Multi-subject Open-set Personalization in Video Generation

Tsai-Shien Chen[1,2,*] Aliaksandr Siarohin[1] Willi Menapace[1] Yuwei Fang[1] Kwot Sin Lee[1]
Ivan Skorokhodov[1] Kfir Aberman[1] Jun-Yan Zhu[3] Ming-Hsuan Yang[2] Sergey Tulyakov[1]
[1]Snap Inc. [2]UC Merced [3]CMU

https://snap-research.github.io/open-set-video-personalization

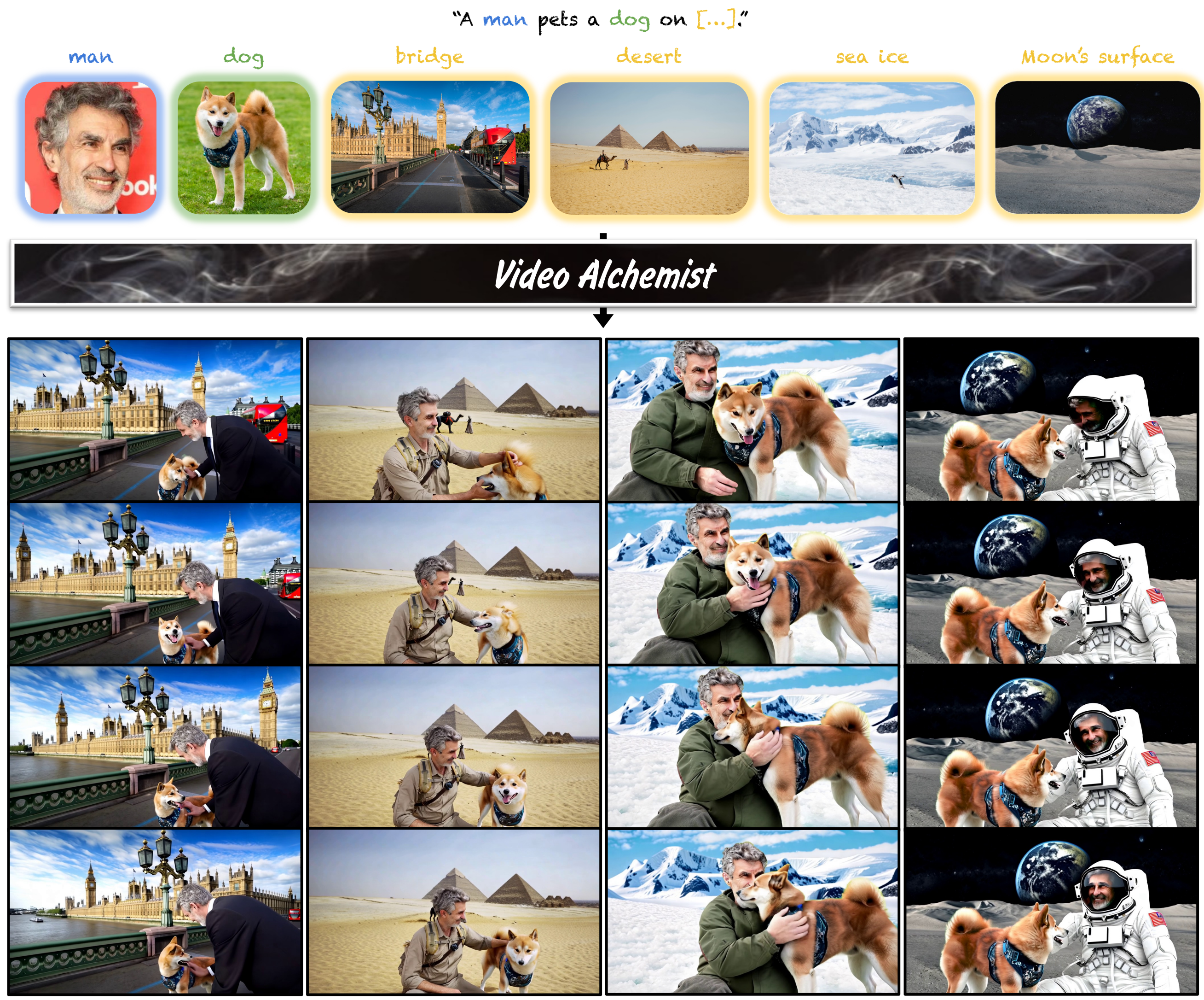

Figure 1. Given a text prompt as well as reference images for each subject (*i.e.*, *man*, *dog*) and background images (*i.e.*, *bridge*, *desert*, *sea ice*, *Moon's surface*), *Video Alchemist* synthesizes natural motions while preserving subject identity and background fidelity.

Abstract

Video personalization methods allow us to synthesize videos with specific concepts such as people, pets, and places. However, existing methods often focus on limited domains, require time-consuming optimization per subject, or support only a single subject. We present Video Alchemist—a video model with built-in multi-subject, open-set personalization capabilities for both foreground objects and background, eliminating the need for time-consuming

*This work was done while interning at Snap.

test-time optimization. Our model is built on a new Diffusion Transformer module that fuses each conditional reference image and its corresponding subject-level text prompt with cross-attention layers. Developing such a large model presents two main challenges: dataset and evaluation. First, as paired datasets of reference images and videos are extremely hard to collect, we sample selected video frames as reference images and synthesize a clip of the target video. However, while models can easily denoise training videos given reference frames, they fail to generalize to new contexts. To mitigate this issue, we design a new automatic data construction pipeline with extensive image augmentations. Second, evaluating open-set video personalization is a challenge in itself. To address this, we introduce a personalization benchmark that focuses on accurate subject fidelity and supports diverse personalization scenarios. Finally, our extensive experiments show that our method significantly outperforms existing personalization methods in both quantitative and qualitative evaluations.

1. Introduction

Diffusion models [28, 63, 64] have enabled us to synthesize realistic videos with natural motions from text prompts [4, 6, 29, 48, 62]. This level of quality and realism paves the way for personalization—the ability to generate videos containing specific objects and people in unseen contexts or backgrounds. Multiple methods have been proposed to generate content with specific people or pets, but they remain limited to closed-set object categories. Some only support human faces [26, 45] or a single subject [32, 78, 80, 88], while others only work on foreground objects [8, 9, 76]. Moreover, many of these methods require costly test-time optimization [45, 78, 80].

In this paper, we present *Video Alchemist*, a video generation model with extensive personalization capabilities. Our model supports the customization of multiple subjects and open-set entities, including both foreground objects and background. Importantly, our method does not require fine-tuning to incorporate new concepts. Figure 1 shows videos personalized for two subjects across four backgrounds. *Video Alchemist* is built on new Diffusion Transformer modules [53] tailored for personalization. Each module uses two cross-attention layers: one to integrate the text prompt describing the entire video and the other to incorporate the embeddings of each reference image. To achieve multi-subject conditioning, we employ a simple yet effective subject-level fusion, blending the word description of each subject with its image embeddings.

But how can we collect data to train our model? Ideally, it requires a dataset of videos and images with many subjects, each captured with varying lighting, background, and pose. Unfortunately, collecting such a dataset for open-set entities is challenging at best and impossible at worst. Alternatively, we can extract reference images and target video clips from the same video. However, this approach comes with a significant drawback—factors unrelated to identity still have a very high correlation across different video frames, leading to what we term the *copy-and-paste* effect. This issue is commonly seen in reconstruction-based methods, such as IP-Adapter [85], as shown in Figure 5. As a result, the model struggles to synthesize diverse videos with unseen backgrounds, lighting, and pose. To alleviate this overfitting, we design a data construction pipeline to automatically extract object segments from target videos and craft personalization-specific data augmentation to ensure that the model focuses on the subject identity of the reference images. Experiments show that training with the proposed augmentation can significantly mitigate the *copy-and-paste* effect, as shown in Figure 6.

Another challenge is the lack of a suitable benchmark for evaluating multi-subject video personalization. Typically, we evaluate video personalization results by computing a similarity score between the generated video and the reference images [32, 58, 85, 88]. Unfortunately, this metric does not apply to multiple entities, as it cannot focus on each subject separately. To address these limitations, we introduce *MSRVTT-Personalization*, a comprehensive and robust evaluation protocol for personalization tasks. This new benchmark facilitates evaluation across various conditioning modes, including conditioning on face crops, single or multiple arbitrary subjects, and combinations of foreground objects and backgrounds. Unlike image-level similarity, we evaluate the subject fidelity of each object segment. The experiments demonstrate that *Video Alchemist* outperforms existing personalization methods regarding both quantitative and qualitative evaluations. In addition, we conduct an extensive ablation study to verify the effectiveness of our proposed components.

Our contributions can be summarized as follows:

- We present *Video Alchemist*, a new video generation model that supports multi-subject, open-set personalization for both foreground objects and background.
- We carefully curate a large-scale training dataset and introduce training techniques to reduce model overfitting.
- We introduce *MSRVTT-Personalization*, a new video personalization benchmark, providing various conditioning modes and accurate measurement of subject fidelity.

2. Related Work

Diffusion Video Models. Diffusion models [28, 29, 56, 63, 64] have demonstrated impressive capabilities in generating realistic images. Building on this success, recent studies have explored their applications in text-conditioned video synthesis [4, 6, 22, 23, 44, 46, 48, 60, 62, 84, 87]. ImagenVideo [60] and Make-A-Video [62] use cascaded temporal and spatial upsamplers for video generation. VideoLDM [4] fine-tunes a pre-trained latent image generator and decoder to produce temporally coherent videos. Dif-

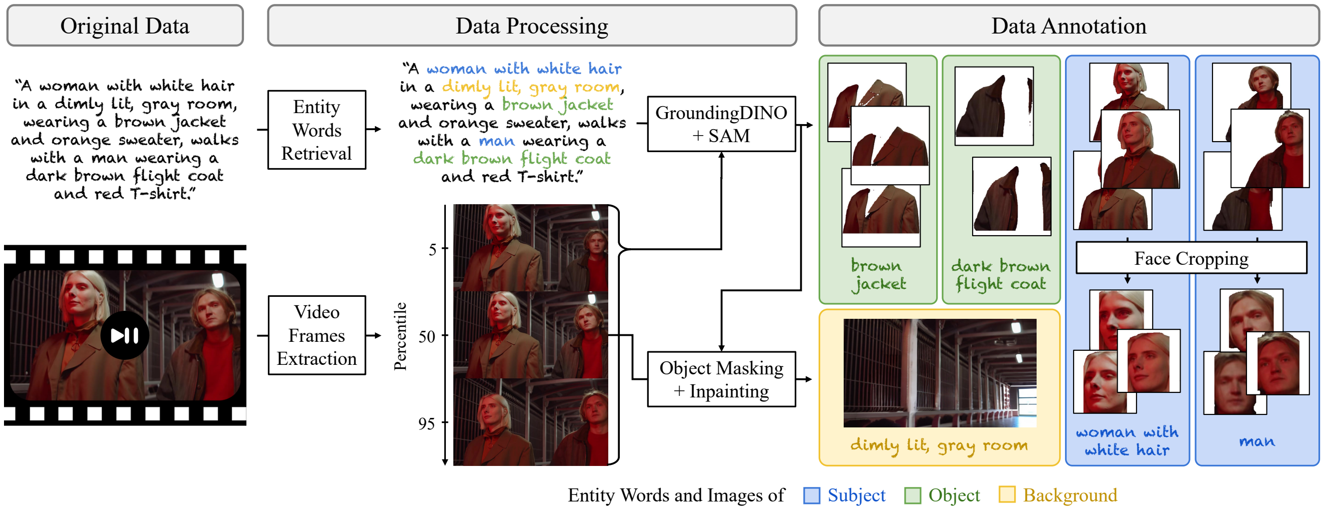

Figure 2. **Dataset collection pipeline for video personalization.** We construct our training dataset using video and caption pairs through three steps. First, we identify three categories of entity words from the caption: subject, object, and background. Next, we use these entity words to localize and segment the target subjects and objects in three selected video frames. Finally, we extract a clean background image by removing the subjects and objects from the middle frame.

ferently from previous models based on the U-Net [57] architecture, SnapVideo [48] adapts the FiT [10] and scales to billion-parameter models. More recently, Sora [6] adopts the Diffusion Transformer [53] to achieve high-resolution, long video synthesis. While these studies have shown significant progress, using text prompts alone confines the generated content to what can be described in words.

Personalized Image Generation. This task aims to customize a generative model to new concepts and subjects using a few input images [2, 19, 24, 25, 33, 36, 51, 58, 61, 68, 73, 75, 85]. For example, DreamBooth [58] optimizes the entire text-to-image model for each subject, while Textual inversion [19] learns a text embedding for each subject and uses the embedding to generate novel images. Custom Diffusion [36] learns to compose multiple concepts, each represented by text embedding and cross-attention weights. However, these optimization-based models require finetuning weights or optimizing embeddings for every new concept, which is inevitably slow and prone to overfitting.

Recent studies have explored encoder-based methods to reduce test-time finetuning [1, 14, 20, 37, 59, 61, 71, 77, 82, 85]. IP-adapter [85] learns a lightweight decoupled cross-attention mechanism for image conditioning. InstanceBooth [61] trains an image encoder to convert reference images into textual tokens and introduces adapter layers to retain identity details. Our model also uses an encoder, but we focus on video personalization with multiple subjects.

Personalized Video Generation. Several works have extended model personalization techniques for videos [12, 18, 26, 32, 42, 45, 47, 76, 78, 80, 86, 88]. DreamVideo [78] uses an optimization-based strategy, training an image adapter to capture the subject's appearance and a motion adapter to model dynamics. In contrast, StoryDiffusion [88] adopts an optimization-free approach with a consistent self-attention mechanism and a semantic motion predictor to ensure smooth transitions and consistent subjects.

However, most of the existing methods focus on limited domains. Some are limited to face personalization [26, 45] or a single subject from specific categories [32, 78, 80, 86, 88], while others focus solely on foreground objects [8, 9, 76]. In contrast, we introduce a video model that supports the customization of multiple open-set entities across both foreground objects and background. Closely related to our work, VideoDrafter [42] achieves open-set video personalization in two stages: text-to-image personalization and first-frame animation. In contrast, our end-to-end method alleviates poor subject consistency in long video synthesis, a notable limitation of first-frame animation.

3. Methodology

Our goal is to learn a generative video model conditioned on a text prompt and a set of images representing each entity word in the prompt.

3.1. Dataset Collection

As shown in Figure 2, we curate our dataset in three steps.

Retrieving Entity Words. To achieve multi-subject personalization, we use a large language model [31] to retrieve multiple entity words from a single caption. Specifically, we define three types of entity words: subject (*e.g.*, human, animal), object (*e.g.*, car, jacket), and background (*e.g.*, room, beach). Subjects and objects are supposed to be clearly visible in the video. Next, we adapt several criteria to filter and enhance the quality of the training dataset. For example, we exclude videos with any subject entity word in plural form (*e.g.*, a group of people, several dogs) to avoid ambiguity in personalization. Another example is that we

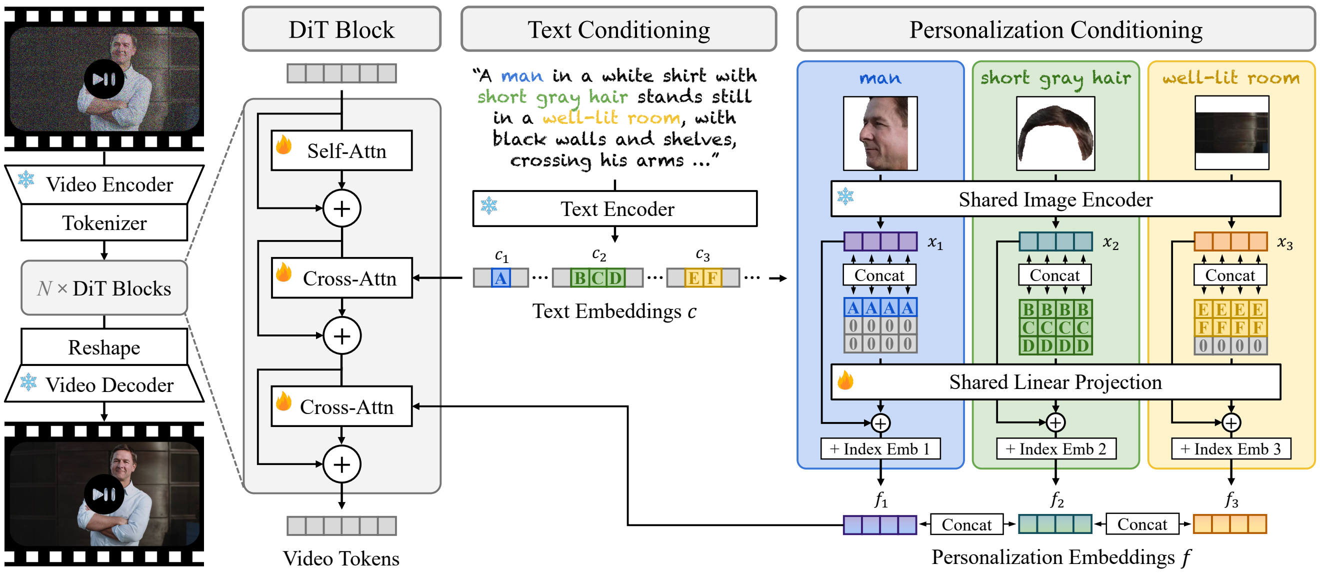

Figure 3. **Model architecture.** Our model is a latent DiT [53], where we first encode a video into video tokens and denoise them with a deep cascade of DiT blocks in the latent space. Each DiT block includes an additional cross-attention operation with personalization embeddings $f = \text{Concat}(f_1, \ldots, f_n, \ldots, f_N)$, where f_n fuses the embeddings of both the reference image x_n and the corresponding entity word c_n. Each square in the figure represents a 1-D token.

remove videos without subject entity words, as their dynamics is often dominated by meaningless camera movements. More details can be found in Appendix A.2.

Preparing Subject Images. Next, we select three frames from the beginning, middle, and end of the video (in the 5%, 50%, and 95% percentiles). The motivation is to capture the target subject or object with different poses and lighting conditions. Subsequently, we apply GroundingDINO [40] to each frame to detect the bounding boxes. These bounding boxes are then used by SAM [35] to segment the mask regions corresponding to each entity. Moreover, for reference images depicting humans, we apply face detection [74] to extract face crops.

Preparing Background Image. Lastly, we create a clean background image by removing the subjects and objects. Since SAM [35] occasionally produces imprecise boundaries, we dilate the foreground mask before applying an inpainting algorithm [56]. We use the background entity word as the positive prompt and use "*Any human or any object, complex pattern, and texture*" as the *negative* prompt. To ensure background consistency, we only use the middle frame of each video sequence.

3.2. Video Personalization Model

We learn *Video Alchemist* by denoising the video using a text prompt, reference images, and their corresponding entity words as conditions.

Video Generation Backbone. As illustrated in Figure 3, our model is a latent Diffusion Transformer (DiT) [53], where we first compress a video into a latent representation using an autoencoder [84] and encode it into a sequence of 1-D video tokens with a tokenizer [34]. Next, we add Gaussian noise to obtain a noisy sample and learn a denoising network following the rectified flow formulation [39, 41].

Our network is a deep cascade of DiT blocks. Unlike vanilla DiT designs, our module supports built-in personalization capability by combining information from both text and image conditioning. Our DiT block includes three layers: one multi-head self-attention [72], followed by two multi-head cross-attention for text and personalization conditioning, respectively. We use RoPE [66] positional embeddings in self-attention due to its effectiveness irrespective of number of video tokens and use self-conditioning [11] to enhance visual quality. We further adopt flash attention [15] and the fused layer norm [49] to accelerate the model training and inference.

Binding of Image and Word Concepts. For multi-subject personalization, the model can be conditioned on different subjects, each represented by one or more reference images. Consequently, providing the binding between corresponding text tokens and image tokens is critical. As shown in the second row of Figure 6, without such binding information, the model tends to apply image conditioning to an incorrect subject, such as placing a reference human face on a dog.

We provide the binding through the form of personalization embeddings $f = \text{Concat}(f_1, \ldots, f_n, \ldots, f_N)$, where f_n encodes information from both the reference image and the corresponding entity word. Here, N is the number of reference images. Specifically, to produce embeddings f_n, we first encode the image as image tokens $x_n \in \mathcal{R}^{l \times d}$, using a shared, frozen image encoder [50]. Here, l denotes the number of tokens per reference image, and d denotes the dimension of each token.

Next, we retrieve word tokens c_n from the text embeddings c (encoded from the text) and flatten c_n into a 1-D embedding. Since the number of tokens of an entity word varies, we zero-pad or truncate the word embeddings to a consistent length. To bind the image and word tokens, we replicate the flattened word tokens l times and concatenate them with the image tokens along the channel axis. Finally, we pass it to a linear projection module, apply a residual connection with the image tokens x_n, and add a learnable image index embedding to separate tokens from different images. Different tokens from the same image will share the same image index embedding.

Personalization Conditioning. The personalization embeddings f are then used to compute cross-attention with video tokens. Although IP-Adapter [85] uses a single decoupled cross-attention layer for both text and image conditioning, we find empirically that separate cross-attention layers perform better in our case. This is likely because our multi-image conditioning introduces a longer sequence of image tokens. Thus, mixing text and image tokens in a shared layer causes the image tokens to dominate, reducing alignment with the text prompt.

We train the model in two stages. In the first stage, we train the model with only one cross-attention for text conditioning. Next, we introduce the additional cross-attention for personalization conditioning and fine-tune the whole model with warmup. Appendix B details model training.

3.3. Reducing Model Overfitting

We learn *Video Alchemist* by denoising the training videos using the selected and segmented frames as conditions. However, this approach often leads to overfitting, where the model learns to focus on the lighting, pose, occlusion, and camera viewpoint of the reference subject (*ref*) rather than its identity. Specifically, we find that:

- If *ref* is high-resolution, the model generates a large subject close to the camera.
- If *ref* is occluded, the model generates other objects that occlude the target subject.
- If *ref* is cropped, the model places the subject at the edge, causing it to be cropped by the video boundary.
- The model often replicates the subject's pose and lighting conditions from *ref*.
- If multiple *ref*s represent the same subject with similar poses, the model produces a subject with minimal motion.

This overfitting often leads to the *copy-and-paste* effect, where the model directly replicates the reference images in the video without introducing pose and lighting variations. This effect is commonly observed in reconstruction-based methods, such as IP-Adapter [85], as shown in Figure 5.

To alleviate these issues, we apply data augmentation to the reference images. Specifically, we use downscaling and Gaussian blurring to prevent overfitting to the image resolution, color jittering and brightness adjustment to mitigate overfitting on the lighting conditions, and random horizontal flip, image shearing, and rotation to weaken overfitting on the subject's pose. The key idea is to guide the model to focus on the subject's identity rather than learning the unintended information leakage from the reference images. More details on the proposed image augmentations can be found in Appendix A.3.

4. Experiments

Section 4.1 introduces *MSRVTT-Personalization*, a comprehensive benchmark for personalization. Section 4.2 provides quantitative and qualitative comparisons with state-of-the-art methods. Section 4.3 discusses the ablation study of our model training and architecture designs. Appendix A contains details of the training dataset and augmentations. Appendix B includes details of model architecture, training, and inference. Finally, we include more generated samples in Appendix C.

4.1. MSRVTT-Personalization Benchmark

Existing methods [58, 78, 85, 88] evaluate subject preservation using image similarity [16, 50, 55] between reference and generated images or videos. However, these metrics are ineffective for multiple subjects, as image-level similarity fails to focus on the target subject. To address this issue, we introduce *MSRVTT-Personalization* to provide a more comprehensive and accurate evaluation of personalization tasks. It supports various conditioning scenarios, including conditioning on face crops, single or multiple subjects, and both foreground objects and backgrounds.

We construct the test benchmark based on MSR-VTT [83] and process the dataset in three steps. First, we use TransNetV2 [65], a shot boundary detection algorithm, to split long videos into multiple clips and apply an internal captioning algorithm to create detailed captions for each clip. Next, we follow the procedure in Section 3.1 to produce annotations for each video-caption pair. Finally, to ensure data quality, we manually select samples that meet the following criteria:

- Video is not an animated still image without meaningful subject motion.
- Video does not contain extensive text overlays.
- The retrieved subjects and objects cover all the main subjects and objects in the video.
- The background image, produced by an inpainting algorithm, has successfully removed foreground objects without generating new objects.

To increase data diversity, we select one clip from each long video and collect 2,130 clips. Figure 4 shows an annotated test sample.

Evaluation Metrics. An ideal personalized video output should align with the text, preserve subject fidelity, and exhibit natural video dynamics. Therefore, we use the following five metrics:

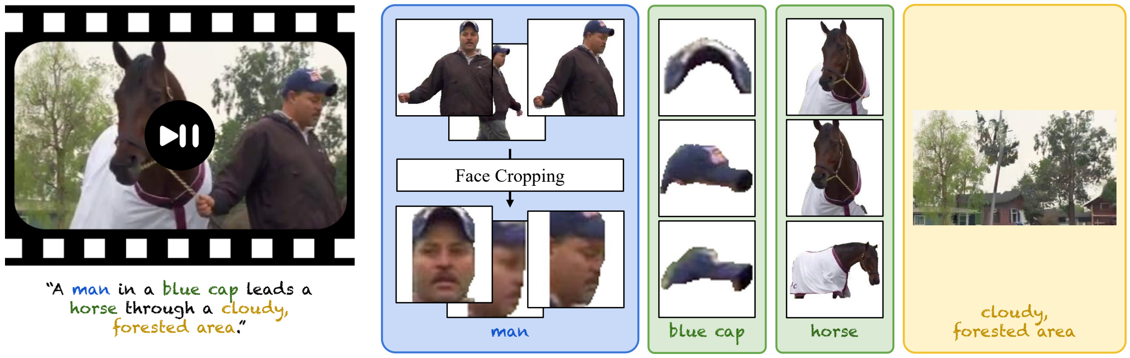

Figure 4. **Test sample in *MSRVTT-Personalization*.** We present a comprehensive video personalization benchmark. Our benchmark supports various modes, including face conditioning, single or multiple subjects conditioning, and foreground and background conditioning.

Table 1. **Quantitative comparison on *MSRVTT-Personalization*.** We compare *Video Alchemist* with state-of-the-art personalization methods across multiple metrics, including text similarity (Text-S), video similarity (Vid-S), subject similarity (Subj-S), face similarity (Face-S), and dynamic degree (Dync-D). The top and bottom tables show the evaluations for subject and face modes, respectively. †For text-to-image models, outputs are treated as single-frame videos without evaluating temporal quality. We evaluate *Video Alchemist* with the videos at 512px $\times$ 288px resolution. We highlight the top two models for the single reference image setting.

Method	Test-time Optimization	Reference Images		Text-S↑	Vid-S↑	Subj-S↑	Dync-D↑
		Subject	Background				
ELITE† [77]	✗	single	✗	0.245	0.620	0.359	-
VideoBooth [32]	✗	single	✗	0.222	0.612	0.395	0.448
DreamVideo [78]	✓	single	✗	0.261	0.611	0.310	0.311
Video Alchemist	✗	single	✗	0.269	0.732	0.617	0.466
DreamVideo [78]	✓	multiple	✗	0.253	0.604	0.256	0.303
Video Alchemist	✗	multiple	✗	0.268	0.743	0.626	0.473
Video Alchemist	✗	multiple	✓	0.254	0.780	0.570	0.506

Method	Test-time Optimization	Reference Images	Text-S↑	Vid-S↑	Face-S↑	Dync-D↑
		Face Crop				
IP-Adapter† [85]	✗	single	0.251	0.648	0.269	-
PhotoMaker† [38]	✗	single	0.278	0.569	0.189	-
Magic-Me [45]	✓	single	0.251	0.602	0.135	0.418
Video Alchemist	✗	single	0.273	0.687	0.382	0.424
PhotoMaker† [38]	✗	multiple	0.275	0.582	0.216	-
Magic-Me [45]	✓	multiple	0.248	0.618	0.153	0.385
Video Alchemist	✗	multiple	0.272	0.694	0.411	0.402

- Text similarity [79]: cosine similarity between the CLIP ViT-L/14 [55] features of the text and the generated frames. It measures how the generated video aligns with the text prompt.
- Video similarity [19]: average cosine similarity between the CLIP ViT-L/14 features of the ground truth and generated frames.
- Subject similarity: average cosine similarity between the DINO ViT-B/16 [7] features of the reference images and the segmented subject of the generated frames. We segment the subjects using Grounding-DINO Swin-T [40] and SAM ViT-B/16 [35].
- Face similarity: average cosine similarity between the ArcFace R100 [16] features of the reference face crops and the generated face crops. We detect generated faces using YOLOv9-C [74].

Table 2. **User preference study.** We show the user preference percentage for subject (left) and face modes (right), respectively.

Method	Preference Ratio↑		Method	Preference Ratio↑	
	Quality	Fidelity		Quality	Fidelity
ELITE [77]	2.7%	0.6%	IP-Adapter [85]	10.4%	20.2%
VideoBooth [32]	0.3%	0.8%	PhotoMaker [38]	37.5%	7.4%
DreamVideo [78]	0.5%	0.5%	Magic-Me [45]	4.4%	4.0%
Video Alchemist	96.5%	98.1%	*Video Alchemist*	47.6%	68.4%

- Dynamic degree [30]: optical flow magnitude between consecutive generated frames. We compute the optical flow using RAFT [67].

Note that video frames with missing subjects or faces are assigned a similarity score of 0. The benchmark is publicly available at https://github.com/snap-research/MSRVTT-Personalization.

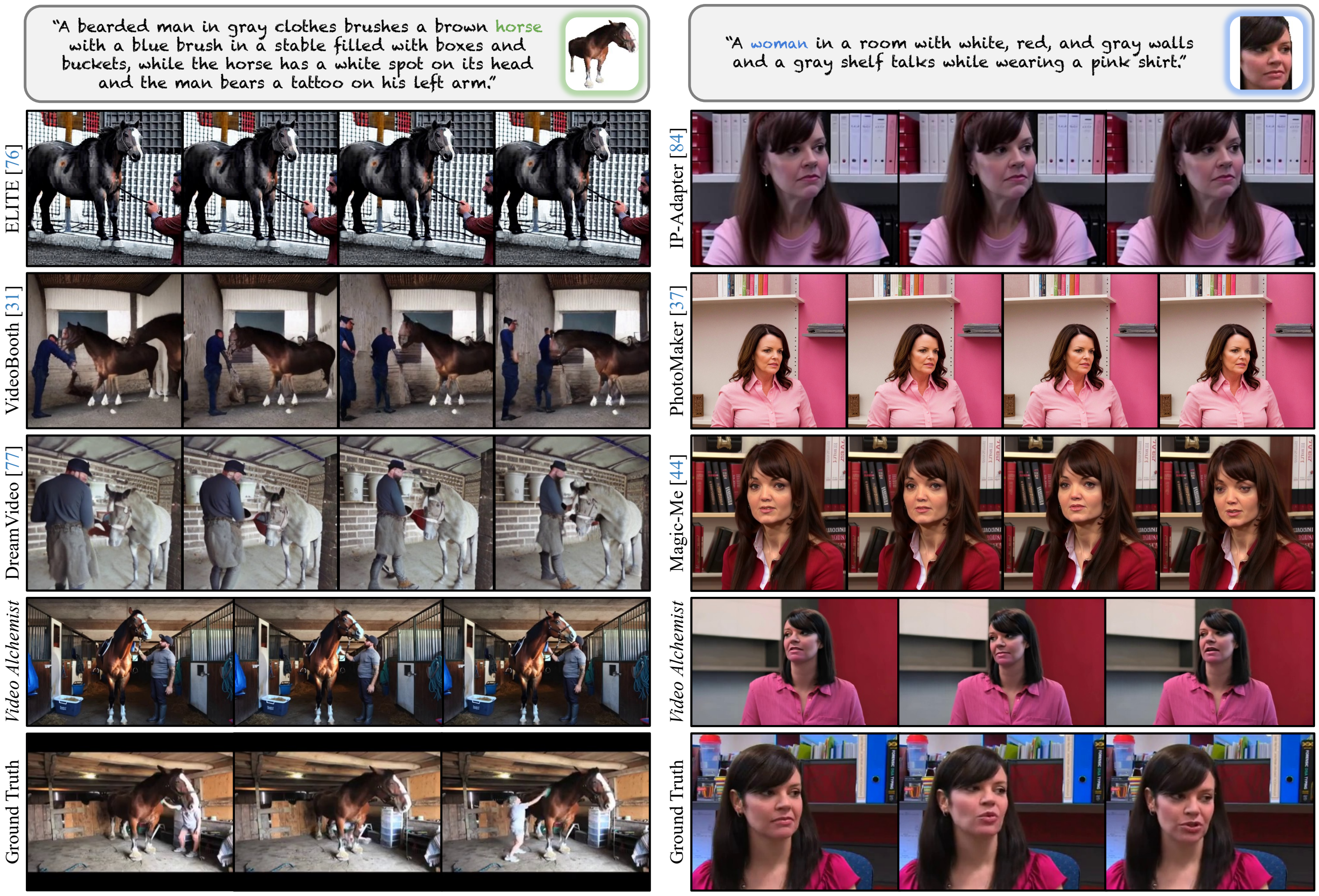

Figure 5. **Qualitative comparison on *MSRVTT-Personalization*.** We use a single reference image to each model for a fair comparison. Compared to existing methods, our results closely match the input text prompt and reference subjects while exhibiting natural motion and pose variations.

4.2. Comparisons with the State-of-the-Arts

In this section, we quantitatively and qualitatively compare *Video Alchemist* with existing personalization models on *MSRVTT-Personalization*.

Experimental Setups. We extensively compare various personalization models, including text-to-image [38, 77, 85] and text-to-video models [32, 45, 78], as well as optimization-based [45, 78] and encoder-based methods [32, 38, 77, 85]. As existing methods use different types of conditional images, we introduce two evaluation modes: subject mode and face mode. Subject mode uses full subject images as input, while face mode uses only face crops. For subject mode, we collect 1,736 test videos with a single subject. For face mode, we collect 1,285 test videos with a single person's face crop.

For text-to-image models [38, 77, 85], we treat the output images as single-frame videos. For optimization-based models [45, 78], we use the default hyperparameters in the official codebase for finetuning. For IP-adapter [85], we use the checkpoint of IP-Adapter-FaceID+. If the model supports multiple reference images, we evaluate it with both single and multiple input images. We additionally evaluate our model with an extra input of a background reference image in the subject mode.

Quantitative Evaluation on *MSRVTT-Personalization*. Table 1 shows the quantitative evaluation results. Compared to the existing open-set personalization methods [32, 77, 78], *Video Alchemist* achieves higher subject fidelity, with a 23.2% higher subject similarity than VideoBooth [32]. Meanwhile, our model achieves the best text alignment and greatest video dynamics. Notably, our open-set model outperforms face-specific models [38, 45, 85] in face fidelity, achieving 11.3% higher face similarity than IP-adapter [85].

Moreover, *Video Alchemist* can generate the target subject or face with higher fidelity when provided with more reference images, demonstrating the advantage of multi-image conditioning. Furthermore, leveraging an extra background reference image, *Video Alchemist* can synthesize a video more similar to the ground truth video, highlighting the effectiveness of our background conditioning. However, more reference images sometimes lead to worse textual alignment, potentially due to the limited flexibility introduced by more reference images.

Table 3. **Ablation study for the subject mode.** We use a single reference image for each model and examine three control factors. The experiments are conducted on the videos at 256px × 144px resolution.

Method	Image Encoder	Use Word Token	Image Augmentations	Text-S↑	Vid-S↑	Subj-S↑	Dync-D↑
Use CLIP	CLIP [55]	✓	✓	0.269	0.768	0.569	0.552
No word token	DINOv2 [50]	✗	✓	0.256	0.790	0.566	0.569
No augmentation	DINOv2 [50]	✓	✗	0.251	0.781	0.609	0.506
Video Alchemist	DINOv2 [50]	✓	✓	0.257	0.790	0.600	0.570

Qualitative Evaluation on *MSRVTT-Personalization*. In Figure 5, we show videos generated by different methods alongside the ground truth videos. More comparisons on various conditional subjects can be found in Appendix C.3. Compared to existing models, our method produces more photorealistic videos with higher fidelity for target subjects.

Human Evaluation. To complement automated evaluation, we conduct a user study to assess visual quality and subject fidelity. We randomly select 200 test samples from the subject and face modes, respectively, and show the conditional image and the results to 5 participants. For each sample, participants are asked to select the one that best preserves the subject details and has the best visual quality.

Table 2 summarizes the results. Our method significantly outperforms the state-of-the-art methods in both visual quality and subject fidelity. Notably, the fidelity scores reported by humans are positively correlated to the scores of subject similarity and face similarity in Table 1, showcasing the effectiveness of the proposed *MSRVTT-Personalization*.

4.3. Ablation Study

In this section, we present an ablation study with three training or architecture choices. The quantitative and qualitative evaluations are shown in Table 3 and Figure 6, respectively.

Different Image Encoders. Inspired by Wu *et al*. [81], we train the models with two image encoders, CLIP [55] and DINOv2 [50], and find that CLIP achieves better text similarity, while DINOv2 performs better in subject similarity. We hypothesize that DINOv2, trained with self-supervised learning objectives, captures unique object features. In contrast, CLIP, designed to bridge visual and textual modalities, focuses on details typically described in the prompt, which can improve the text-image alignment.

Necessity of Binding Image and Word Concepts. In Section 3.2, we propose a mechanism to bind the concepts of images and the corresponding entity words. Without such binding, the model may incorrectly apply image conditions to the wrong subject. For example, the model places a reference human face on a dog as in the second row of Figure 6. This misalignment also results in missing subjects and lower subject similarity.

Effect of Data Augmentation. In Section 3.3, we introduce data augmentation to reduce model learning. Without augmentation, the model suffers from the *copy-and-paste* issue. Although this helps to achieve higher subject similarity, it degrades dynamic degree and decreases text similarity. Specifically, although the prompt in Figure 6 is `a woman is smiling ...`, the synthetic subject in the third row does not *smile*. Instead, it replicates the same facial expression as in the reference image.

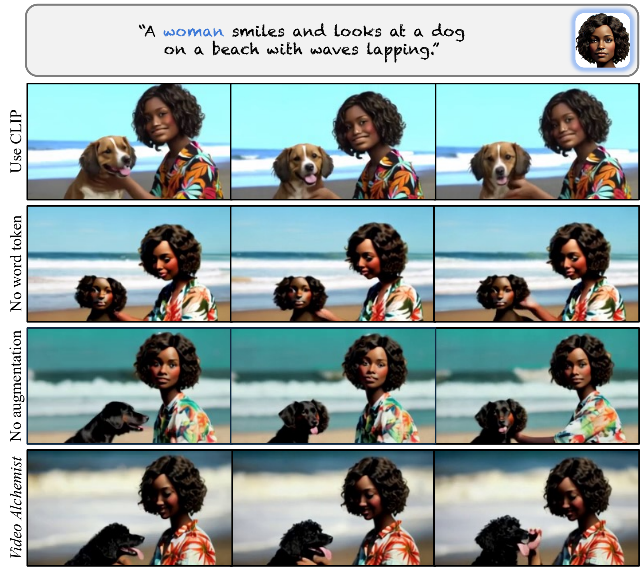

Figure 6. **Qualitative results of the ablation study.** From top to bottom, we show that 1) *Video Alchemist* achieves better subject fidelity using DINOv2 [50] as the image encoder; 2) it correctly binds the conditional image and entity word with the usage of word tokens; 3) it mitigates the *copy-and-paste* effect and synthesizes text-aligned videos via the proposed data augmentation. The reference image is synthesized by DALL·E 3 [3].

5. Conclusion

We have presented *Video Alchemist*, a video personalization model that supports multi-subject and open-set personalization capabilities for both foreground objects and background without requiring test-time optimization. It is built on a Diffusion Transformer module that integrates conditional images with their subject-level prompts through cross-attention layers. With our dataset curation and data augmentation, we have reduced model overfitting on undesirable properties of the reference images. In addition, we have introduced a new benchmark for evaluating personalization models across various conditioning scenarios. Experimental results show that our method outperforms existing methods in quantitative and qualitative measures.

Acknowledgments. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

We thank Ziyi Wu, Moayed Haji Ali, and Alper Canberk for their helpful discussions.

References

[1] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia*, 2023. 3

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, 2023. 3

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023. 8

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 16

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. Technical report, OpenAI, 2024. 2, 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

[8] Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*, 2023. 2, 3

[9] Hong Chen, Xin Wang, Yipeng Zhang, Yuwei Zhou, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Disenstudio: Customized multi-subject text-to-video generation with disentangled spatial control. In *ACM MM*, 2024. 2, 3

[10] Ting Chen and Lala Li. Fit: Far-reaching interleaved transformers. *arXiv preprint arXiv:2305.12689*, 2023. 3

[11] Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*, 2023. 4

[12] Tsai-Shien Chen, Chieh Hubert Lin, Hung-Yu Tseng, Tsung-Yi Lin, and Ming-Hsuan Yang. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*, 2023. 3

[13] 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. 13

[14] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 3

[15] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*, 2022. 4

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 5, 6

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 16

[18] Yuwei Fang, Willi Menapace, Aliaksandr Siarohin, Tsai-Shien Chen, Kuan-Chien Wang, Ivan Skorokhodov, Graham Neubig, and Sergey Tulyakov. Vimi: Grounding video generation through multi-modal instruction. In *EMNLP*, 2024. 3

[19] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3, 6

[20] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 2023. 3

[21] Peng Gao, Le Zhuo, Ziyi Lin, Chris Liu, Junsong Chen, Ruoyi Du, Enze Xie, Xu Luo, Longtian Qiu, Yuhang Zhang, et al. Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945*, 2024. 16

[22] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 2

[23] 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

[24] Cusuh Ham, Matthew Fisher, James Hays, Nicholas Kolkin, Yuchen Liu, Richard Zhang, and Tobias Hinz. Personalized residuals for concept-driven text-to-image generation. In *CVPR*, 2024. 3

[25] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *ICCV*, 2023. 3

[26] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 2, 3

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 15, 16

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 2

[30] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 6

[31] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3, 13

[32] 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 *CVPR*, 2024. 2, 3, 6, 7

[33] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia*, 2024. 3

[34] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 4, 6

[36] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[37] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2023. 3

[38] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 6, 7

[39] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 4

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4, 6

[41] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 4, 16

[42] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 3

[43] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 15

[44] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2

[45] Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *arXiv preprint arXiv:2402.09368*, 2024. 2, 3, 6, 7

[46] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. Text-guided synthesis of eulerian cinemagraphs. In *SIGGRAPH Asia*, 2023. 2

[47] Joanna Materzynska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. In *ACCV*, 2024. 3

[48] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*, 2024. 2, 3

[49] Nvidia. Fused layer norm, 2018. 4

[50] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 4, 5, 8, 15

[51] Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024. 3

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 16

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3, 4, 15

[54] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 15

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6, 8, 15

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4
[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, 2015. 3
[58] 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 *CVPR*, 2023. 2, 3, 5
[59] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 3
[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2
[61] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 3
[62] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 2
[63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2
[64] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2
[65] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 5, 13
[66] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4
[67] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 6
[68] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM TOG*, 2024. 3
[69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 13
[70] Torchmetrics. Clip score - pytorch-metrics, 2024. 25
[71] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia*, 2023. 3
[72] A Vaswani. Attention is all you need. In *NeurIPS*, 2017. 4
[73] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 3
[74] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv:2402.13616*, 2024. 4, 6
[75] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 3
[76] Zhao Wang, Aoxue Li, Enze Xie, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 2, 3
[77] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, 2023. 3, 6, 7
[78] 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, 3, 5, 6, 7
[79] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 6
[80] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 2, 3
[81] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey R Allen, and Thomas Kipf. Neural assets: 3d-aware multi-object scene synthesis with image diffusion models. In *NeurIPS*, 2024. 8
[82] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *IJCV*, 2024. 3
[83] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5
[84] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4, 15
[85] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 5, 6, 7, 25
[86] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024. 3
[87] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video

generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 2

[88] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 2, 3, 5