Scaling Zero-Shot Reference-to-Video Generation

Zijian Zhou[1,2], Shikun Liu[1], Haozhe Liu[1], Haonan Qiu[1], Zhaochong An[1], Weiming Ren[1], Zhiheng Liu[1], Xiaoke Huang[1]
Kam-Woh Ng[1], Tian Xie[1], Brandon Han[1], Yuren Cong[1], Hang Li[1], Chuyan Zhu[1], Aditya Patel[1], Tao Xiang[1], Sen He[1]
[1]Meta AI, [2]King's College London

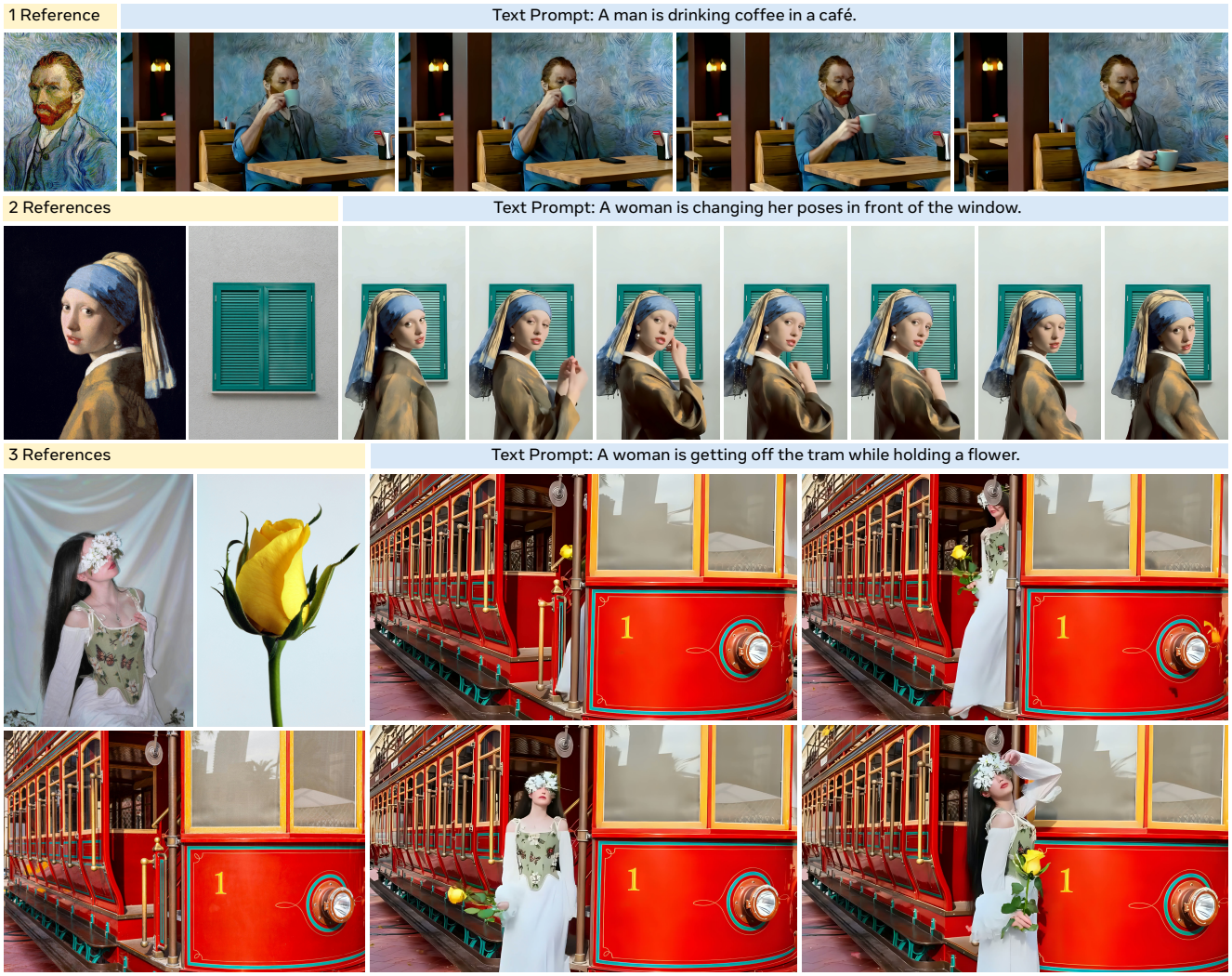

Figure 1. Saber is a zero-shot reference-to-video method trained only on video-text pairs. It preserves identity and appearance while coherently integrating single/multiple references into videos guided by text prompts.

Abstract

Reference-to-video (R2V) generation aims to synthesize videos that align with a text prompt while preserving the subject identity from reference images. However, current R2V methods are hindered by the reliance on explicit reference image-video-text triplets, whose construction is highly expensive and difficult to scale. We bypass this bottleneck by introducing Saber, a scalable zero-shot framework that requires no explicit R2V data. Trained exclusively on video-text pairs, Saber employs a masked training strategy and a tailored attention-based model design to learn identity-consistent and reference-aware representations. Mask augmentation techniques are further integrated to mitigate copy-paste artifacts common in reference-to-video generation. Moreover, Saber demonstrates remarkable general-

ization capabilities across a varying number of references and achieves superior performance on the OpenS2V-Eval benchmark compared to methods trained with R2V data.

1. Introduction

Reference-to-video (R2V) generation synthesizes videos that align with a given text prompt while preserving the identity and appearance of subjects in reference images. This task represents a crucial step toward personalized video generation, enabling applications such as customized storytelling [33, 44, 53] and virtual avatars [11, 13, 36, 49]. Despite recent progress in text-to-video (T2V) and image-to-video (I2V) generation [12, 15, 17, 22, 26, 42, 47, 50], R2V remains uniquely challenging as it must simultaneously ensure semantic alignment with text and maintain high-fidelity subject identity from the reference images.

Existing R2V methods [9, 10, 19, 20, 23, 27, 46, 52] typically rely on constructing explicit R2V datasets (*e.g.*, OpenS2V-5M [48] and Phantom-Data [7]) that contain triplets of reference images, videos, and text prompts. Building such datasets involves complex pipelines for data collection, annotation, clustering and filtering, which are costly and difficult to scale. Moreover, the limited diversity of reference images in these datasets restrict generalization, making it difficult to handle unseen subject categories.

We propose Saber, a scalable, zero-shot framework that bypasses this data bottleneck. Saber is trained solely on large-scale video-text pairs, the same data paradigm used for T2V and I2V models. This design allows Saber to leverage abundant video-text datasets [5, 39, 43], completely eliminating the need for bespoke R2V data construction.

To this end, our method introduces a masked training strategy that uses randomly sampled and partially masked video frames as reference images during training, where the randomness of masking provides diverse reference conditions and improves generalization across subject categories. This process compels the model to learn identity- and appearance-consistent representations from the reference context, effectively simulating the R2V task without R2V data. This strategy is complemented by a tailored attention mechanism, guided by attention masks, which directs the model to focus on reference-aware features while suppressing background noise. To further enhance visual fidelity and mitigate the copy-paste artifacts which is common in reference-to-video generation [10, 18, 27], we integrate a series of spatial mask augmentations, effectively improving the visual quality of the generated videos.

Saber's design is inherently scalable. It naturally supports a varying number of reference images (see Fig. 1 and Fig. 4), without additional data preparation or modification to the training pipeline, allowing for richer, multi-subject customization. The stochasticity of the masked training strategy also allows Saber to robustly handle multiple reference views of the same subject (see Fig. 7).

We evaluate Saber on the OpenS2V-Eval [48] benchmark, where it consistently outperforms models [9, 10, 19, 20, 23, 27, 52] that were explicitly trained on R2V data. In addition, by simply adjusting the masking ratio during training, Saber can adapt to references depicting either foreground subjects or background scenes (see Fig. 1).

Our contributions are summarized as follows:

- We introduce Saber, the first zero-shot R2V framework that eliminates the need for explicit R2V data through masked training on video-text pairs.
- Saber surpasses previous R2V-data-trained methods on OpenS2V-Eval [48] and demonstrates strong generalization and scalability, paving the way for future research in scaling reference-to-video generation.

2. Related Work

2.1. Video Generation

The rapid progress of diffusion models [35] has greatly advanced video generation. Early methods [2, 3, 14] extended pre-trained text-to-image models [32, 35] with temporal modules to synthesize videos. More recently, large-scale models based on Diffusion Transformer [31] and trained on massive video-text datasets [5, 43] have achieved state-of-the-art, high-fidelity video generation [4, 12, 22, 29, 30, 42, 45, 47, 50]. Despite these advances, existing methods mainly focus on text-to-video and image-to-video tasks. While fine-grained, subject-driven control, as required by reference-to-video generation, remains a significant challenge, the complex, costly data construction for R2V datasets [27, 48] makes the large-scale training seen in T2V and I2V infeasible.

2.2. Reference-to-Video Generation

Building on the progress of text-to-video and image-to-video models [15, 17, 22, 42, 47], reference-to-video generation [18–20, 23, 27, 49] has seen significant advancement. Early studies [6, 11, 28, 36, 49, 54] mainly focused on human reference images, termed identity-preserving video generation, where facial or body features are injected into models to maintain identity consistency. Later, reference images extended from humans to various objects and backgrounds [18, 20, 27], allowing more flexible control. Some works [27, 48] also refer to this task as subject-consistent or subject-to-video generation, which is equivalent to R2V.

Representative works include Phantom [27], which learns cross-modal alignment with a joint text-image injection model using image-video-text triplet data. VACE [20] introduces a context adapter to process reference images and enable temporal-spatial feature interaction within a unified framework. SkyReels-A2 [10] builds an image-text joint embedding model to inject multi-element representations, balancing consistency and coherence. HunyuanCus-

tom [18] employs a LLaVA-based [25] fusion module and an image ID enhancement module to strengthen multimodal understanding and identity consistency. MAGREF [9] uses region-aware masking and pixel-wise concatenation for effective multi-reference interaction. PolyVivid [19] adds a 3D-RoPE enhancement and attention-inherited identity injection to reduce identity drift. BindWeave [23] leverages an MLLM [1] to link complex prompts with visual subjects, improving video generation quality.

However, a critical limitation unites these approaches: these methods rely on explicit reference image-video-text triplet datasets, which are costly and difficult to construct. Datasets such as OpenS2V-5M [48] and Phantom-Data [7] require complex construction pipelines, including candidate extraction, low-quality sample filtering, sample clustering, cross-pair matching, and expensive API calls for reference image generation. Such processes result in uncontrolled data quality, poor scalability, and high construction complexity. In contrast, we propose a zero-shot R2V framework trained solely on video-text pairs, achieving strong performance on public benchmarks.

3. Preliminary

3.1. Video Generation Models

Video generation models [15, 17, 22, 42, 47] have achieved remarkable progress and gained broad attention. Among these, the Wan Video series (*e.g.*, Wan2.1 [42]) is one of the most popular open-source frameworks. Our method builds on the Wan2.1-14B model [42], which consists of a variational autoencoder (VAE) [21], a transformer backbone [31, 41] and a text encoder (*i.e.*, umt5-xxl [8]). The VAE encodes videos into temporally and spatially compressed latents $\mathbf{z}_0$ and decodes them back to pixel space, reducing token count and computation. Wan2.1 trains the diffusion model Ψ using Flow Matching (FM) [24], where the forward process linearly interpolates between data and noise. For a time step $t \in [0, 1]$, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ is added to $\mathbf{z}_0$ to obtain $\mathbf{z}_t$, following $\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\epsilon$. The model is optimized to predict the target velocity with the following objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{z}_0,\epsilon,t,c}\left[\|(\mathbf{z}_0 - \epsilon) - \Psi_\theta(\mathbf{z}_t, t, c)\|_2^2\right], \quad (1)$$

where θ denotes the learnable parameters of the diffusion model Ψ, and c represents the condition features derived from the given text prompt and reference images.

3.2. Task Definition and Notations

Given K reference images $\{\mathbf{I}_k \in \mathbb{R}^{H_k \times W_k \times 3}\}_{k=1}^K$ and a text prompt $\mathbf{P}$, the reference-to-video method generates a video $\{\mathbf{I}_f \in \mathbb{R}^{H \times W \times 3}\}_{f=1}^F$ whose subjects preserve the identities and appearances of those in the reference images while following the instructions in text prompt $\mathbf{P}$. Here, H_k and W_k denote the height and width of the corresponding k-th reference image $\mathbf{I}_k$, while F, H, and W represent the number of frames, height, and width of the generated video.

4. Method

Our goal is to train a diffusion model Ψ_θ capable of generating videos $\{\mathbf{I}_f\}_{f=1}^F$ that preserve the identity and appearance of subjects in the given reference images $\{\mathbf{I}_k\}_{k=1}^K$ while following the provided text prompt $\mathbf{P}$. Previous methods [18–20, 23, 27] rely on reference image-video-text triplets, which are costly and hard to scale. In contrast, Saber achieves R2V capabilities using only video-text pairs, the same data paradigm used for T2V and I2V training.

Our core idea of Saber is to simulate the R2V task by replacing the explicitly collected reference images with randomly masked frames during training. This masked training strategy is supported by two key components to enhance robustness and visual quality: i) a series of mask augmentations designed to mitigate copy-paste artifacts, and ii) a tailored attention mechanism that guides the model to focus on relevant reference features.

We first introduce the construction of masked frames in Sec. 4.1, including mask generation and augmentation. Next, we present the model's architecture design in Sec. 4.2, detailing the input format and transformer-based attention mechanism. Finally, Sec. 4.3 describes the zero-shot R2V inference process.

4.1. Masked Frames as Reference

A standard R2V model learns to extract identity and appearance features from reference images $\{\mathbf{I}_k\}_{k=1}^K$ and inject them into the generated video $\{\mathbf{I}_f\}_{f=1}^F$. However, existing R2V datasets [7, 48] mainly consist of humans and common objects, leading to limited subject diversity and poor generalization. To address this, instead of relying on pre-collected reference images, we use randomly masked frames as dynamic substitutes during training. This strategy naturally introduces highly diverse reference samples, allowing the model to learn more effective subject integration and achieve stronger generalization.

As shown in Fig. 2, for each k-th reference image $\mathbf{I}_k$ randomly sampled from the video, we first use a **mask generator** to produce a binary mask $\mathbf{M}_k \in \{0, 1\}^{H \times W}$. To mitigate the copy-paste issue [27] in R2V tasks, we perform **mask augmentation** to disrupt spatial correspondence between the masked references and their corresponding video frames. Specifically, we apply an identical set of spatial augmentations to both $\mathbf{I}_k$ and $\mathbf{M}_k$, producing $\bar{\mathbf{I}}_k$ and $\bar{\mathbf{M}}_k$. The masked frame $\hat{\mathbf{I}}_k$ is then obtained as $\hat{\mathbf{I}}_k = \bar{\mathbf{I}}_k \odot \bar{\mathbf{M}}_k$. This process is repeated to create the full set of K masked frames $\{\hat{\mathbf{I}}_k\}_{k=1}^K$ that serve as the reference condition.

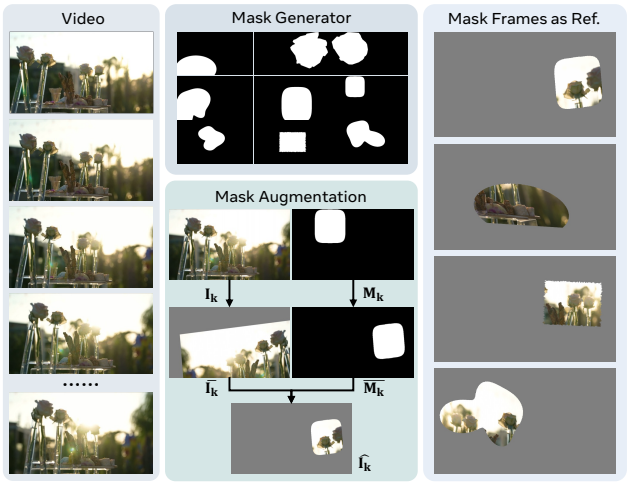

Figure 2. **Masked reference generation.** Given a video, the mask generator produces diverse random masks, which are then applied to each randomly sampled video frame with mask augmentation.

Mask Generator. We randomly select one mask type from predefined shape categories (*e.g.*, ellipse, Fourier blob, convex/concave polygon, *etc.*) to generate a binary mask $\mathbf{M} \in \{0,1\}^{H\times W}$ with a target foreground area ratio $r \in [r_{\min}, r_{\max}]$. Specifically, we first randomly select a foreground center. To ensure that the generated mask meets the desired foreground area ratio r, we define a continuous scale parameter for each shape category, where the mask's foreground area increases monotonically with the scale. A bisection search over the scale is then performed to satisfy the area ratio constraint. When pixel discretization prevents an exact match, small topology-preserving adjustments are applied: "growth" dilates background boundary pixels, while "shrinkage" erases background boundary pixels. This design ensures controllable foreground area ratios while maintaining diversity in mask shapes. Several mask examples are illustrated in Fig. 2 Top.

Mask Augmentation. We apply random affine transformations, including rotation, scaling, shear, translation, and optional horizontal flip, to both the image $\mathbf{I}_k$ and its mask $\mathbf{M}_k$, ensuring the masked region remains fully inside the frame. Transformation parameters are uniformly sampled within predefined ranges and validated to avoid boundary overflow. The same affine transformation is applied to the image and mask using bilinear and nearest-neighbor interpolation, respectively.

The reference code of the mask generator and augmentation are provided in the supplementary materials.

4.2. Model Design

After obtaining the masked frames $\{\hat{\mathbf{I}}_k\}_{k=1}^K$ as reference images, we detail our model design for the R2V task. We adopt a simple yet effective **input format** by concatenating reference images along the temporal dimension at the end

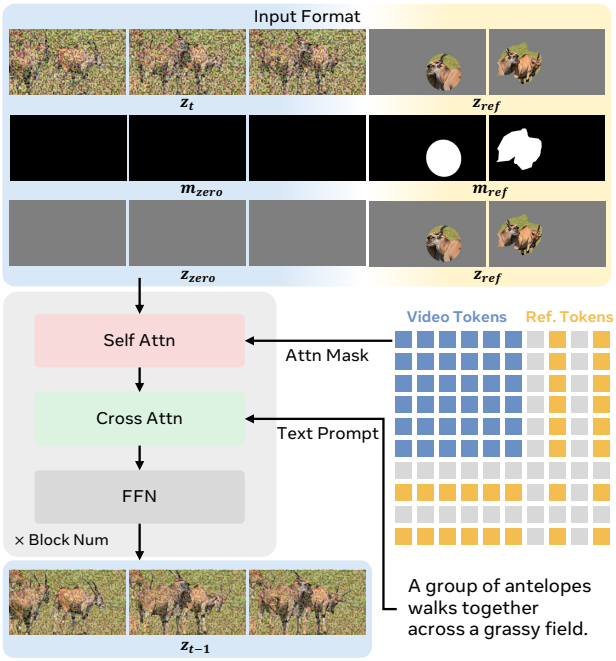

Figure 3. **Model design overview.** Masked frames serve as reference images and are concatenated to the video tokens in latent space. Self-attention enables interaction between video and reference tokens under the attention mask, while cross-attention incorporates text guidance for semantic alignment. The VAE, text encoder, and timestep components are omitted for clarity.

of the target video frames in latent space. This allows the model to manage the interaction between the target video latents and reference latents through the **attention mechanism** in each transformer block.

Input Format. Given a video-text pair $\{\mathbf{I}_f\}_{f=1}^F$ and $\mathbf{P}$, and the masked frames $\{\hat{\mathbf{I}}_k\}_{k=1}^K$ as reference images, we use the VAE to encode the video from pixel space into latent space, obtaining $\mathbf{z}_0 = \{\mathbf{z}_{\hat{f}} \in \mathbb{R}^{h\times w\times d}\}_{\hat{f}=1}^{\hat{F}}$. Here, $\hat{F} = \lfloor (F-1)/4 \rfloor + 1$, where 4 is the temporal compression ratio of the Wan2.1 VAE [42], h, w, and d denote the height, width, and feature dimension of the video latent, respectively. We obtain $\mathbf{z}_t$ with time step t following Sec. 3.1. For the reference images, we individually encode each $\hat{\mathbf{I}}_k$ using the VAE to obtain $\mathbf{z}_{\text{ref}} = \{\mathbf{z}_k \in \mathbb{R}^{h\times w\times d}\}_{k=1}^K$. Accordingly, each $\mathbf{M}_k$ is resized to match the latent space resolution, producing $\mathbf{m}_{\text{ref}} = \{\mathbf{m}_k \in \{0,1\}^{h\times w\times 4}\}_{k=1}^K$, where 0 indicates non-reference and 1 indicates reference region. The transformer input $\mathbf{z}_{\text{in}}$ is defined as in Eq. 2:

$$\mathbf{z}_{\text{in}} = \mathsf{cat}\begin{bmatrix} \mathsf{cat}[& \mathbf{z}_t & \mathbf{z}_{\text{ref}} &]_{\text{temporal}} \\ \mathsf{cat}[& \mathbf{m}_{\text{zero}} & \mathbf{m}_{\text{ref}} &]_{\text{temporal}} \\ \mathsf{cat}[& \mathbf{z}_{\text{zero}} & \mathbf{z}_{\text{ref}} &]_{\text{temporal}} \end{bmatrix}_{\text{channel}}, \tag{2}$$

where $\mathsf{cat}[\cdot]_{\text{temporal}}$ and $\mathsf{cat}[\cdot]_{\text{channel}}$ denote concatenation along the temporal and channel dimensions, respectively. $\mathbf{z}_{\text{zero}}$ is the VAE-encoded latent of a zero-value video, and

Table 1. **Quantitative results on the OpenS2V-Eval [48] benchmark.** Saber outperforms both closed-source and explicitly trained R2V methods, achieving the highest overall score in a zero-shot setting. It also attains the best NexusScore for subject consistency and competitive performance on GmeScore and NaturalScore.

Method	Total Score ↑	Aesthetics ↑	MotionSmoothness ↑	MotionAmplitude ↑	FaceSim ↑	GmeScore ↑	NexusScore ↑	NaturalScore ↑
Closed-source commercial R2V methods								
Pika2.1 [38]	51.88%	46.88%	87.06%	24.71%	30.38%	69.19%	45.40%	63.32%
Vidu2.0 [40]	51.95%	41.48%	90.45%	13.52%	35.11%	67.57%	43.37%	65.88%
Kling1.6 [37]	56.23%	44.59%	86.93%	41.60%	40.10%	66.20%	45.89%	74.59%
Explicit R2V data-based training methods								
SkyReels-A2 [10]	52.25%	39.41%	87.93%	25.60%	45.95%	64.54%	43.75%	60.32%
MAGREF [9]	52.51%	45.02%	93.17%	21.81%	30.83%	70.47%	43.04%	66.90%
Phantom-14B [27]	56.77%	46.39%	96.31%	33.42%	51.46%	70.65%	37.43%	69.35%
VACE-14B [20]	57.55%	47.21%	94.97%	15.02%	55.09%	67.27%	44.08%	67.04%
BindWeave [23]	57.61%	45.55%	95.90%	13.91%	53.71%	67.79%	46.84%	66.85%
Zero-shot R2V methods								
Saber (Ours)	57.91%	42.42%	96.12%	21.12%	49.89%	67.50%	47.22%	72.55%

$\mathbf{m}_{\text{zero}}$ is an all-zero mask, both shaped to match the temporal dimensions of the video part. Note that $\mathbf{z}_{\text{ref}}$ remains noise-free to preserve accurate conditioning.

Attention Mechanism. After obtaining the transformer input $\mathbf{z}_{\text{in}}$, we encode the text prompt $\mathbf{P}$ into text features $\mathbf{z}_{\text{P}}$ and jointly feed them, along with the time step t, into the transformer. Each transformer block consists of a self-attention, cross-attention, and feed-forward (FFN) modules.

In self-attention, the video and reference parts of $\mathbf{z}_{\text{in}}$ interact with each other. To avoid attending to non-reference regions, each $\mathbf{M}_k$ is resized to match the flattened latent shape to form an attention mask, where video tokens are bi-directionally attended, and only valid reference regions are attended in the reference part.

The self-attention output is then passed to the cross-attention module to interact with $\mathbf{z}_{\text{P}}$. Here, video tokens are guided by the text prompt, while reference tokens learn their semantic alignment, enabling the integration of reference image information under textual constraints. The FFN module refines the results, with the time step t injected into the latents for the control of the time step. After multiple transformer blocks, the transformer model outputs the predicted latent $\mathbf{z}_{t-1}$. The model design is shown in Fig. 3.

4.3. Zero-Shot Inference

In this section, we present the approach that enables the model trained with masked frames to perform zero-shot R2V inference. During inference, for each reference image $\mathbf{I}_k$, we first use a pre-trained object segmenter [34, 51] to extract the foreground subject region mask $\mathbf{M}_k$. We then normalize the reference image $\mathbf{I}_k$ to the range of $[-1, 1]$ and fill the masked background regions with zeros (color gray). Notably, this segmentation step is flexible. If a reference image is intended to provide a background scene rather than a foreground subject, segmentation is skipped. In this case, we use the full, unmasked reference image and an all-ones mask $\mathbf{M}_k$, treating the entire image as the reference region.

Both $\mathbf{I}_k$ and $\mathbf{M}_k$ are processed by a resize-and-padding operation, which scales $\mathbf{I}_k$ from its original size (H_k, W_k) to fit within the target video size (H, W) while preserving the aspect ratio and pads the remaining area with zeros to produce a centered reference image of size (H, W). Finally, the processed reference image and mask are fed into the model following the input format in Sec. 4.2 and are used for prediction following the Wan inference pipeline [42].

5. Experiments

5.1. Datasets, Metrics and Implementation Details

Datasets. Benefiting from the masked training strategy, Saber is trained exclusively on video-text pair datasets, enabling the use of data from T2V and I2V sources. Specifically, we employ the ShutterStock Video [39] dataset and generate captions for all video clips using Qwen2.5-VL-Instruct [1], thus constructing the corresponding video-text pairs for training.

Metrics. To ensure fair comparison, we adopt the OpenS2V-Eval [48] benchmark and follow its official protocol for fine-grained evaluation of reference-to-video generation. The benchmark contains 180 prompts across seven categories, spanning single-reference (face, human, entity) and multi-reference (multi-face, multi-human, human-entity) scenarios. We report automated metrics where higher scores indicate better performance, including Aesthetics for visual quality, MotionSmoothness for temporal coherence, MotionAmplitude for motion magnitude, and FaceSim for identity preservation. In addition, we use three OpenS2V-Eval metrics, NexusScore, NaturalScore, and GmeScore, which measure subject consistency, naturalness, and text-video alignment, respectively.

Implementation Details. Saber is finetuned from the Wan2.1-14B [42] model using our proposed masked training strategy on video-text pair datasets. For the mask generator, we adopt a probabilistic sampling strategy for the foreground area ratio r: with 10% probability, we set $r \in [0, 0.1]$ to simulate minimal and no reference information, enabling the model to handle varying numbers of reference images; with 80% probability, we set $r \in [0.1, 0.5]$ to

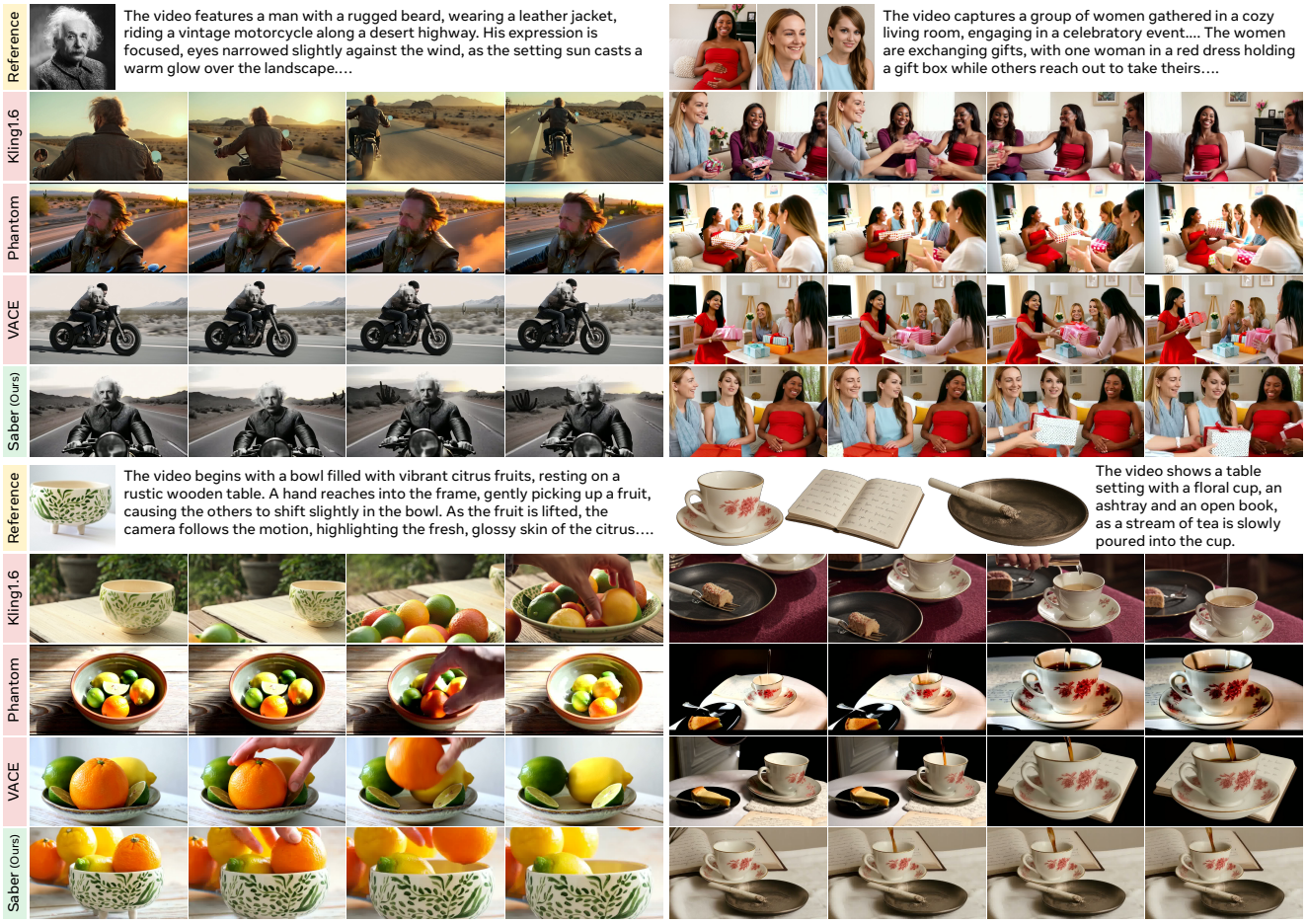

Figure 4. **Qualitative comparison with existing R2V methods.** We compare Saber with Kling1.6 [37], Phantom [27], and VACE [20] across four scenarios: single/multiple human and object references. Saber accurately preserves subject identity and appearance, integrates multiple references coherently, and generates smoother, more visually consistent videos.

represent typical primary subjects; and with the remaining 10% probability, we set $r \in [0.5, 1.0]$ to help the model learn from large reference images or background scenes. For mask augmentation, we randomly apply rotation within $[-10, 10]$ degree, scaling in the range $[0.8, 2.0]$, horizontal flipping with 50% probability, and shearing within $[-10, 10]$ degree. We found these augmentations to be empirically effective at overcoming copy-paste artifacts. We train our model with the objective defined in Eq. 1, using the AdamW optimizer with $1e^{-5}$ learning rate and a global batch size of 64. During inference, we use BiRefNet [51] to segment the foreground subjects from the reference images. Following the standard setting of Wan2.1 [42], we generate videos with 50 denoising steps and a CFG [16] guidance scale of 5.0.

5.2. Quantitative Results

We follow Yuan et al. [48] and conduct a comprehensive evaluation on OpenS2V-Eval [48] benchmark, with the results presented in Tab. 1. The table compares three types of methods: closed-source commercial R2V methods, explicitly trained R2V methods, and our zero-shot R2V approach. Compared with the closed-source commercial method Kling1.6 [37], our model achieves a 1.68% higher total score. Among methods trained on explicit R2V datasets, our method surpasses Phantom [27] by 1.14%, VACE [20] by 0.36%, and BindWeave [23] by 0.30%. While these methods rely on costly explicit R2V datasets that are difficult to scale, our approach uses only text-video pairs with a masked training strategy, achieving the best overall performance in a zero-shot setting.

Among all sub-metrics, NexusScore best represents R2V performance by measuring subject consistency. Saber achieves the highest NexusScore, exceeding Phantom by 9.79%, VACE by 3.14%, and BindWeave by 0.36%. This shows that the masked training strategy effectively learns subject features from video-text pairs in a zero-shot setting, outperforming all R2V-data-based models. Our method also achieves competitive results on GmeScore (text-video alignment) and NaturalScore (video naturalness).

Table 2. **Ablation study.** i) Masked training outperforms training on the OpenS2V-5M [48] (w/o masked training), demonstrating the advantage of the masked training strategy. ii) Using only a single mask type reduces total score, while combining all types performs best, showing the importance of mask diversity. iii) Fixing the foreground area ratio ($r = 0.3$) leads to a further drop, indicating limited mask variation harms generalization.

Method	Total Score ↑	GmeScore ↑	NexusScore ↑	NaturalScore ↑
Saber	57.91%	67.50%	47.22%	72.55%
w/o masked training	56.24%	67.27%	45.33%	70.19%
ellipse only	54.56%	67.98%	40.28%	72.54%
fourier only	56.33%	67.25%	44.82%	72.46%
polygon only	56.49%	67.41%	45.24%	72.21%
fixed $r = 0.3$	51.73%	67.12%	39.20%	69.55%

5.3. Qualitative Results

We further conduct a qualitative comparison between Saber and other methods (Kling1.6 [37], Phantom [27] and VACE [20]) across various visual scenarios, as shown in Fig. 4. i) In the top-left (single human reference), both Kling1.6 and Phantom fail to embed the reference subject into the generated video, leading to inconsistent facial appearances. VACE suffers from a copy-paste issue, directly overlaying the face from the reference image. In contrast, Saber generates a video with a consistent and text-aligned facial identity. ii) In the bottom-left (single object reference), Kling1.6 produces a bowl with an incorrect leg structure, while Phantom and VACE fail to capture both the shape and appearance of the bowl from the reference. In contrast, our method, benefiting from the rich diversity of masked frames during masked training, accurately integrates the bowl's shape and appearance into the generated video. iii) In the top-right (multiple human references), Kling1.6 embeds only one subject, Phantom duplicates the same identity twice, and VACE fails to inject facial information from the references. In contrast, Saber incorporates all three subjects and generates a coherent, natural video. iv) In the bottom-right (multiple object references), Kling1.6 produces incorrect patterns on the cup, while Phantom and VACE generate correct cup textures but omit the ashtray, with VACE also showing temporal discontinuity. Saber, on the other hand, generates all referenced objects correctly and maintains smooth, consistent video quality.

5.4. Ablation Study

We conduct a series of ablation studies to analyze the key components of Saber, including the masked training strategy, mask generator, mask augmentation, and attention mask in the attention mechanism.

The Effect of Masked Training. To evaluate the effect of masked training, we finetune our model on the OpenS2V-5M [48] dataset using the same architecture. As shown in Tab. 2, masked training improves the total score by +1.67%, indicating that it strengthens subject representation learning and reduces overfitting to specific reference cues.

Figure 5. **Effect of mask augmentation.** Without mask augmentation, the model shows copy-paste artifacts by directly copying reference content. Applying augmentation enables more natural and coherent video generation.

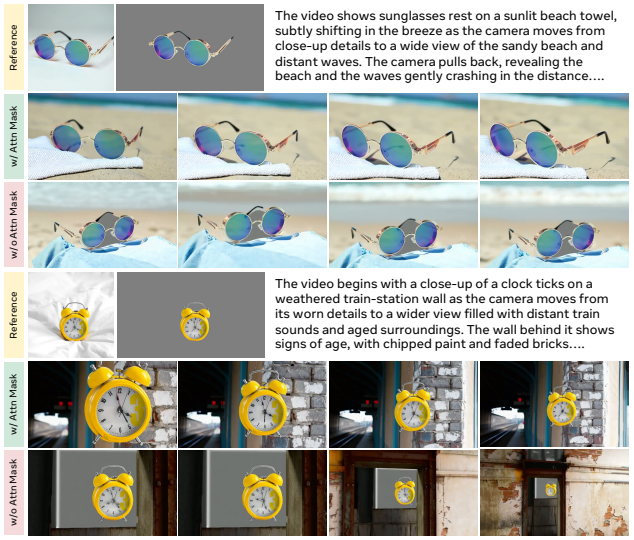

Figure 6. **Effect of the attention mask.** Removing the attention mask introduces gray artifacts around subjects, while applying it ensures clean separation from the gray background and smoother, more natural video results.

The Effect of Mask Generator. In Tab. 2, we first analyze the mask type by training the model using only one type at a time. Using only ellipse, Fourier, or polygon masks reduces total score by 3.35%, 1.58%, and 1.42%, respectively, whereas combining all types yields the best results, showing that mask diversity is crucial for masked training. We also fix the foreground area ratio r to 0.3, which leads to a 6.18% drop, indicating that restricting mask variation limits generalization.

The Effect of Mask Augmentation. We evaluate the effect of mask augmentation by training the model with and without it. As shown in Fig. 5, without augmentation, the model exhibits severe copy-paste artifacts, directly placing the T-shirt upright on the rock. With augmentation (rotation, scaling, flipping, and shearing), the T-shirt naturally lies on the rock surface, resulting in more realistic and coherent compositions. This demonstrates that geometric diversity in masking is crucial for natural video generation.

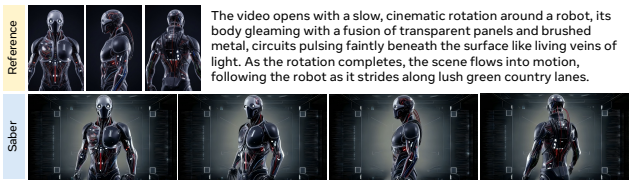

Figure 7. **Qualitative results on multiple reference images of the same subject.** Given the front, side, and back views of a robot as reference, Saber correctly recognizes them as the same subject and integrates multi-view appearance features into a coherent video, accurately preserving fine structural and surface details.

The Effect of the Attention Mask. Finally, we examine the role of the attention mask in our attention mechanism, which constrains reference-video token interaction. As shown in Fig. 6, removing the attention mask introduces visible gray artifacts around the subject regions, as the model fails to correctly extract subjects from masked reference images (*e.g.*, gray areas behind sunglasses or clocks). Incorporating the attention mask effectively resolves these issues, leading to cleaner subject separation, smoother blending, and improved overall video quality.

5.5. Emergent Abilities

In this section, we explore several interesting capabilities of Saber that emerged from its training strategy, demonstrating robustness beyond the standard R2V task.

Single Subject Multiple Views. We test Saber's ability to handle multiple reference images corresponding to different views of the same subject. As shown in Fig. 7, we use the front, side, and back views of a robot as reference inputs to Saber. The results show that Saber successfully understands that all reference images depict the same subject (the robot) and integrates the appearance feature from different views into a single coherent video subject. Despite the robot's complex surface details and wiring structure, Saber accurately captures and synthesizes these visual characteristics into the generated video.

Cross-modal Alignment. We also evaluate the model's alignment between reference images and text prompts by swapping subject descriptions and observing the corresponding video changes. As shown in Fig. 8, when altering prompts such as "a man wearing a blue shirt seated" (Type 1) and "a man wearing a black vest seated" (Type 2), Saber correctly aligns subjects with their descriptions. Similarly, when swapping the positions of a man and woman in the second case, the model accurately reflects the change. This demonstrates that, as described in Sec. 4.2, the interaction between video and reference tokens through self-attention, followed by cross-attention with text prompt features, enables robust reference image-text alignment.

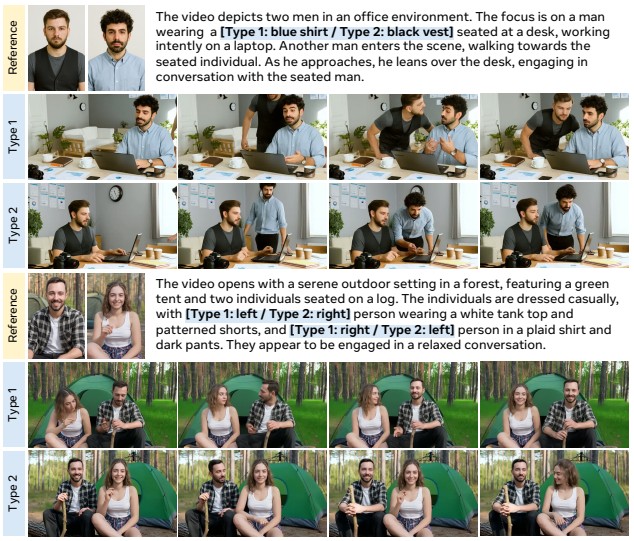

Figure 8. **Qualitative results of cross-modal alignment between reference images and text prompts.** By swapping subject descriptions in the prompts (*e.g.*, clothing color or subject positions), Saber accurately reflects the corresponding visual changes, demonstrating robust alignment between reference images and textual descriptions through its attention mechanisms.

6. Conclusion

In this work, we present Saber, a scalable zero-shot framework for reference-to-video generation that eliminates the need for explicitly R2V datasets. Trained solely on large-scale video-text pairs, Saber leverages a masked training strategy, a tailored attention mechanism, and mask augmentation to achieve identity-consistent, natural, and coherent video generation. It further scales to multiple references, supporting both multi-identity and multi-view inputs without additional data preparation or changes to the training pipeline. Extensive experiments on the OpenS2V-Eval benchmark demonstrate that Saber consistently outperforms methods trained on explicit R2V data. These results show that effective R2V models can be trained without dedicated datasets, paving the way for future research in scalable and generalizable reference-to-video generation.

Limitations. While Saber achieves strong zero-shot performance and scalability, several limitations remain. First, R2V generation may collapse when the number of reference images increases significantly (*e.g.*, 12), resulting in fragmented compositions where references are combined without coherent understanding. Second, Saber primarily focuses on identity preservation and visual coherence, while fine-grained motion control and temporal consistency under complex prompts remain challenging. Future work can explore more effective integration of numerous reference images into unified video generation, as well as adaptive guidance to further improve controllability and realism in reference-to-video generation.

References

[1] 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*, 2025. 3, 5

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *CVPR*, 2024. 2

[4] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *CVPR*, 2025. 2

[5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024. 2

[6] Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *ECCV*, 2024. 2

[7] Zhuowei Chen, Bingchuan Li, Tianxiang Ma, Lijie Liu, Mingcong Liu, Yi Zhang, Gen Li, Xinghui Li, Siyu Zhou, Qian He, et al. Phantom-data: Towards a general subject-consistent video generation dataset. *arXiv preprint arXiv:2506.18851*, 2025. 2, 3

[8] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023. 3

[9] Yufan Deng, Xun Guo, Yuanyang Yin, Jacob Zhiyuan Fang, Yiding Yang, Yizhi Wang, Shenghai Yuan, Angtian Wang, Bo Liu, Haibin Huang, et al. Magref: Masked guidance for any-reference video generation. *arXiv preprint arXiv:2505.23742*, 2025. 2, 3, 5

[10] Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025. 2, 5

[11] 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. *arXiv preprint arXiv:2509.01362*, 2025. 2

[12] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 2

[13] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[15] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 3

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[17] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023. 2, 3

[18] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025. 2, 3

[19] Teng Hu, Zhentao Yu, Zhengguang Zhou, Jiangning Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Polyvivid: Vivid multi-subject video generation with cross-modal interaction and enhancement. In *NeurIPS*, 2025. 2, 3

[20] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *ICCV*, 2025. 2, 3, 5, 6, 7

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3

[23] 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. *arXiv preprint arXiv:2510.00438*, 2025. 2, 3, 5, 6

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3

[26] Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, et al. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280*, 2024. 2

[27] 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 *ICCV*, 2025. 2, 3, 5, 6, 7

[28] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2024. 2

[29] Zhiheng Liu, Ka Leong Cheng, Xi Chen, Jie Xiao, Hao Ouyang, Kai Zhu, Yu Liu, Yujun Shen, Qifeng Chen, and Ping Luo. Manganinja: Line art colorization with precise reference following. In *CVPR*, 2025. 2

[30] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. In *CVPR*, 2025. 2

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[33] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 2

[34] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 5

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[36] Liao Shen, Wentao Jiang, Yiran Zhu, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. *arXiv preprint arXiv:2510.14255*, 2025. 2

[37] Kling Team. Kling1.6 elements to video. https://app.klingai.com/global/image-to-video/multi-id/new, 2025. 5, 6, 7

[38] Pika Team. Pika2.1 consistent character video. https://pollo.ai/consistent-character-video, 2025. 5

[39] ShutterStock Team. Shutterstock video dataset. https://www.shutterstock.com. 2, 5

[40] Vidu Team. Vidu2.0 reference to video. https://www.vidu.com/ai-reference-to-video, 2025. 5

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[42] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 4, 5, 6

[43] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *CVPR*, 2025. 2

[44] Xiyu Wang, Yufei Wang, Satoshi Tsutsui, Weisi Lin, Bihan Wen, and Alex Kot. Evolving storytelling: benchmarks and methods for new character customization with diffusion models. In *ACM MM*, 2024. 2

[45] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024. 2

[46] Bowen Xue, Qixin Yan, Wenjing Wang, Hao Liu, and Chen Li. Stand-in: A lightweight and plug-and-play identity control for video generation. *arXiv preprint arXiv:2508.07901*, 2025. 2

[47] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 3

[48] Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025. 2, 3, 5, 6, 7

[49] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *CVPR*, 2025. 2

[50] Yifu Zhang, Hao Yang, Yuqi Zhang, Yifei Hu, Fengda Zhu, Chuang Lin, Xiaofeng Mei, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Waver: Wave your way to lifelike video generation. *arXiv preprint arXiv:2508.15761*, 2025. 2

[51] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. 5, 6

[52] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Nanxuan Zhao, Jing Shi, and Tong Sun. Sugar: Subject-driven video customization in a zero-shot manner. *arXiv preprint arXiv:2412.10533*, 2024. 2

[53] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024. 2

[54] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. In *CVPR*, 2025. 2