CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities

Tao Wu [1], Yong Zhang [2*], Xintao Wang [2,4], Xianpan Zhou [3], Guangcong Zheng [1], Zhongang Qi [4], Ying Shan[2,4], Xi Li [1*]

[1]College of Computer Science and Technology, Zhejiang University
[2]Tencent AI Lab
[3]Polytechnic Institute, Zhejiang University
[4]ARC Lab, Tencent PCG

taowucs@zju.edu.cn, {norriszhang,xintaowang}@tencent.com, {zhouxianpan,guangcongzheng}@zju.edu.cn, {zhongangqi, yingsshan}@tencent.com, xilizju@zju.edu.cn

Abstract

Customized video generation aims to generate high-quality videos guided by text prompts and subject's reference images. However, since it is only trained on static images, the fine-tuning process of subject learning disrupts abilities of video diffusion models (VDMs) to combine concepts and generate motions. To restore these abilities, some methods use additional video similar to the prompt to fine-tune or guide the model. This requires frequent changes of guiding videos and even re-tuning of the model when generating different motions, which is very inconvenient for users. In this paper, we propose CustomCrafter, a novel framework that preserves the model's motion generation and conceptual combination abilities without additional video and fine-tuning to recovery. For preserving conceptual combination ability, we design a plug-and-play module to update few parameters in VDMs, enhancing the model's ability to capture the appearance details and the ability of concept combinations for new subjects. For motion generation, we observed that VDMs tend to restore the motion of video in the early stage of denoising, while focusing on the recovery of subject details in the later stage. Therefore, we propose Dynamic Weighted Video Sampling Strategy. Using the pluggability of our subject learning modules, we reduce the impact of this module on motion generation in the early stage of denoising, preserving the ability to generate motion of VDMs. In the later stage of denoising, we restore this module to repair the appearance details of the specified subject, thereby ensuring the fidelity of the subject's appearance. Experimental results show that our method has a significant improvement compared to previous methods.

Code — https://github.com/WuTao-CS/CustomCrafter

Introduction

With the development of diffusion models and multimodal, text-to-video generation has made significant progress (Brooks et al. 2024; Yang et al. 2023, 2024a,b; Miao et al. 2023; Su et al. 2023a,b). Challenges still arise when users want to generate videos of specific subjects. Customized video generation needs to simultaneously satisfy three requirements: consistent subject appearance, free concept combination, and smooth motion generation. Concept combination refers to the ability to combine the learned specific subject with other different concepts. For example, as shown in Figure 1, when we learn about a specific guitar, we hope that this guitar can be combined with other concepts (e.g., a person) to generate videos. Numerous studies (Gal et al. 2023; Ruiz et al. 2023; Han et al. 2023; Gu et al. 2024b; Kumari et al. 2023; Ruiz et al. 2024; Huang et al. 2024b,a) have proposed many methods for customized image generation and have achieved good results. When these approaches are applied directly to customized video generation, they often fail to generate videos well. These methods damage the conceptual combination ability and motion generation ability of the text-to-video model during the fine-tuning process, which means that the subject learned by the model cannot be combined with other concepts and the motion in the generated video tends to be static.

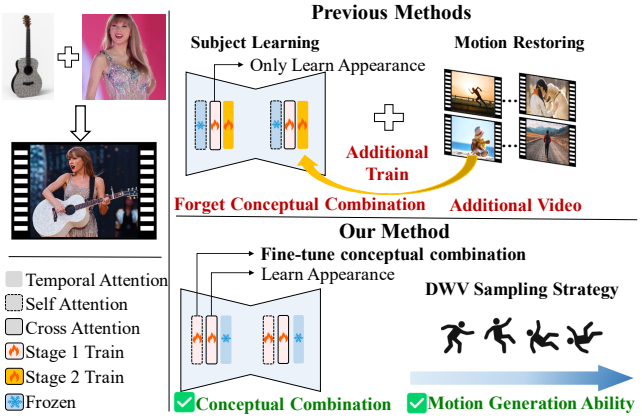

Figure 1: Comparison of our approach with previous work. Our method can better learn the appearance of the subject while preserving the concept combination ability and motion generation ability, only requires one stage of training without additional videos.

Recent methods for customized video generation have noticed these issues. Some works (Wei et al. 2024; He

*Corresponding author.

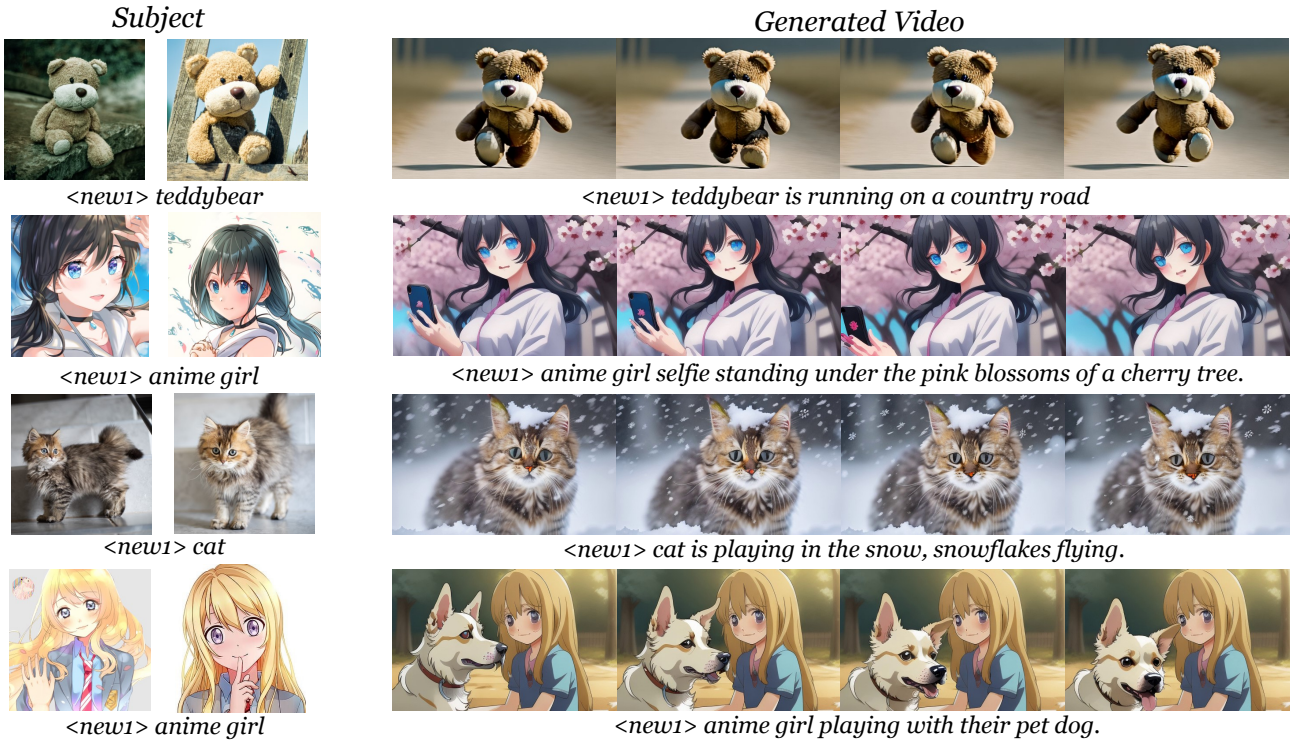

Figure 2: Visualization for our CustomCrafter. Our approach allows customization of subject identity and movement patterns to generate the desired video with text prompt by preserving motion generation and conceptual combination abilities.

et al. 2023) have recognized the decline in motion fluidity and conceptual combination abilities. These methods believe that the damage to the model's concept combination ability and motion generation ability cannot be recovered, which is caused by using images to fine-tune the VDMs during subject learning. So, as shown in Figure 1, they use further fine-tuning of the model parameters with additional videos similar to the content described in the prompt (Wei et al. 2024), or use videos to guide the video generation process (He et al. 2023). However, these methods require retrieving similar text prompts from massive video libraries to generate different prompts for the same subject. It is often necessary to frequently change the guiding videos or even re-fine-tune the model, leading to additional training, which brings great inconvenience to users. Considering the above issues, one question naturally arises: **Is it possible to generate videos of a specified subject only by performing subject learning, while preserving the model's inherent abilities of concept combination and motion generation?**

In this paper, we introduce a novel framework, CustomCrafter, which preserves the model's motion generation and conceptual combination abilities without the need for additional video and fine-tuning to recover these abilities. Through experiments, we have observed that: (1) Numerous studies (Cao et al. 2023; Gu et al. 2024a; Nam et al. 2024) have indicated that for image models, self-attention often significantly affects the ability to combine concepts. And self-attention plays a crucial role in preserving geometric and shape for subject (Liu et al. 2024). We have observed that this phenomenon remains applicable to VDMs. Furthermore, influenced by (Kumari et al. 2023), existing work only updates spatial cross-attention during subject learning. (2) During the video diffusion models (VDMs) generation process, we can observe that in different timesteps of denoising process, the model repair content has a certain tendency. The early stages of the denoising process tend to restore the layout of each frame and the motion, whereas the later stages focus on the recovery of object detail.

Therefore, to address the issue of decreased concept combination ability, we propose the Spatial Subject Learning Module, which can update the weights of both the spatial cross-attention and self-attention layers during fine-tuning. This improves the model's ability to capture the appearance of new subjects, while also improving the model's ability to combine new subjects with other concepts. Regarding the decline in motion generation ability, we design the Spatial Subject Learning Module to be pluggable, and propose the Dynamic Weighted Video Sampling Strategy, which improves the model's inference process. When generating videos, by leveraging the pluggable nature of the subject learning module, we can preserve the model's motion generation ability by reducing the influence of subject learning on the stages that tend to restore motion. Then, during the stages that tend to repair details, we can restore the influence of the subject learning module. This ensures the consistency of the subject's appearance, thereby enabling the generation of videos of a specified subject using the model's inherent motion generation ability. As shown in Figure 2,

we can generate high-quality videos of a specified subject by preserving the inherent abilities of VDMs of concept combination and motion generation and only performing subject learning. Our method does not require the introduction of additional videos as guidance or repeated fine-tuning of the model. By preserving the inherent knowledge of the text-to-video model, we can conveniently generate videos of specified objects that align with the prompt.

We provide qualitative, quantitative, and user study results that demonstrate the superiority of our method. Our contributions are summarized as follows:

- As far as we know, we are the first to discover and use the property of VDMs' denoising process to decouple appearance and motion to improve customized generation.
- We propose a subject learning method that can learn the appearance of the subject better and effectively preserve the ability to combine new subjects with other concepts.
- We introduce a sampling strategy that can preserve the motion generation of VDMs without using additional videos to guide or fine-tune the model.

Related Work

Text-to-Video Diffusion Models

Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Jiang et al. 2024) have recently emerged as a trend in generative models, particularly in the domain of text-to-image (T2I) generation (Ramesh et al. 2022; Rombach et al. 2022; Zhang et al. 2024; Wu et al. 2024a; Wang et al. 2024a; Wu et al. 2024b; Huang et al. 2024a; Zheng et al. 2023). For video generation, Video Diffusion Model has been introduced to model video distributions. Pioneering work to utilize a space-time-factored U-Net for video modeling in pixel space for unconditional video generation was done by VDM (Ho et al. 2022; Dou et al. 2024). AnimateDiff (Guo et al. 2024) further advanced the field of text-to-video generation by incorporating a motion module into the Stable Diffusion model. Following this, LVDM (He et al. 2022; Chen et al. 2023) suggest extending LDM to model videos in the latent space of an autoencoder. This method gradually became mainstream and derived many methods, including ModelScope (Wang et al. 2023a), LAVIE (Wang et al. 2023c), PYOCO (Ge et al. 2023), VideoFactory (Wang et al. 2023b), VPDM (Yu et al. 2023), and VideoCrafter2 (Chen et al. 2024a).

Customized Generation on Diffusion Models

Numerous studies (Wei et al. 2023; Li, Li, and Hoi 2024) have proposed many methods for custom image generation and achieved good results. Most current work focuses on subject customization with a few images (Han et al. 2023; Gu et al. 2024b; Shi et al. 2024). Moreover, some works study the more challenging multi-subject customization task (Yuan et al. 2024; Xiao et al. 2024; Ma et al. 2024; Chen et al. 2024b). Despite significant progress in customized image generation, customized video generation is still under exploration. Although there have been initial attempts to customize video diffusion, such as VideoAssembler (Zhao et al. 2023), VideoBooth (Jiang et al. 2023),ID-Animator (He et al. 2024) and CustomVideo (Wang et al. 2024b), which use reference images to personalize the video diffusion model while preserving the identity of the subject. These approaches focus on addressing the problem of generating videos with a similar subject appearance, neglecting disruption to motion and conceptual combination abilities. DreamVideo (Wei et al. 2024) first decouples the learning process for subject and motion. Animate-A-Story (He et al. 2023) refers to the depth information of the additional video to guide motion generation. However, the use of additional video data and the need to retrain the model according to different prompts bring great inconvenience to users.

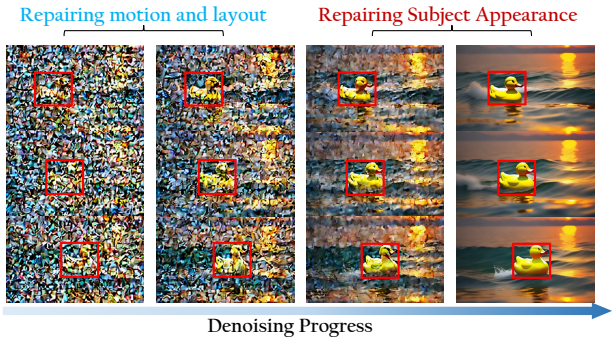

Figure 3: Visualization of video denoising process. The motion is formed in early stages of the denoising process, and the subject's appearance emerges in later stages.

Preliminary

Video diffusion models (VDMs) (Wang et al. 2023a; He et al. 2022; Guo et al. 2024; Chen et al. 2024a)2 are designed for video generation tasks by extending image diffusion models to adapt to video data. VDMs learn a video data distribution by the gradual denoising of a variable sampled from a Gaussian distribution. First, a learnable autoencoder (consisting of an encoder $\mathbf{E}$ and a decoder $\mathbf{D}$) is trained to compress the video into a smaller latent space representation. Then, a latent representation $z = \mathbf{E}(x)$ is trained instead of a video x. Specifically, the diffusion model ϵ_θ aims to predict the added noise ϵ at each timestep t based on the text condition c, where $t \in \mathcal{U}(0, 1)$. The training objective can be simplified as a reconstruction loss:

$$\mathcal{L}_{video} = \mathbb{E}_{z,c,\epsilon\sim\mathcal{N}(0,\mathrm{I}),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c, t\right)\right\|_2^2\right], \quad (1)$$

where $z \in \mathbb{R}^{F\times H\times W\times C}$ is the latent code of video data with F, H, W, C being frame, height, width, and channel, respectively. τ_θ presents a pre-trained text encoder. c is the text prompt for input video. A noise-corrupted latent code z_t from the ground-truth z_0 is formulated as $z_t = \lambda_t z_0 + \sigma_t \epsilon$, where $\sigma_t = \sqrt{1-\lambda_t^2}$, λ_t and σ_t are hyperparameters to control the diffusion process. In this work, we have selected the VideoCrafter2 (Chen et al. 2024a) as our base model.

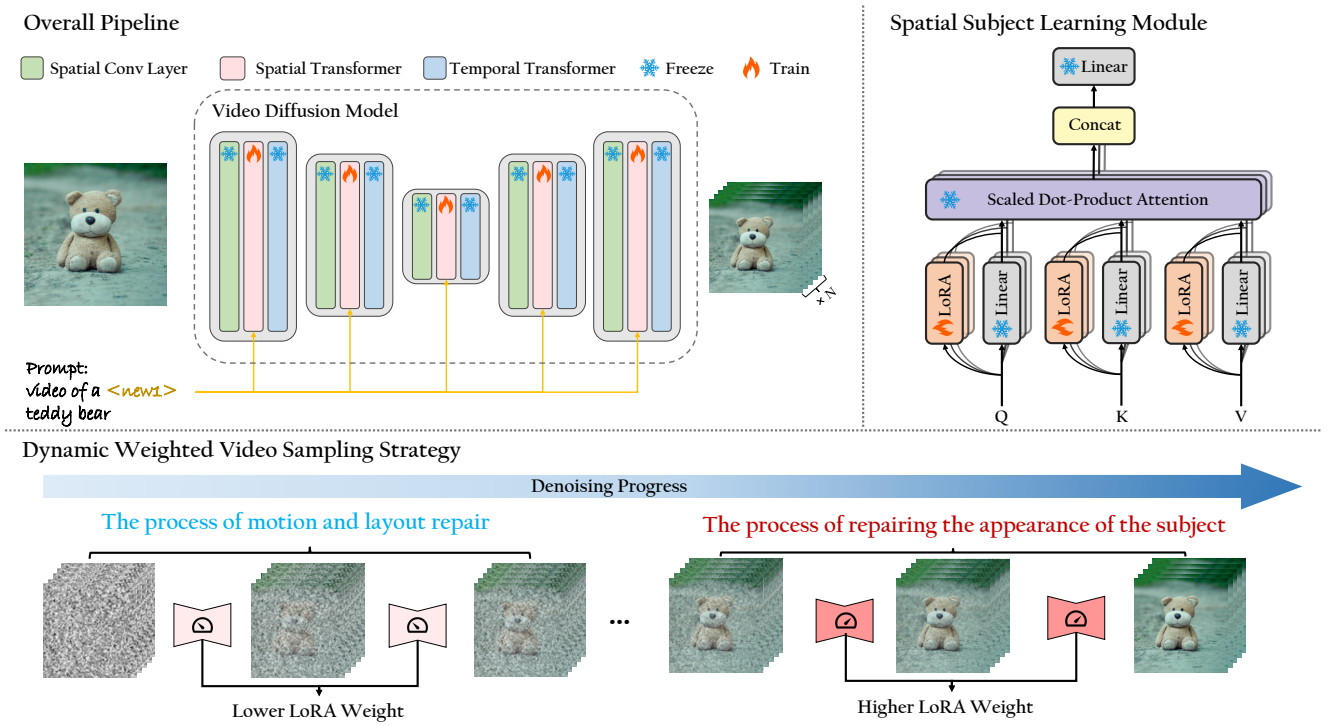

Figure 4: Overall review of CustomCrafter. For subject learning, we adopt LoRA to construct Spatial Subject Learning Module, which update the Query, Key, and Value parameters of attention layers in all Spatial Transformer models. In the process of generating videos, we divide the denoising process into two phases: the motion layout repair process and the subject appearance repair process. By reducing the influence of the Spatial Subject Learning Module in the motion layout repair process, and restoring it in the subject appearance repair process to repair the details of the subject.

Method

In this section, we first analyze the reasons for the impairment of conceptual combination and motion capabilities in models during the subject learning and outline our findings about the VDMs. Then, we introduce the Spatial Subject Learning Module, which used to learn the appearance details of the subject. Finally, we introduce our Dynamic Weighted Video Sampling Strategy, which can generate high-quality videos without additional video guidance or training after the model has learned the appearance of the subject.

Explore the Video Diffusion Model

Given a pre-trained VDM, our goal is to enable the model to learn the subject's appearance from a number of images and the corresponding text prompt. To achieve this goal, most existing works update part of the model to learn subject's appearance. However, since the training data only consist of images and the VDMs require a relatively long training period to learn a new concept, the model inevitably forgets its motion generation and concept combination ability. Regarding the issue of conceptual combination, many studies on diffusion models point out (Cao et al. 2023; Nam et al. 2024) that for text-to-image diffusion models, self-attention significantly affects the geometric and shape for subject and the ability to combine concepts. We found the same phenomenon in video diffusion models. However, current methods tend to fine-tune only the parameters of cross-attention during training. This results in the model being unable to learn the appearance of new subjects and reduces its ability to combine new subjects with other concepts.

For motion generation, as shown in Figure 3, we observed that the duck's movement is formed early in the denoising process. The later stage of the denoising process is to enhance the appearance of the subject based on the established motion. Therefore, during the video generation process, the models repair content with a certain tendency. The VDMs tend to restore the overall layout and motion in the early stages of the denoising process, while focusing on the recovery of object detail in the later stages. Therefore, our approach is to utilize a plug-and-play module to facilitate subject learning. By reducing the influence of this module on the denoising generation process in the early stages of inference, we can mitigate the disruption to the motion generation capability of VDMs. In the later stages of the denoising process, where object detail recovery occurs, we increase the influence of this module. This approach allows us to preserve the original model's video motion generation ability while generating high-quality appearance of new subjects.

Spatial Subject Learning Module

For learning the appearance of new subject, we constructed a Spatial Subject Learning Module. During training, we re-

peat a single picture of the object N times to turn the picture into a still N frame video. As shown in Figure 4, our fine-tuning parameters can be divided into two parts. First, following textual inversion (Gal et al. 2023), we employ a new token V^* and learn a new token embedding vector in the CLIP text encoder to represent a new concept. For example, a specified teddy bear can be represented as $V^*teddybear$. The second part pertains to the spatial transformer module of the video diffusion model. We fine-tuned both the cross-attention module and the self-attention module in the spatial transformer blocks of VDM to ensure that the model has the ability to combine new subjects with other concepts while learning the appearance of the subject. To achieve a plug-and-play effect, we adopted the Low-Rank Adaptation (Hu et al. 2021) (LoRA) method for fine-tuning. LoRA applies a residue path of two low-rank matrices $B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times k}$ in the attention layers, whose original weight is $W_0 \in \mathbb{R}^{d\times k}$, $r \ll \min(d, k)$. The new forward path is as follows:

$$W = W_0 + \lambda \Delta W = W_0 + \lambda BA, \tag{2}$$

where λ is a coefficient adjusting the strength of the added LoRA. In this paper, we insert LoRA layers into the query, key, and value corresponding to W_q,W_k,W_v in both the cross-attention and self-attention modules for fine-tuning.

Dynamic Weighted Video Sampling Strategy

To alleviate the decrease in motion generation ability, as shown in Algorithm 1, we propose Dynamic Weighted Video Sampling Strategy. We notice that during the generation process of the video diffusion model, in the early stage of denoising, the model tends to restore the video's motions. In the later stage of the denoising process, the model tends to restore the details of the generated video. Therefore, in the first K steps of the denoising process, we adjust the weight λ of all LoRA modules in our Spatial Subject Learning Module to a smaller value λ_s, to ensure that the model is almost unaffected by motion stagnation and the decrease in the combination ability of concepts caused by the parameters of the subject's overfit. In the later stage of the denoising process, we restore the weight λ of the LoRA modules to a higher value λ_l, allowing the model to further repair the specific details of each frame of the subject, thereby generating high-quality videos of the specified subject.

Model Training Strategy

Inspired by previous work (Ruiz et al. 2023; Kumari et al. 2023), during training, we use class-specific prior preservation to mitigate overfitting issues in the training process, to enhance the diversity of the generated videos. The loss for prior preservation is formulated as the following:

$$\mathcal{L}_{pr} = \mathbb{E}_{z^{pr},c^{pr},\epsilon\sim\mathcal{N}(0,\mathrm{I}),t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t^{pr}, c^{pr}, t\right)\right\|_2^2\right], \tag{3}$$

where z^{pr} is the latent code of the input regularized video, c^{pr} is the text condition for the input regularized video. In training, our total loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{video} + \alpha\mathcal{L}_{pr}, \tag{4}$$

where α is a hyper-parameter to adjust the relative weight of prior-preservation.

Algorithm 1: Dynamic Weighted Video Sampling Strategy

Input: A source prompt $\mathcal{P}$, a random seed s, a small LoRA weight λ_s used in Phase 1, a large LoRA weight λ_l used in Phase 2, and the delimitation point k.
Output: latent code for generating video.
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed s;
$Change(DM, \lambda, \lambda_s)$; /* Change λ to λ_s */
for $t = T, T-1, \ldots, 1$ **do**
  if *t == (T - K)* **then**
    $Change(DM, \lambda, \lambda_l)$ /* Change λ to λ_l */
  end
  $z_{t-1} \leftarrow DM(z_t, \mathcal{P}, t, s)$
end
Return: z_0

Method	CLIP-T↑	CLIP-I ↑	DINO-I↑	T. Cons. ↑
Custom Diffusion	0.289	0.759	0.546	0.990
Custom Diffusion*	0.286	0.769	0.583	0.992
DreamVideo	0.298	0.724	0.489	0.992
DreamVideo*	0.295	0.748	0.536	0.993
Ours	**0.318**	**0.786**	**0.627**	**0.994**

Table 1: Comparison with the existing methods. Note that Custom Diffusion* and DreamVideo* in the table represent the results we get after extending the number of training steps in the original paper.

Experiments

Experimental Setup

Datasets and Protocols For subject customization, we select subjects from image customization papers (Ruiz et al. 2023; Kumari et al. 2023) for a total of 20 subjects. For each subject, we use ChatGPT to generate 10 related prompts, which are used to test the generation of specified motion videos for the subject. All experiments use VideoCrafter2 as the base model. When learning the subject, we use the AdamW optimizer, set the learning rate to 3×10^{-5} and the weight decay to 1×10^{-2}. We perform 10,000 iterations on 4 NVIDIA A100 GPUs. For the Class-specific Prior Preservation Loss, similar to (Ruiz et al. 2023; Kumari et al. 2023),we collected 200 images from LAION-400M (Schuhmann et al. 2021) for each subject as regularization data and set α to 1.0. During the inference process, we use DDIM (Song, Meng, and Ermon 2020) for 50-step sampling and classifier-free guidance with a cfg of 12.0 to generate videos with a resolution of 512×320. For all subjects, to facilitate experimentation and comparison, we uniformly set λ_s and λ_l to 0.4 and 0.8 respectively, and set K to 5 based on our observation. In actual use, these parameters can be adjusted by the user.

Baselines Given that different base model are chosen in the current field of video customization, we reproduce Custom Diffusion (Kumari et al. 2023) and DreamVideo (Wei et al. 2024) based on VideoCrafter2. Since our methods do not introduce additional videos as guidance, to ensure fairness, we only reproduce the subject learning part of

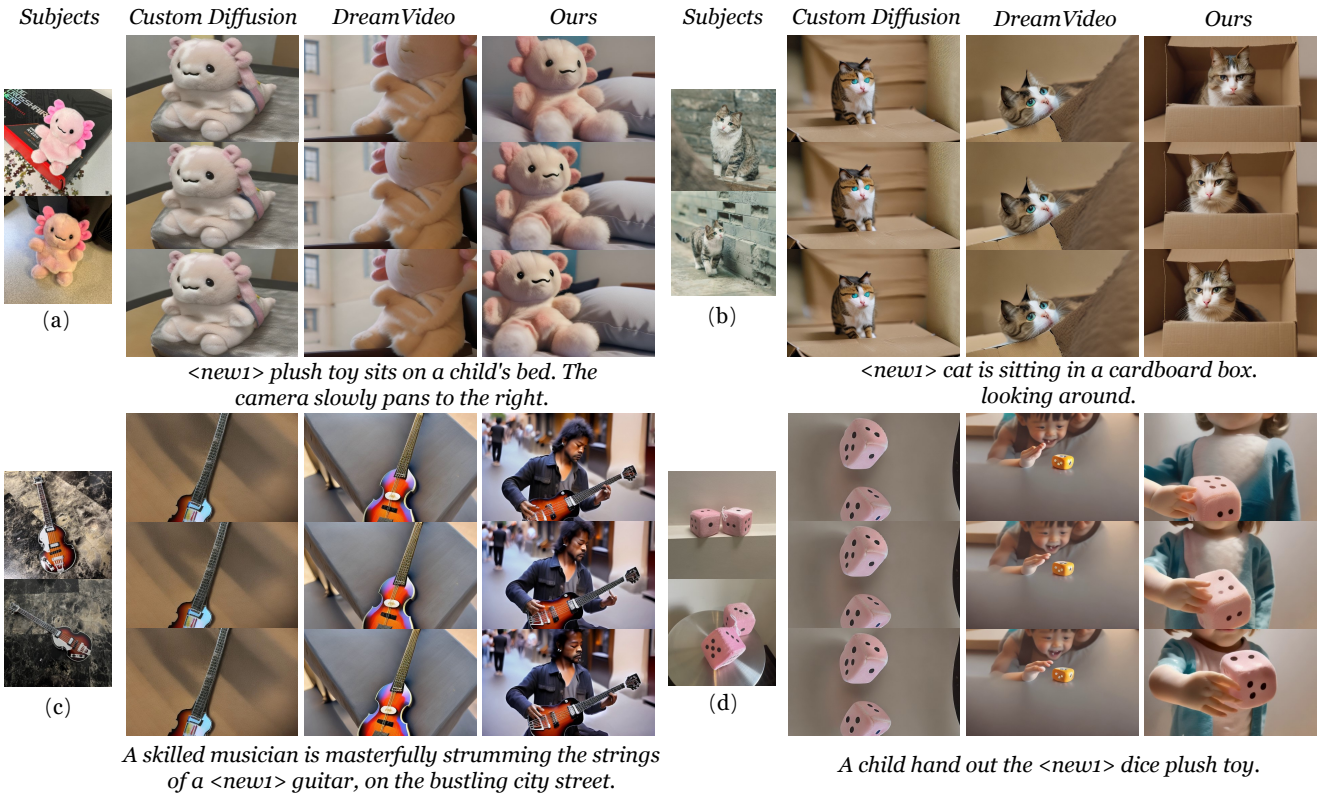

Figure 5: Qualitative comparison of customized video generation with both subjects and motions. Without guidance from additional videos, our method significantly outperforms in terms of concept combination.

DreamVideo for fair comparison. In addition, considering that VDMs need more steps to learn the appearance of the subject, and the default settings of Custom Diffusion and DreamVideo cannot fit the subject appearance features well, we accordingly extend the training steps of these methods.

Evaluation Metrics Follow (Wei et al. 2024; Wang et al. 2024b), we evaluate our approach with the following four metrics: (1) *CLIP-T* calculates the average cosine similarity between CLIP (Radford et al. 2021) image embeddings of all generated frames and their text embedding. (2) *CLIP-I* measures the visual similarity between the generated and target subjects. We computed the average cosine similarity between the CLIP image embeddings of all generated frames and the target images. (3) *DINO-I* (Ruiz et al. 2023), another metric to measure visual similarity using ViTS/16 DINO (Zhang et al. 2022). Compared to CLIP, the self-supervised training model encourages the distinguishing features of individual subjects. (4) *Temporal Consistency* (Esser et al. 2023), we compute CLIP image embeddings on all generated frames and report the average cosine similarity between all pairs of consecutive frames.

Quantitative Results

We trained 20 subjects using Custom Diffusion, DreamVideo and our method, respectively. After training, we used each method to generate videos for each subject using 10 prompts, employing the same random seed and denoising steps. The results, as shown in Table 1, indicate that our method outperforms existing methods in all four metrics. The degree of text alignment and subject fidelity have been significantly improved. The temporal consistency of the generated videos is roughly equivalent to that of other methods. The metrics used to evaluate the subject fidelity, CLIP-I and DINO-I, have improved by 1.7% and 4.4%, respectively, compared to existing methods. The degree of text alignment has improved by 1.5% compared to the previous best result.

Qualitative Results

We also visualized some results for qualitative analysis. We used the prompt of dynamic videos to generate videos of specified subjects, observing the subject fidelity in the generated videos and the motion fluency. As shown in Figure 5(a), when we want to generate a video of a specified plush toy sitting on a child's bed and the camera slowly pans to the right, we find that existing methods overfit reference image during training. Without guidance from additional videos, the generated motions are almost static. However, our method can generate videos with fluent motions and right concept combination. Besides, in Figure 5(b), only our method correctly generates the conceptual combination of the cat and the cardboard box and the motion of "looking around" with high subject fidelity. Furthermore, in Figure 5(c), when we want to generate a video of a musician playing a given guitar,

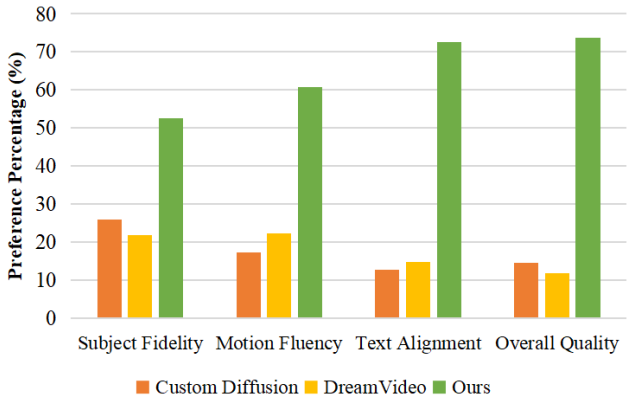

Figure 6: User Study. Our CustomCrafter achieves the best human preference compared with other comparison methods

we find that existing methods greatly damage the model's ability to combine concepts. They cannot generate a musician playing the guitar, and the motion is "frozen". Similarly, in Figure 5(d), when we want to generate a video of a child handing out the dice toy, a similar situation occurs. Our method successfully generated the combination of concept of a child and a toy dice, and has smooth motions. Therefore, without guidance from additional videos, our method significantly outperforms existing methods in terms of concept combination ability and motion naturalness, and has better subject fidelity. Please refer to the supplementary material for more visualizations and demonstration videos.

User Study

To further validate the effectiveness of our method, we conducted a human evaluation of our method and existing methods without using additional video data as guidance. We invited 20 professionals to evaluate the 30 sets of generated video results. For each group, we provided subject images and videos generated using the same seed and the same text prompt under different methods for comparison. We evaluated the quality of the generated videos in four dimensions: Text Alignment, Subject Fidelity, Motion Fluency, and Overall Quality. Text Alignment evaluates whether the generated video matches the text prompt. Subject Fidelity measures whether the generated object is close to the reference image. Motion Fluency is used to evaluate the quality of the motions in the generated video. Overall Quality is used to measure whether the quality of the generated video overall meets user expectations. As shown in Figure 6, our method has gained significantly more user preference in all metrics, proving the effectiveness of our method.

Ablation Study

In this section, we construct ablation studies to validate the effectiveness of each component. As shown in Table 2, we choose Custom Diffusion as the baseline to present the quantitative results of our designed Spatial Subject Learning Module and Dynamic Weighted Video Sampling Strategy. It can be observed that using our Spatial Subject Learning

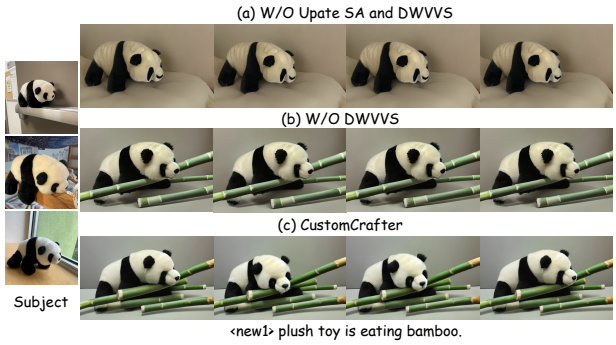

Figure 7: Effect of each design of our method. It can be seen that each of our designs has achieved the expected effect.

SSLM	DWVSS	CLIP-T↑	CLIP-I↑	DINO-I↑	T.Cons.↑
		0.286	0.769	0.583	0.992
✓		0.294	0.790	0.631	0.993
✓	✓	0.318	0.786	0.627	0.994

Table 2: Ablation Study. "SSLM" is Spatial Subject Learning Module, "DWVSS" is Dynamic Weighted Video Sampling Strategy.

Module achieves better results on the CLIP-I and DINO-I metrics, which measure the subject fidelity. This suggests that compared to previous work, our method is more capable of capturing the details of the subject. The Dynamic Weighted Video Sampling Strategy, due to modifications in the process, may result in a slight impairment of the subject's appearance similarity. However, the motions can be significantly improved, substantially enhancing the text alignment. In addition, we use visualization results to demonstrate the effectiveness of our method. As shown in Figure 7, when we only update the parameters of cross-attention, we find that the model's ability to combine concepts is poor and it cannot generate a simple concept combination of pandas and bamboo that matches the prompt. However, when we use our Spatial Subject Learning Module module to update both cross-attention and self-attention, and do not use the Dynamic Weighted Video Sampling Strategy, the generated video's ability to combine concepts is improved, but the VDM cannot generate fluent motions that match the text prompt. After adopting our sampling strategy, the generated video has almost no significant loss in subject fidelity, but the naturalness of the motion has greatly improved.

Conclusion

In this paper, we introduce our CustomCrafter, a novel framework for customized video generation. This approach does not require additional video to repair motion generation ability. We first designed a Spatial Subject Learning Module, which updates the Spatial Attention for subject learning. Simultaneously, we proposed a Dynamic Weighted Video Sampling Strategy, which improves the model's inference process to restore the motion generation capability of VDMs. Through experiments, we have demonstrated that our method is better than existing methods.

Acknowledgements

This work is supported in part by National Science Foundation for Distinguished Young Scholars under Grant 62225605, Project 12326608 supported by NSFC, Natural Science Foundation of Shanghai under Grant 24ZR1425600, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01020), as well as the Ningbo Science and Technology Innovation Project (No.2024Z294), and sponsored by CCF-Tencent Rhino-Bird Open Research Fund and Research Fund of ARC Lab, Tencent PCG.

References

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proc. CVPR*, 22560–22570.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2310.19512*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024a. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proc. CVPR*, 7310–7320.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024b. Anydoor: Zero-shot object-level image customization. In *Proc. CVPR*, 6593–6602.

Dou, H.; Li, R.; Su, W.; and Li, X. 2024. GVDIFF: Grounded Text-to-Video Generation with Diffusion Models. *arXiv preprint arXiv:2407.01921*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *Proc. ICCV*, 7346–7356.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proc. ICLR*.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. In *Proc. ICCV*.

Gu, J.; Wang, Y.; Zhao, N.; Fu, T.-J.; Xiong, W.; Liu, Q.; Zhang, Z.; Zhang, H.; Zhang, J.; Jung, H.; et al. 2024a. Photoswap: Personalized subject swapping in images. *Proc. NeurIPS*, 36.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024b. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Proc. NeurIPS*.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *Proc. ICLR*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proc. ICCV*, 7323–7334.

He, X.; Liu, Q.; Qian, S.; Wang, X.; Hu, T.; Cao, K.; Yan, K.; Zhou, M.; and Zhang, J. 2024. ID-Animator: Zero-Shot Identity-Preserving Human Video Generation. *arXiv preprint arXiv:2404.15275*.

He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent Video Diffusion Models for High-Fidelity Video Generation with Arbitrary Lengths. *arXiv preprint arXiv:2211.13221*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Proc. NeurIPS*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. In *Proc. NeurIPS*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, L.; Wang, W.; Wu, Z.-F.; Dou, H.; Shi, Y.; Feng, Y.; Liang, C.; Liu, Y.; and Zhou, J. 2024a. Group Diffusion Transformers are Unsupervised Multitask Learners. *arXiv preprint arXiv:2410.15027*.

Huang, L.; Wang, W.; Wu, Z.-F.; Shi, Y.; Dou, H.; Liang, C.; Feng, Y.; Liu, Y.; and Zhou, J. 2024b. In-Context LoRA for Diffusion Transformers. *arXiv preprint arXiv:2410.23775*.

Jiang, R.; Zheng, G.-C.; Li, T.; Yang, T.-R.; Wang, J.-D.; and Li, X. 2024. A Survey of Multimodal Controllable Diffusion Models. *Journal of Computer Science and Technology*, 39(3): 509–541.

Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2023. VideoBooth: Diffusion-based Video Generation with Image Prompts. *arXiv preprint arXiv:2312.00777*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proc. CVPR*, 1931–1941.

Li, D.; Li, J.; and Hoi, S. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Proc. NeurIPS*, 36.

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In *Proc. CVPR*, 7817–7826.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2024. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, 1–12.

Miao, P.; Su, W.; Wang, G.; Li, X.; and Li, X. 2023. Self-Paced Multi-Grained Cross-Modal Interaction Modeling for Referring Expression Comprehension. *IEEE Trans. Image Process.*

Nam, J.; Kim, H.; Lee, D.; Jin, S.; Kim, S.; and Chang, S. 2024. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proc. CVPR*, 8100–8110.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proc. CVPR*, 22500–22510.

Ruiz, N.; Li, Y.; Jampani, V.; Wei, W.; Hou, T.; Pritch, Y.; Wadhwa, N.; Rubinstein, M.; and Aberman, K. 2024. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proc. CVPR*, 6527–6536.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proc. CVPR*, 8543–8552.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Su, W.; Miao, P.; Dou, H.; Fu, Y.; and Li, X. 2023a. Referring expression comprehension using language adaptive inference. In *Proc. AAAI*, volume 37, 2357–2365.

Su, W.; Miao, P.; Dou, H.; Wang, G.; Qiao, L.; Li, Z.; and Li, X. 2023b. Language adaptive weight generation for multi-task visual grounding. In *Proc. CVPR*, 10857–10866.

Wang, C.; Guo, Z.; Duan, Y.; Li, H.; Chen, N.; Tang, X.; and Hu, Y. 2024a. Target-Driven Distillation: Consistency Distillation with Target Timestep Selection and Decoupled Guidance. *arXiv preprint arXiv:2409.01347*.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, W.; Yang, H.; Tuo, Z.; He, H.; Zhu, J.; Fu, J.; and Liu, J. 2023b. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *arXiv preprint arXiv:2305.10874*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023c. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*.

Wang, Z.; Li, A.; Xie, E.; Zhu, L.; Guo, Y.; Dou, Q.; and Li, Z. 2024b. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*.

Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proc. CVPR*, 6537–6549.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proc. ICCV*, 15943–15953.

Wu, T.; Li, X.; Qi, Z.; Hu, D.; Wang, X.; Shan, Y.; and Li, X. 2024a. SphereDiffusion: Spherical Geometry-Aware Distortion Resilient Diffusion Model. In *Proc. AAAI*, volume 38, 6126–6134.

Wu, Y.; Zhou, X.; Ma, B.; Su, X.; Ma, K.; and Wang, X. 2024b. Ifadapter: Instance feature control for grounded text-to-image generation. *arXiv preprint arXiv:2409.08240*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *Int. J. Comput. Vis.*, 1–20.

Yang, L.; Shen, D.; Cai, C.; Yang, F.; Li, S.; Zhang, D.; and Li, X. 2024a. Solving token gradient conflict in mixture-of-experts for large vision-language model. *arXiv preprint arXiv:2406.19905*.

Yang, L.; Zhao, H.; Yu, Y.; Zeng, X.; and Li, X. 2024b. RCS-Prompt: Learning Prompt to Rearrange Class Space for Prompt-based Continual Learning. In *Proc. ECCV*.

Yang, L.; Zhou, X.; Li, X.; Qiao, L.; Li, Z.; Yang, Z.; Wang, G.; and Li, X. 2023. Bridging cross-task protocol inconsistency for distillation in dense object detection. In *Proc. ICCV*, 17175–17184.

Yu, S.; Sohn, K.; Kim, S.; and Shin, J. 2023. Video probabilistic diffusion models in projected latent space. In *Proc. CVPR*, 18456–18466.

Yuan, G.; Cun, X.; Zhang, Y.; Li, M.; Qi, C.; Wang, X.; Shan, Y.; and Zheng, H. 2024. Inserting Anybody in Diffusion Models via Celeb Basis. *Proc. NeurIPS*, 36.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, Y.; Xing, J.; Lo, E.; and Jia, J. 2024. Real-world image variation by aligning diffusion inversion chain. *Proc. NeurIPS*, 36.

Zhao, H.; Lu, T.; Gu, J.; Zhang, X.; Wu, Z.; Xu, H.; and Jiang, Y.-G. 2023. Videoassembler: Identity-consistent video generation with reference entities using diffusion model. *arXiv preprint arXiv:2311.17338*.

Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proc. CVPR*, 22490–22499.