ImPoster : Text and Frequency Guidance for Subject Driven Action Personalization using Diffusion Models

Divya Kothandaraman, Kuldeep Kulkarni*, Sumit Shekhar*, Balaji Vasan Srinivasan*, Dinesh Manocha
University of Maryland College Park, Adobe Research*

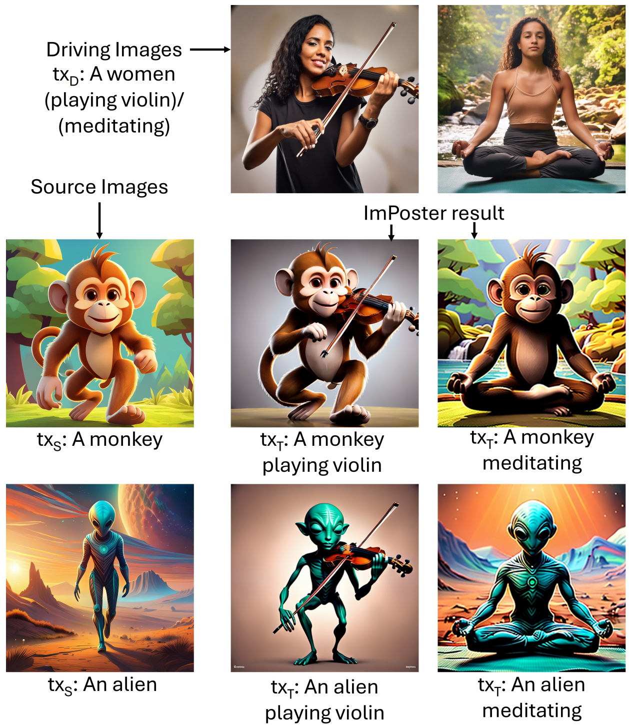

Figure 1: *Given a single "source" image and a "single" driving image and the corresponding text descriptions, ImPoster generates an image of the source subject performing driving action.* We show how ImPoster is able to make a monkey and an alien meditate and play violin.

Abstract

We present ImPoster, a novel algorithm for generating a target image of a 'source' subject performing a 'driving' action. The inputs to our algorithm are a single pair of a source image with the subject that we wish to edit and a driving image with a subject of an arbitrary class performing the driving action, along with the text descriptions of the two images. Our approach is completely unsupervised and does not require any access to additional annotations like keypoints or pose. Our approach builds on a pretrained text-to-image latent diffusion model and learns the characteristics of the source and the driving image by finetuning the diffusion model for a small number of iterations. At inference time, ImPoster performs *step-wise text prompting* i.e. it denoises by first moving in the direction of the image manifold corresponding to the driving image followed by the direction of the image manifold corresponding to the text description of the desired target image. We propose a novel diffusion guidance formulation, *image frequency guidance*, to steer the generation towards the manifold of the source subject and the driving action at every step of the inference denoising. Our frequency guidance formulations are derived from the frequency domain properties of images. We extensively evaluate ImPoster on a diverse set of source-driving image pairs to demonstrate improvements over baselines. To the best of our knowledge, ImPoster is the first approach towards achieving both subject-driven as well as action-driven image personalization. The code and data are available at https://github.com/divyakraman/ ImPosterDiffusion2024.

Proceedings of the 31st International Conference on Computational Linguistics, pages 11013–11028
January 19–24, 2025. ©2025 Association for Computational Linguistics

1 Introduction

Monkeys seldom meditate! Neither do they play the violin! Generative AI has enabled the creation of images that cannot be easily enacted or photographed in real life. More often than not, applications such as animation, movie creation, etc, require a particular subject performing a specific action. For example, given a 'source image' of a dog and a 'driving image' of a cat drinking water from a mug, we might want to generate an image of the dog drinking water from a mug in exactly the same manner as the cat. Thus, the generated dog needs to have the same identity as the 'source subject'. Its pose or the style with which it drinks water from the mug needs to resemble the 'driving pose', i.e. the pose of the cat in the driving image. Consequently, the specific action depicted in the generated image and the identity of the subject performing the action are not arbitrary - rather, they are dictated by the source and the driving images.

Prior work on the closely related problem of motion transfer (Zhao and Zhang, 2022; Tao et al., 2022; Siarohin et al., 2021, 2019b; Shalev and Wolf, 2022) condition on SMPL pose information (Yoon et al., 2021), keypoints, etc in a supervised/ unsupervised manner and train on large datasets to transfer the action depicted by a subject in a driving image or video to the subject in the source image. Such methods however do not generalize well to arbitrary pairs of source and driving images and often require a lot of data for training. Thus, it is useful to develop an approach that can directly work on a single pair of source-driving images in a completely unsupervised manner.

Text is an excellent auxiliary modality to guide the generation process, and the recent progress in text-to-image diffusion models (Rombach et al., 2022; Saharia et al., 2022a; Ho et al., 2022) motivates their usage for conreolled image generation. Prior work on personalized text-guided image editing such as DreamBooth (Ruiz et al., 2022), IMAGIC (Kawar et al., 2022b), custom diffusion (Kumari et al., 2022), InstantBooth (Shi et al., 2023), (Gal et al., 2023), ELITE (Wei et al., 2023) and SUTI (Chen et al., 2023b) are able to add a specific subject to the model while finetuning the diffusion model to generate various actions dictated by the text input. However, the user does not have any control over the exact manner in which they might like the subject to perform the action in the generated image. For instance, a dog can drink water from a mug in various ways - we want the model to be able to generate a specific pose corresponding to the action, as defined by the driving image. On the other hand, prior work such as blended diffusion (Avrahami et al., 2022) and DiffusionCLIP (Kim et al., 2022) are able to generate a specific action and change the style of the image. However, they are unable to make a source subject perform a specific driving action.

Main contributions. We propose an algorithm, ImPoster, for generating an image of a specific subject from a source image performing a specific action described in the driving image. Given a single source-driving image pair and the corresponding text descriptions, our method can perform 'body' transformations to the source subject, as dictated by the action depicted in the driving image. Our method builds on a pre-trained text-to-image diffusion model to perform test-time optimization/ inference and does not require any additional information such as pose or keypoints. Our method, ImPoster , first learns the characteristics of the source and the driving image by finetuning the pretrained diffusion model on the source and driving corresponding text-image pair. Our inference guidance methods, which form the novel contributions of this paper, enable the generation of the desired target image:

1. **Stepwise text prompting.** ImPoster generates the desired target image of the source subject performing the specific driving action using a *stepwise text prompting* strategy. Since geometric or structural information is much harder to generate, ImPoster lays the foundation for generating the driving action by first moving towards the direction of the manifold of the driving image. This step is followed by switching directions and denoising by moving in the direction of the text description corresponding to the desired target image.

2. **Image frequency guidance.** While the stepwise text prompting method provides a strong prior for the driving action, it is insufficient to generate an accurate image of the source subject. To alleviate this issue, we delve into the frequency domain representation of an image, a powerful space for understanding how the human brain interprets natural images (Oppenheim and Lim, 1981; Xu et al., 2021; Yang and Soatto, 2020; Yang et al., 2020). Moti-

vated by guidance techniques (Ho and Salimans, 2022; Bansal et al., 2023) in diffusion models, we harness the amplitude map of the Fourier transform of image features and present frequency amplitude guidance to preserve the characteristics of the source subject. In order to prevent the loss of the specific driving action (or structure) while extracting source subject characteristics, frequency amplitude guidance is complemented by frequency phase guidance. This is inspired by the fact that the phase of the Fourier transform of an image is representative of its geometry and structure. Frequency amplitude guidance and frequency phase guidance, termed as *frequency guidance*, guide the diffusion denoising towards generating the source subject performing the specific driving action.

We apply ImPoster on a wide variety of source-driving image pairs on a curated dataset with 120 source-driving image pairs. ImPoster can make an elephant read a book (described by a human reading a book), a monkey meditate and perform push-ups (described by a human performing meditating and doing pushups), a teddy bear play guitar (described by a human playing guitar), etc. We also show the effectiveness of stepwise inference and our frequency guidance method, along with qualitative and quantitative comparisons against prior work metris such as CLIP Score, SSCD, DINO and a new metric to quantify the alignment with driving action, phase score, defined using the phase of the Fourier transform.

In summary, the contributions of this paper are as follows: (i) We formalize a novel task of generating images consistent with a source subject and driving action; (ii) We propose ImPoster, a novel diffusion models based method that can address this pragmatic task; (iii) We curate a dataset of 120 source-driving image pairs; (iv) In order to quantitatively establish the correspondence between the generated image and driving action, we propose a new metric 'Phase score' based on the phase of the Fourier transform; (v) Finally, our exhaustive experimental results reveal large gains over baselines; thereby setting a new benchmark for the task.

2 Related work

Exemplar image animation. A great amount of literature has been dedicated to exemplar image animation (Shalev and Wolf, 2022; Siarohin et al., 2021, 2019b; Tao et al., 2022; Zhao and Zhang, 2022; Siarohin et al., 2019a, 2018; Tao et al., 2022) transfer motion characteristics from a driving video to a source image. These methods at the core of them transfer motion at frame level from a driving frame onto the source image, similar to our method. However, these models are restricted in their domain to only human and almost always require additional annotations like keypoints, 2D poses, 3D poses or require computation of optical flow.

Diffusion models for Text-based image editing/ personalization. Recent progress in generative AI has shown promising results using diffusion models (Nichol et al., 2021; Wang et al., 2022a; Su et al., 2022; Sasaki et al., 2021; Saharia et al., 2022a,b; Yang et al., 2022; Preechakul et al., 2022; Zhang et al., 2023) for performing non-trivial operations, such as posture changes and multiple objects editing. Text-based image editing and personalization approaches (Ruiz et al., 2022; Kawar et al., 2022b; Kumari et al., 2022; Balaji et al., 2022; Gal et al., 2023; Kothandaraman et al., 2023a; Huang et al., 2023; Brooks et al., 2022; Kothandaraman et al., 2024a; Zheng et al., 2022; Shi et al., 2023; Zhang et al., 2022; Ma et al., 2023b; Gal et al., 2022; Shi et al., 2023; Wei et al., 2023; Chen et al., 2023b; Kumari et al., 2022; Han et al., 2023; Qiu et al., 2023; Ma et al., 2023a; Xiao et al., 2023) add a subject to the diffusion using a few images of the subject followed by using text to manipulate the image to obtain the desired output. The methods for personalized or subject-driven based image editing can be categorized into two broad categories. First category of works (Ruiz et al., 2022; Kawar et al., 2022b; Wei et al., 2023; Kumari et al., 2022; Gal et al., 2023; Yang et al., 2022) perform fine-tune a pre-trained model on a small number of images that involve the subject and then perform inference optimization with the text depicting the action as the input. While these methods are successful in generating images with the subject performing an action depicted by the text, the user does not have control over the specific pose or imitable characteristic. The second category of works (Shi et al., 2023; Chen et al., 2023b) address the shortcomings of fine-tuning the model and propose methods that are purely based on inference-time optimization and hence much faster. However, even they suffer from the same drawbacks in not being able to give users the control of the action depicted. Multi-concept

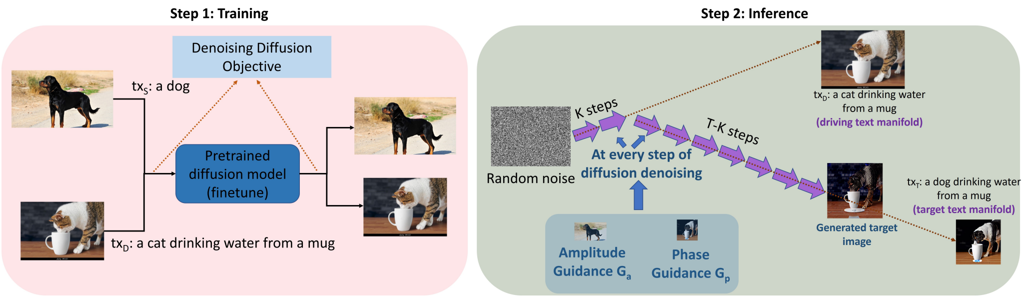

Figure 2: Given a single source-driving image pair, ImPoster generates an image of the source subject performing the action depicted in the driving image. ImPoster first finetunes the text-to-image diffusion model on the source-driving image pair. At inference, ImPoster begins by first denoising in the direction of the driving image manifold followed by moving towards the manifold corresponding to the desired target image. At every step of the inference, frequency guidance steers the generation of an image with source subject characteristics and driving action.

customization methods (Kumari et al., 2022; Ma et al., 2023a; Xiao et al., 2023; Han et al., 2023) are not very effective either. For instance, one of the closest works to ours is Custom Diffusion (Kumari et al., 2022) that enable generating images by fusing multiple concepts from multiple images. However, this method is not well suited to generate images involving concepts of actions.

Recent methods on motion customization (Wu et al., 2023; Kothandaraman et al., 2024b; Chen et al., 2023a) of diffusion models are able to transfer the motion from a video to subjects. However, these methods rely on temporal properties of videos to tranfer the motion. Our problem statement is tangential to video motion customization, wherein the goal is to customize the action from an image.

Guidance methods for diffusion. Guidance methods (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Kawar et al., 2022a; Wang et al., 2022b; Chung et al., 2022a; Lugmayr et al., 2022; Chung et al., 2022b; Graikos et al., 2022; Kothandaraman et al., 2023b) have been used to control and guide diffusion denoising. One of the first guidance methods is classifier and classifier-free guidance (Dhariwal and Nichol, 2021; Ho and Salimans, 2022) that reinforce the class of the object in the generated image. Bansal et. al. (Bansal et al., 2023) proposed universal guidance to guide the generation using segmentation maps, sketches, etc. Kothandaraman et. al. (Kothandaraman et al., 2023b) proposed a mutual information based guidance method to generate high fidelity images from various viewpoint.

3 Method

Given a source image I_S and a driving image I_D, ImPoster generates an image of the source subject performing the driving action. We assume access to the text descriptions tx_S and tx_D for the source and driving images respectively. The text description, tx_T, corresponding to the desired target image I_T is a modified combination of tx_S and tx_D. For example, if tx_S is "A dog" and tx_D is "A cat drinking water from a mug", tx_T would be "A dog drinking water from a mug". We assume no access to any training dataset, pose information (such as SMPL), keypoints, etc. Our method is completely unsupervised and works on a single source-driving image pair.

An overview of ImPoster is as follows. To add knowledge pertaining to the source and the driving image, we finetune the diffusion model on I_S and I_D using tx_S and tx_D, respectively. During inferencing, we begin by denoising by first moving in the direction of the manifold corresponding to I_D followed by moving in the direction of the manifold corresponding to tx_T. Such a mechanism, which we term **stepwise text prompting** creates a prior for pose information followed by denoising to generate the desired subject performing the action. Further, to reinforce the generation of an image of the source subject performing the driving action, at every inference step, we apply a novel **image frequency guidance** strategy to explicitly steer the denoising towards the desired driving action and source subject characteristics. We now turn to describe our method in detail.

3.1 Training

A pretrained text-to-image diffusion model such as stable diffusion is trained on massive text-image data. However, it does not contain information specific to the source and driving image. In order to enable the diffusion model reconstruct I_S and I_D from tx_S and tx_D, respectively, we finetune the diffusion model for n_{tr} iterations on the source and driving text-image pairs using the denoising diffusion objective function (Ho et al., 2020). Specifically, let θ denote the U-net parameters, L denote the denoising diffusion objective, and e_S, eD denote the text embeddings corresponding to the source and driving texts. At each iteration $i < n_{tr}$ of finetuning, we perform the following optimizations:

$$\begin{aligned} &\min_{\theta} \sum_{t=T}^{0} L(f(x_t, t, e_S; \theta), I_S), \\ &\min_{\theta} \sum_{t=T}^{0} L(f(x_t, t, e_D; \theta), I_D). \end{aligned} \tag{1}$$

3.2 Inference

3.2.1 Stepwise text prompting

A key question that we need to keep in mind while developing an effective inference strategy is: *How should the noisy image be perturbed so that it better corresponds with the driving action and the source subject ?* Moving in the direction of the driving image manifold will ensure the reconstruction of the driving action accurately, however, biased towards the driving subject. Similarly, moving in the direction of the source image manifold will ensure the reconstruction of the source subject accurately, however, biased towards the source action. The image manifold corresponding to tx_T generally does not contain the desired target image – i.e. given tx_T, it is unable to generate the desired target image by naive diffusion denoising. On one hand, a pretrained diffusion model conditioned on tx_T is inclined to generate a wide variety of images, most not fully conformant with the specific characteristics of the source subject as well as the specific pose delineated by the driving action. On the other hand, there could be a bias towards reconstructing the source or driving image as well.

To solve the aforementioned issues, we propose a stepwise inference strategy. Let the number of inference steps be denoted by T_i. Starting from random noise, we denoise using tx_D for the first K iterations following by denoising using tx_T for the next $T - K$ iterations. This implies that the generation process begins by first moving towards the driving image manifold followed by the manifold corresponding to the target text. In the first few iterations, the sampler denoises to reconstruct the driving image, creating a strong prior for the driving action. In the subsequent iterations, denosing to move towards the target text manifold ensures that the generated image contains the source subject, as dictated by the target text.

On one hand, the target text points towards the driving action. On the other, denoising using tx_D in the first few iterations creates a strong prior for the structure of the generated image. This ensures that the generated image has the driving action. Note that we denoise with tx_D first followed by tx_T and vice versa. Diffusion works by sequentially denoising from random noise and is a slow process i.e. the updates in each step are small. Therefore, it is easier for the model to first generate the pose and then denoise to obtain subject characteristics. It is much more difficult to modify the pose than it is to modify the characteristics or identity of the subject in the image.

Despite the stepwise inference strategy, the generated image might still contain certain characteristics of the driving subject or source action. To alleviate this issue, we propose a novel frequency guidance method, which reinforces the model to stick to the driving action and source subject at every step of the generation process.

3.2.2 Frequency guidance

The frequency domain representation of an image provides rich information (Oppenheim and Lim, 1981) about the image. The amplitude of the 2D spatial Fourier transform of an image is representative of the intensities of different frequencies in the image, and represents changes in the spatial domain. It contains information about the geometrical structure of features in the image. The phase of the 2D spatial Fourier transform of an image represents the location of these features which help the human eye understand the image; it contains information regarding the edges, contours, etc. It is possible to reconstruct the grayscale counterpart of an image with just its phase representation. The amplitude information, along with the phase, helps in reconstructing the characteristics (colors, attributes, texture, identity) of the scene entities. However, the amplitude alone cannot generate a meaningful image; phase information is very important.

In light of this, it can be inferred that the phase of the Fourier transform of the driving image is indicative of the driving action. Similarly, the amplitude of the Fourier transform of the source image is indicative of the knowledge required to reconstruct the source image. We use the Fourier domain representation to **guide** the generation process. Frequency domain properties of an image propagate to its feature space (computed by a neural network). Hence, we can apply frequency guidance on the latent feature representation space of stable diffusion (Rombach et al., 2022). Classifier-free guidance (Ho and Salimans, 2022) and universal guidance (Bansal et al., 2023) mathematically derive guidance as,

$$\tilde{\epsilon_\theta}(z_t, t) = \epsilon_\theta(z_t, t) + s(t) \times \nabla_{z_t} G. \quad (2)$$

where z_t is the denoised latents at timestep t, ∇_{z_t}G is the gradient of the guidance function and $s(t)$ controls the strength of the guidance for each sampling step. An appropriately weighted version of this noise is subtracted from the latents computed in the previous timestep to obtained updated latents. Consequentially, rather than adjusting the noise predicted by the diffusion model using the gradients of the guidance function, we can directly modify the computed latents using a suitable guidance function using the gradient of the computed amplitude and phase functions w.r.t. the latents. At every step of the inference sampling process, we modify the computed latents z_t as,

$$\tilde{z_t} = z_t - s_a \times \nabla_{z_t} G_a - s_p * \times \nabla_{z_t} G_p, \quad (3)$$

where s_a and s_p are the scaling factors for the frequency amplitude guidance and frequency phase guidance functions, G_a and G_p, respectively. Note that z_t is computed using the noise predicted by the diffusion model at each timestep. G_a is the $L2$ distance between the amplitude of the generated latents and the amplitude of the latents of the source image. G_p is the $L2$ distance between the phase of the generated latents and the phase of the latents of the driving image. G_a drives the generation towards the source subject at every step of the sampling. G_p reinforces the driving pose at every step of the sampling and prevents any distortion in pose that may be caused by G_a.

4 Experiments and Results

Dataset and implementation details. Since there is no dataset for this task, we collect a dataset with 15 driving actions and 8 source subject, resulting in a total of 120 source-driving image pairs. The source subjects and driving actions are generated using Adobe Firefly. This dataset curation is inspired by prior papers that defined new tasks in diffusion personalization such as DreamBooth (Ruiz et al., 2022)(30 subjects), Custom Diffusion (Kumari et al., 2022)(5 concepts, 8 prompts for the multi-concept setting), and Concept Decomposition (Vinker et al., 2023)(15 pairs for application 1, 13 concepts for application 2) and is also comparable in size to these prior datasets. Consistent with prior work, we generate results corresponding to 5 different random seeds for each source-driving pair, resulting in a total of 600 images being used for all quantitative analysis.

We finetune the diffusion model for $n_{tr} = 500$ iterations. We set s_a to $1e-6$, $s_p = 1e-3$, $k = 5$ and $T = 50$. We use an image size of 512×512. Our model takes about $51/2$ minutes per image pair on one NVIDIA RTX A5000 GPU. We use the Stable Diffusion 2.1 model as the backbone. We add LoRA (Hu et al., 2021) layers to the backbone model for the finetuning step, the rest of the model is frozen.

Qualitative results.

State-of-the-art comparisons. Closest related to our work are image editing or personalization approaches. The most recent and effective methods in this domain are DreamBooth (Ruiz et al., 2022)(CVPR 2023) and Custom Diffusion (Kumari et al., 2022)(CVPR 2023), inspired by which, we compare with an enhanced baseline, Baseline++. Baseline++ takes in a single source image and driving image along with their corresponding text prompts and finetunes the diffusion model (similar to DreamBooth (Ruiz et al., 2022) + LoRA (Hu et al., 2021)). Next, it uses a pragmatic combination of the text prompts corresponding to the source and the driving image (tx_T) to generate the target image (similar to Custom Diffusion (Kumari et al., 2022)).

We show comparisons with Baseline++ in Figure 3. Baseline++, due to bias issues with respect to the pose in the source image, and inability to effectively generate the driving action, is unable to generate the source subject performing driving action. In contrast, ImPoster holistically distils the driving action and the characteristics of the source subject through its stepwise text prompting and image frequency guidance strategies to achieve a good bias-variance trade-off and generate the source subject performing driving action.

Quantitative Results: Comparisons with Baseline++ and Ablations In concordance with prior work (Ruiz et al., 2022; Kawar et al., 2022b; Kumari et al., 2022) on diffusion models for text-based image editing/ personalization, we evaluate our

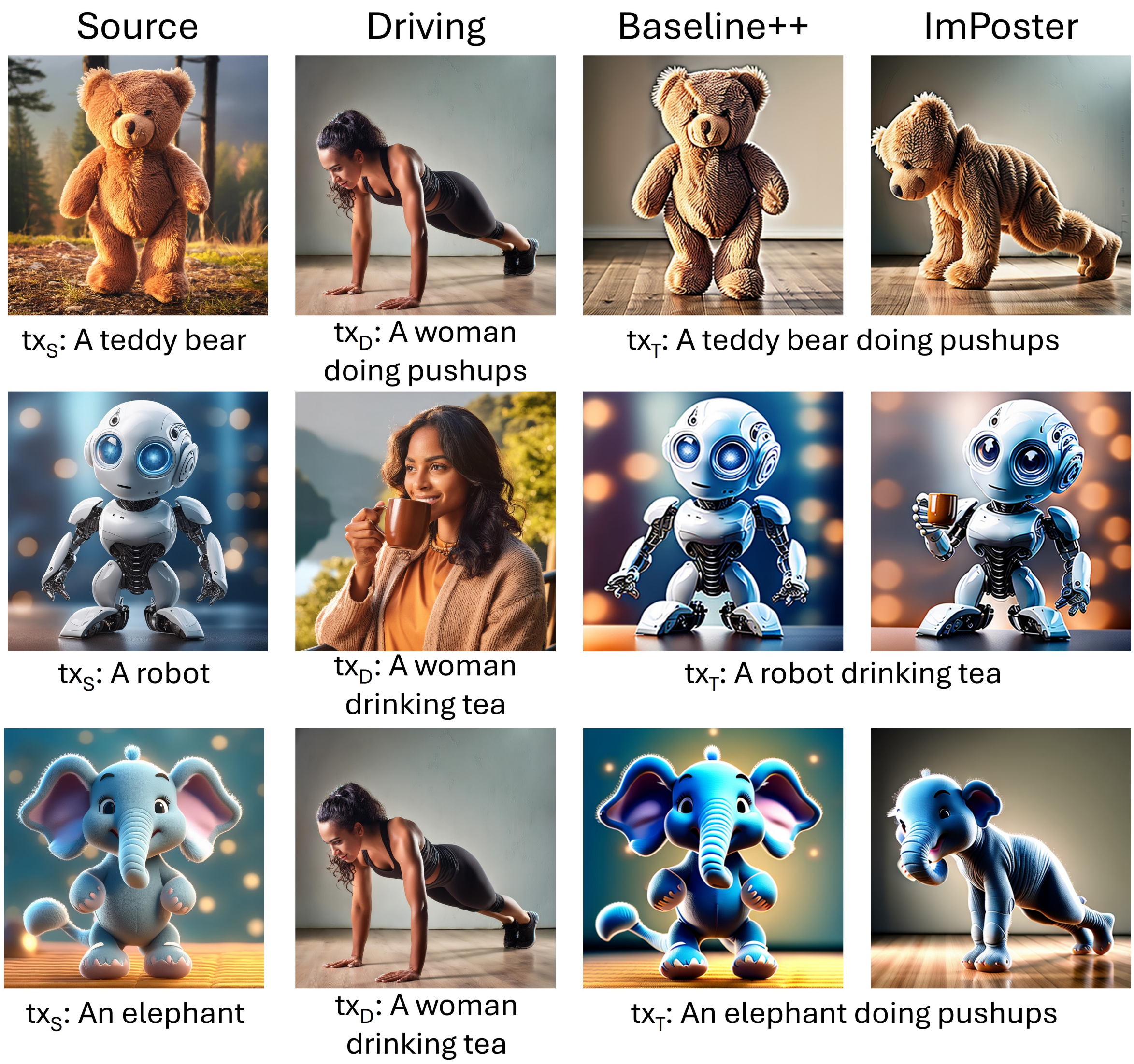

Figure 3: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues. Please see the appendix for more results generated using ImPoster and comparisons with Baseline++.

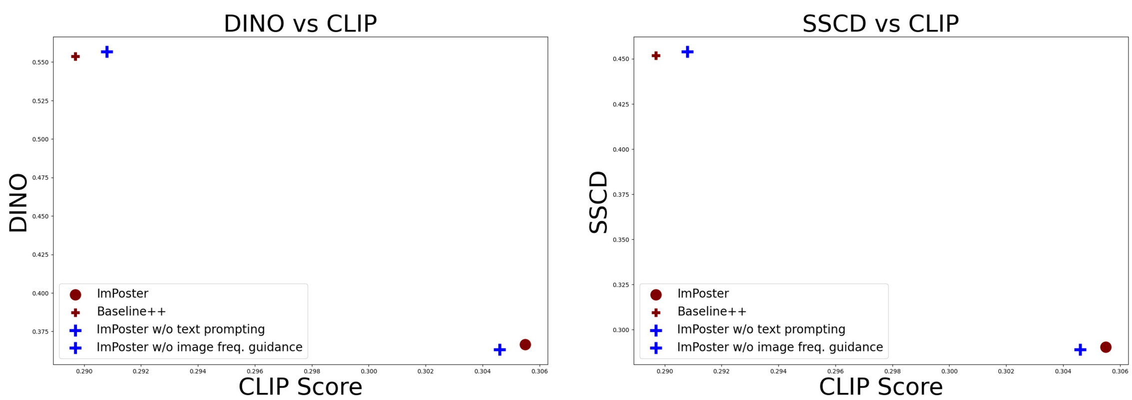

Figure 4: ImPoster is able to successfully transfer the driving motion while retaining the characteristics of the source subject, and achieves a better trade-off between driving action (CLIP/Phase score) and source subject (SSCD/DINO) than prior work, as also evidenced by our qualitative results. Stepwise text prompting creates a prior for the driving action to enable the model generate the driving action. The image frequency guidance formulations help the model in preserving the characteristics of the source subject, while reinforcing the driving action.

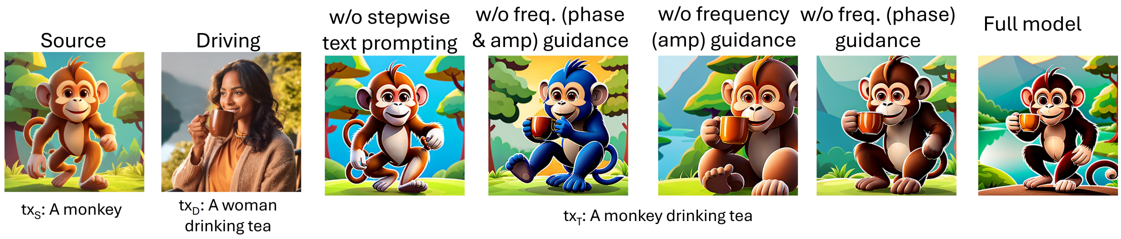

Figure 5: **Ablations.** Without stepwise text prompting, there is no prior for the driving action, which inhibits the model from generating the driving action accurately. The frequency (amplitude and phase) guidance methods help in generating the characteristics of the source subject (here, monkey) accurately - notice that there are changes to the color of the monkey (column 6), missing details in the fingers (column 4), changes in the size of the monkey (column 5).

method using the following quantitative metrics. We compute the averages using all 600 generated images. The results are in Figure 4.

1. **Text alignment:** We evaluate the alignment of the generated target image with the target text using the CLIP score (Ruiz et al., 2022). Higher CLIP scores indicate higher alignment with text. Here, CLIP Score gives us a broad overview of the alignment of the generated image with the subject and action. ImPoster achieves a far higher CLIP score than Baseline++, indicating its ability to effectively transfer the driving action to the source subject.

2. **Fidelity:** We evaluate the fidelity of the subject in the generated target image the source subject using self-supervised similarity metrics - SSCD (Pizzi et al., 2022) and DINO scores (Caron et al., 2021). Baseline++ is unable to generate the driving action, and simply replicates the source image. This results in it having a higher value of SSCD and DINO score than ImPoster, which is able to transfer the driving action, as well as preserve the characteristics of the source image.

3. **Phase score:** To evaluate the correctness of the action or pose generated in the target image (as dictated by the driving image), we define a new metric called the phase score. As per classical computer vision literature (Oppenheim and Lim, 1981), the phase of the Fourier transform of the image is indicative of the action depicted in the image. We compute phase score as the cosine similarity between the phase of the Fourier transform of the driving image and the generated target image. Higher cosine score indicates higher similarity. ImPoster achieves a higher phase score of 0.7529 as compared to Baseline++'s phase score of 0.7515, indicating its ability to transfer the driving action effectively.

Ablation analysis. a. Overall alignment with driving action and source subject: ImPoster achieves a higher CLIP score than all ablations, indicating the usefulness of each component of our model towards transferring the driving action to the source subject, while preserving its characteristics. The stepwise text prompting and frequency guidance methods (including phase frequency guidance and amplitude frequency guidance) are complementary to each other and work in a holistic manner to achieve the desired goal. **b. Effectiveness of stepwise text prompting:** The stepwise text prompting strategy provides a crucial signal for generating the driving image, without which, the model achieves an overall low CLIP score as well as Phase score. Similar to Baseline++, the model without the stepwise text prompting strategy has a tendency to replicate the source image without the driving action, resulting in a higher SSCD or DINO score. **c. Effectiveness of image frequency guidance:** Without the frequency guidance functions, the SSCD and DINO scores drop, indicating its effectiveness in preserving fidelity w.r.t. source subject while transferring the driving action.

Please refer to the appendix for (i) more qualitative results, (ii) failure cases.

5 Conclusions

We propose the usage of infusing text and image frequency for the task of method for the image generation task of subject-driven action personalization. To the best of our knowledge, ours is the first approach towards this task. Our method is successful in achieving a wide range of image animations such as making a monkey meditate and play the violin, as also validated by our quantitative

Figure 6: **Failure cases.** Our method is ineffective in a few cases - it is unable to make a rabbit perform yoga or dance, and a cat perform yoga or meditate. We believe that this is due to bias-variance trade-off issues, that remain unresolved even after the application of our stepwise text prompting and image frequency guidance strategies. Usage of stronger vision and language backbones in the text to image generation pipeline, as they are made open-source to the community, can help the model disentagle features of the image better to alleviate these bias-variance trade-off issues. Besides, further research on this problem can lead to the development of newer methods that can help improve these results.

results on our curated dataset using various metrics, including the newly proposed 'phase score'. We hope our paper inspires further research in the area.

6 Limitations and Future Work

While our method is able to execute large non-trivial non-rigid pose transfers to specific subjects, as defined by a driving image, it is still limited in the ability of change that it can bring forth. For instance, it is unable to make a rabbit or a cat do yoga, indicating that there is scope for further research. Another direction for future research, also a widespread issue in the text-based image personalization literature, is the development of better quantitative metrics for comprehensively measuring subject fidelity and the desired edit. Since self-supervised methods such as SSCD and DINO measure image-level similarity, they result in a high value even if there is absolutely no action transfer – this leads to inappropriate evaluation of the models' capability in performing the desired edit (driving action transfer in our case), while maintaining source subject fidelity. More directions for future work including investigating our frequency guidance strategy for other image editing and personalization applications, extension to video applications and scenarios that involve more than one source subject or driving action in the scene.

Societal impact. Research on identification of fake imagery is essential to prevent the malicious use of our method.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Universal guidance for diffusion models. *arXiv preprint arXiv:2302.07121*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Hong Chen, Xin Wang, Guanning Zeng, Yipeng Zhang, Yuwei Zhou, Feilin Han, and Wenwu Zhu. 2023a.

Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*.

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. 2023b. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2022a. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. 2022b. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*.

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. 2022. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022a. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022b. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435.

Divya Kothandaraman, Ming Lin, and Dinesh Manocha. 2024a. Prompt mixing in diffusion models using the black scholes algorithm. *arXiv preprint arXiv:2405.13685*.

Divya Kothandaraman, Kihyuk Sohn, Ruben Villegas, Paul Voigtlaender, Dinesh Manocha, and Mohammad Babaeizadeh. 2024b. Text prompting for multi-concept video customization by autoregressive generation. *arXiv preprint arXiv:2405.13951*.

Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. 2023a. Aerial diffusion: Text guided ground-to-aerial view translation from a single image using diffusion models. *arXiv preprint arXiv:2303.11444*.

Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. 2023b. Aerialbooth: Mutual information guidance for text controlled aerial view synthesis from a single image. *arXiv preprint arXiv:2311.15478*.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2022. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. 2023a. Subject-diffusion: Open domain personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2307.11410*.

Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. 2023b. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Alan V Oppenheim and Jae S Lim. 1981. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. 2022. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. 2023. Controlling text-to-image diffusion by orthogonal finetuning. *arXiv preprint arXiv:2306.07280*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. 2021. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*.

Yoav Shalev and Lior Wolf. 2022. Image animation with perturbed masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3647–3656.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32.

Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*.

Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3637–3646.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13.

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. 2022a. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. 2022b. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings* *of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2023. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2022. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*.

Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. 2020. Phase consistent ecological domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9011–9020.

Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095.

Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. 2021. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2022. Sine: Single image editing with text-to-image diffusion models. *arXiv preprint arXiv:2212.04489*.

Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666.

Wanfeng Zheng, Qiang Li, Xiaoyan Guo, Pengfei Wan, and Zhongyuan Wang. 2022. Bridging clip and stylegan through latent alignment for image editing. *arXiv preprint arXiv:2210.04506*.

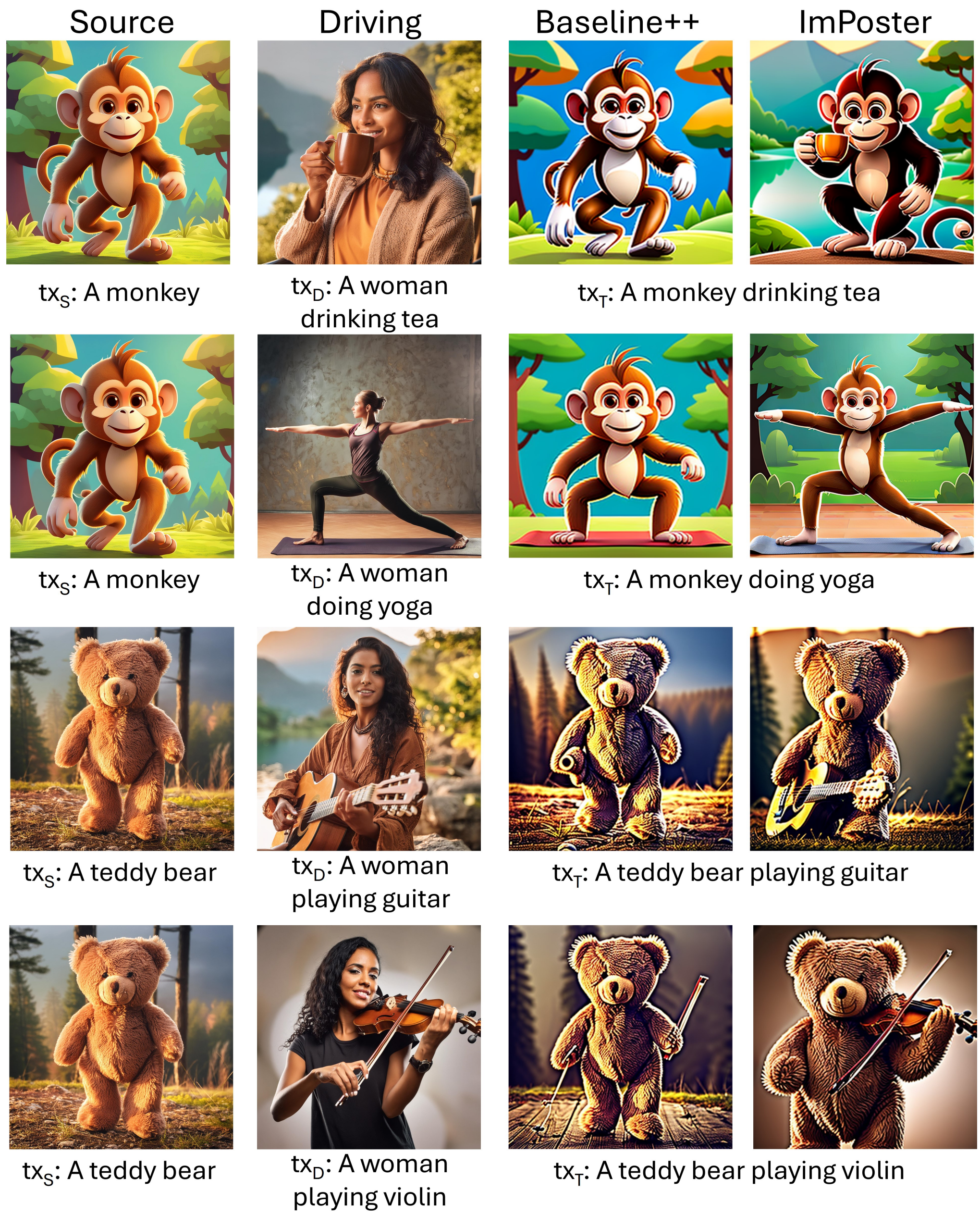

Figure 7: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

Method	CLIP	Phase Score	DINO	SSCD
Baseline++	0.2897	0.7515	0.5538	0.4517
ImPoster	0.3055	0.7529	0.3665	0.2903
Effectiveness of Stepwise text prompting: set k=0	0.2908	0.7516	0.5568	0.454
Effectiveness of image frequency guidance: set $s_a = 0, s_p = 0$	0.3046	0.7529	0.3633	0.2890
Effectiveness of image frequency (amp) guid. : set $s_p = 0$	0.3052	0.7528	0.3632	2884
Effectiveness of image frequency (phase) guid. : set $s_a = 0$	0.3062	0.7529	0.3629	2877

Table 1: ImPoster is able to successfully transfer the driving motion while retaining the characteristics of the source subject, and achieves a better trade-off between driving action (CLIP/Phase score) and source subject (SSCD/DINO) than prior work, as also evidenced by our qualitative results. Stepwise text prompting creates a prior for the driving action to enable the model generate the driving action. The image frequency guidance formulations help the model in preserving the characteristics of the source subject, while reinforcing the driving action.

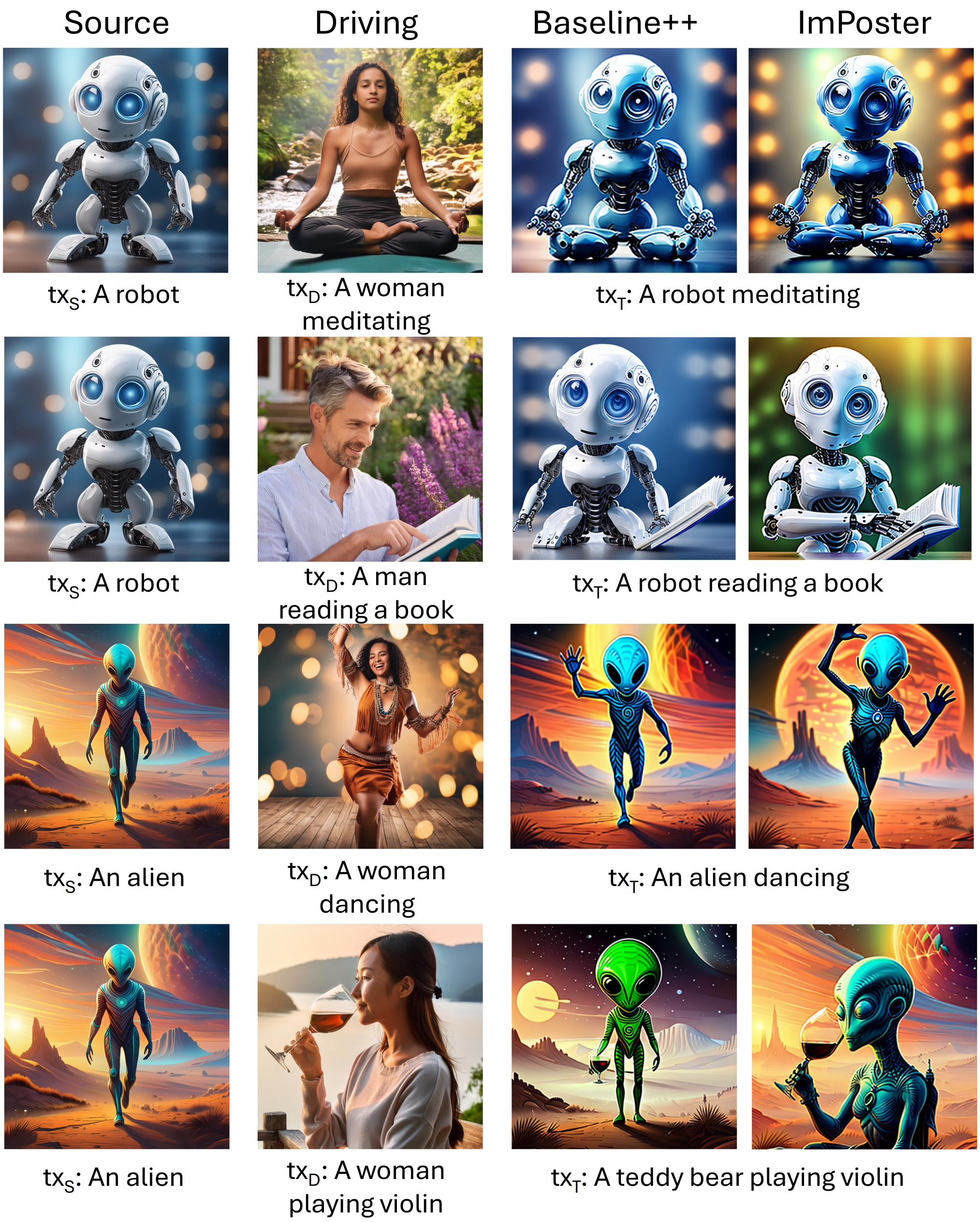

Figure 8: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

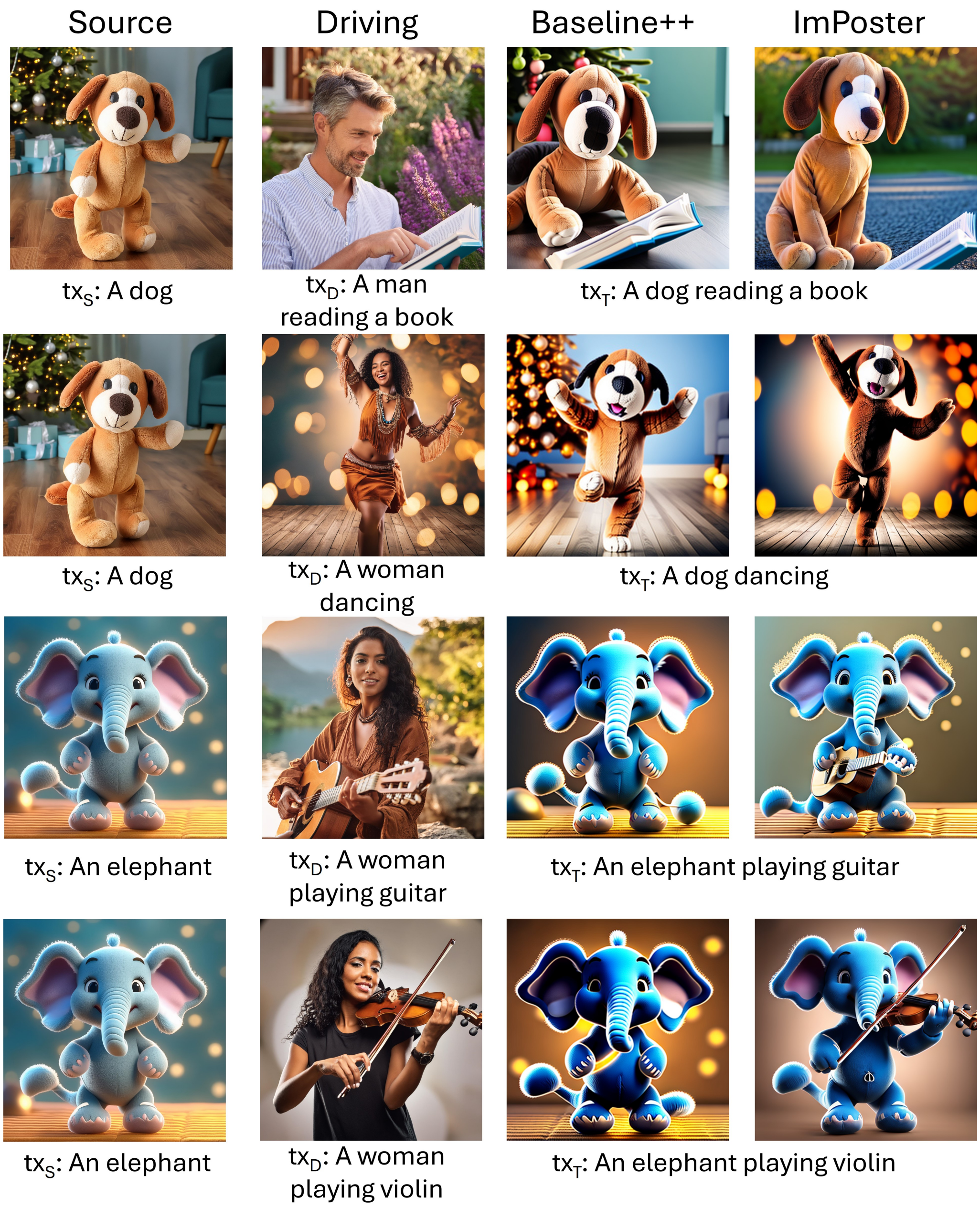

Figure 9: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.

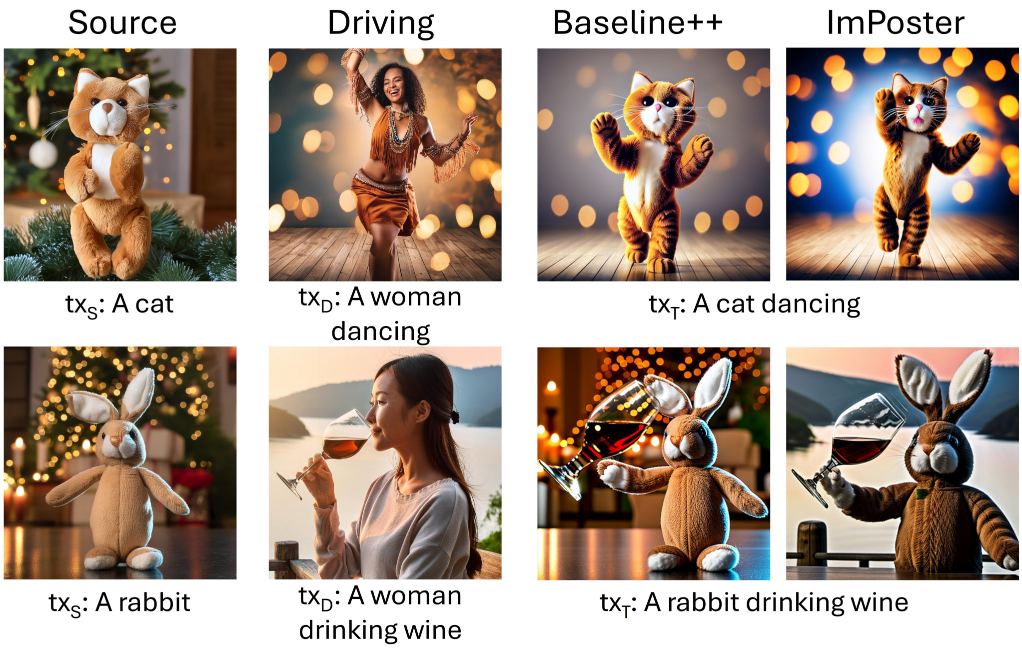

Figure 10: ImPoster is able to successfully transfer the driving action to a source subject, while maintaining its characteristics. In contrast, Baseline++ is unable to generate the driving action for the given source subject due to bias issues.