** |

and abrupt transitions. In contrast, our method successfully leverages reference texture details to reconstruct an appearance consistent with the actual garment. We also provide more qualitative results generated by our MVI2V in Fig. 6, where representative multi-view examples are shown to highlight the appearance consistency of our model across different views.

The quantitative results, presented in Tab. 1, are consistent with our qualitative findings. Token-Concat incorporates multi-view information, but due to the coupling of different modalities, it yields only a marginal improvement in the garment consistency metrics compared to the baseline, while causing a noticeable degradation in video quality metrics. Concat-ID mitigates this issue by introducing a dedicated time projection layer for reference tokens to differentiate between the two modalities. This approach alleviates the decline in video quality and achieves further gains in Garment Consistency. Nevertheless, the effectiveness of this disentanglement strategy remains limited. In contrast, our method employs an additional dedicated stream to completely disentangle tokens from different modalities. This design prevents cross-modality interference, thus achieving the highest reference accuracy and simultaneously enhancing the overall quality of the generated videos. Furthermore, MVI2V-In-House also demonstrates significant metric improvements over the baseline. This validates the generalizability of our method across both single-stream and dual-stream architectures. Qualitative comparisons with the in-house baseline are provided in the supplementary material.

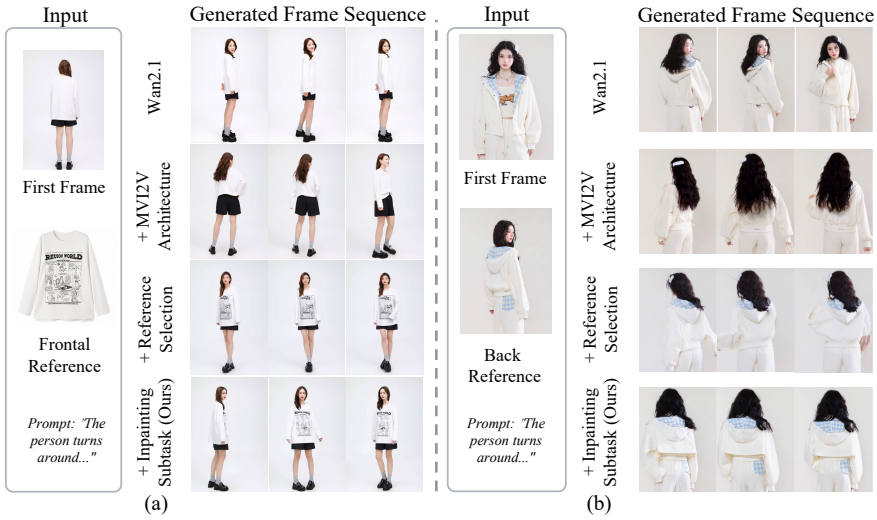


**Fig. 7:** Qualitative ablation results using garment (a) or person (b) images as reference. Our method behaves better on texture inconsistencies (a) or loss (b) problems.

### 5.3 Ablation Study

We evaluate the incremental addition of the three core components of MVI2V: (1) the MVI2V architecture, (2) our reference frame selection strategy, and (3) the inpainting subtask, building upon the Wan2.1 baseline. The quantitative results are shown in Tab. 2.

**Table 3:** Quantitative analysis of image-to-video capability preservation on the original VBench I2V benchmark.

| Model | Subject Consistency | Background Consistency | Motion Smoothness |
|---|---|---|---|
| Wan2.1 | 0.938 | 0.969 | 0.980 |
| MVI2V-Wan2.1 | 0.953 | 0.969 | 0.987 |
| In-House | 0.965 | 0.964 | 0.994 |
| MVI2V-In-House | 0.959 | 0.958 | 0.993 |

Specifically, we start by training the MVI2V architecture with simple random reference frame selection strategy and no inpainting subtask (+MVI2V Architecture). This naive extension of the baseline yields only a marginal improvement in the garment consistency metrics. This is primarily because a simple random reference frame sampling strategy is prone to selecting redundant views or those closely resembling the initial first frame, leading to inefficient training. In contrast, when our proposed reference selection strategy is employed (+Reference Selection), the garment consistency metrics show a significant boost. Finally, introducing the proposed inpainting subtask (+Inpainting Subtask) compels the model to rely more on the information from reference images, further enhancing its reference capability. This demonstrates that achieving multi-view consistency in I2V generation is not merely a model-driven


**Fig. 9:** Image-to-Video generation results of our MVI2V-Wan2.1 model.

task, but rather necessitates targeted adaptations across the model architecture, data construction, and training strategy.

Fig. 7 illustrates the qualitative results of our ablation study. It is observed that, compared to the baseline, employing MVI2V architecture alone is insufficient for the model to adhere to reference images. With the addition of our proposed reference frame selection strategy, the model gains some reference capability, but still suffers from texture inconsistencies or loss. Finally, by incorporating the inpainting sub-task, the model's referencing ability is maximized, successfully generating videos where the garment appearance is consistent with the reference image.

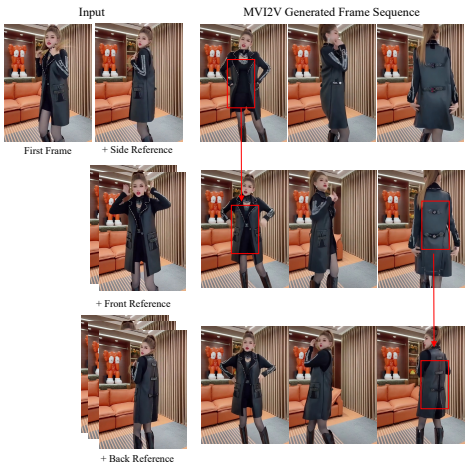


**Fig. 8:** Qualitative results with an increasing number of reference views.

**Impact of the Number of Reference Views.** As shown in Fig. 8, the consistency of the generated video improves as additional reference viewpoints are provided, since the model can better cover self-occluded regions and disambiguate appearance details. In most real-world scenarios, however, we observe that two complementary images (front and back views) already provide a good generation quality.

**Preservation of Foundational I2V Capabilities** Our MVI2V models can also function as an I2V model, retaining the capability for single-image to video generation. To validate this, we evaluated their I2V performance on the original Vbench benchmark. Tab. 3 compares the Vbench metrics of the MVI2V models with its baseline version. The extended models show comparable

performance across all metrics. This indicates that our MVI2V training process does not compromise the original I2V capabilities. Fig. 9 shows one visualization of single-image to video generation result of our MVI2V-Wan2.1 model.

## 6 Conclusion

In this paper, we introduce the task of multi-view guided image-to-video (I2V) generation and propose a comprehensive framework to train a competent model. Starting from a pretrained I2V model, we extend its architecture and incorporate an inpainting-based training strategy to facilitate efficient learning. We also devise an effective data curation pipeline tailored for this task. Given reference images of a person or a garment from various viewpoints, our model can generate dynamic person videos of the subject with a consistent multi-view appearance.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62072382), Alibaba Group through Alibaba Innovative Research Program, and the Yango Charitable Foundation.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021)
3. Chen, J.K., Bansal, A., Vo, M.P., Wang, Y.X.: Dress&dance: Dress up and dance as you like it-technical preview. arXiv preprint arXiv:2508.21070 (2025)
4. Chen, J.K., Bansal, A., Vo, M.P., Wang, Y.X.: Virtual fitting room: Generating arbitrarily long videos of virtual try-on from a single image–technical preview. arXiv preprint arXiv:2509.04450 (2025)
5. Chong, Z., Zhang, W., Zhang, S., Zheng, J., Dong, X., Li, H., Wu, Y., Jiang, D., Liang, X.: Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation (2025), `https://arxiv.org/abs/2501.11325`
6. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
7. Gong, L., Zhu, Y., Li, W., Kang, X., Wang, B., Ge, T., Zheng, B.: Atomovideo: High fidelity image-to-video generation. ArXiv **abs/2403.01800** (2024)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. ArXiv **abs/2006.11239** (2020)
9. Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021)

10. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023), website: https://humanaigc.github.io/animate-anyone/
11. Ju, X., Ye, W., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Xu, Q.: Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907 (2025)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Recognition (CVPR) (2018)
13. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
14. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024), accessed: 2026-03-01
15. Li, G., Zheng, S., Zhang, H., Chen, J., Luan, J., Ou, B., Zhao, L., Li, B., Jiang, P.T.: Magictryon: Harnessing diffusion transformer for garment-preserving video virtual try-on. arXiv preprint arXiv:2505.21325 (2025)
16. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
17. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision (2024)
18. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. ArXiv **abs/2210.02747** (2022)
19. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
20. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. ArXiv **abs/2209.03003** (2022)
21. Peebles, W.S., Xie, S.: Scalable diffusion models with transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 4172–4182 (2022)
22. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10674–10685 (2021)
24. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
25. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017)
26. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
27. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. ArXiv **abs/2308.06571** (2023)

28. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)
29. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
30. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. ArXiv **abs/2501.04001** (2025)
31. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In: International Conference on Machine Learning (2025)
32. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:2503.07027 (2025)
33. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. ArXiv **abs/2503.21755** (2025)
34. Zhong, Y., Yang, Z., Teng, J., Gu, X., Li, C.: Concat-id: Towards universal identity-preserving video synthesis. arXiv preprint arXiv:2503.14151 (2025)
35. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. ArXiv **abs/2211.11018** (2022)