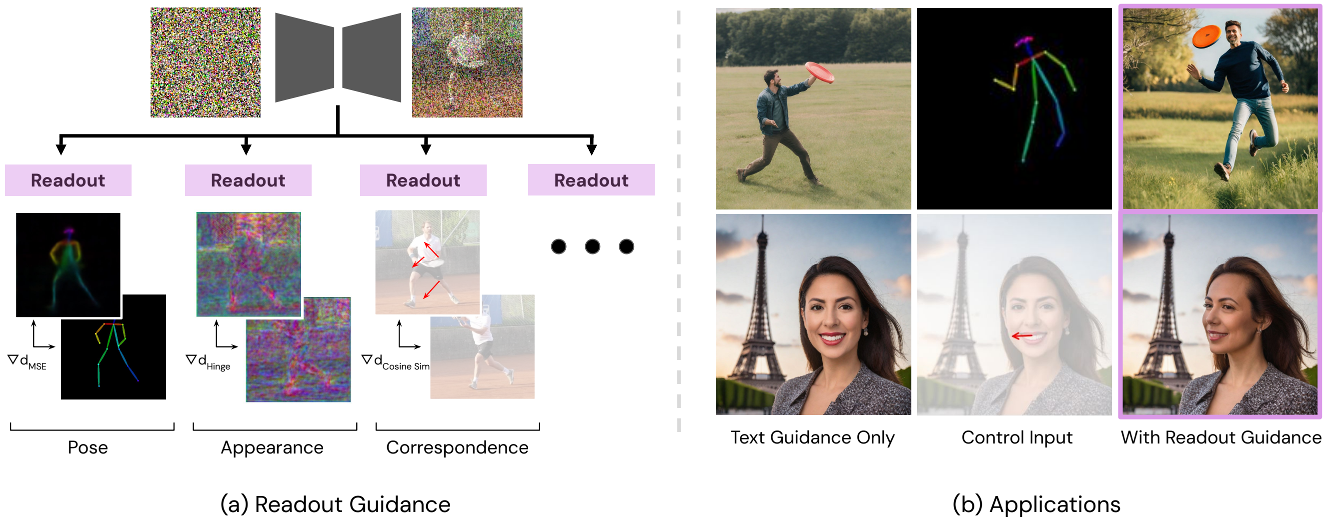
Readout Guidance: Learning Control from Diffusion Features

Grace Luo[1,2] Trevor Darrell[2] Oliver Wang[1] Dan B Goldman[1] Aleksander Holynski[1,2]

[1]Google Research [2]UC Berkeley



(a) Readout Guidance

(b) Applications

Figure 1. Given a frozen pre-trained text-to-image diffusion model [50], we learn parameter-efficient *readout heads* to interpret relevant signals, or *readouts*, from the intermediate network features. These readouts can be single-image concepts such as pose and depth, or relative concepts between two images, such as appearance similarity and correspondence. We use the readouts for sampling-time guidance to enable controlled image generation.

Abstract

We present Readout Guidance, a method for controlling text-to-image diffusion models with learned signals. Readout Guidance uses readout heads, *lightweight networks trained to extract signals from the features of a pre-trained, frozen diffusion model at every timestep. These readouts can encode single-image properties, such as pose, depth, and edges; or higher-order properties that relate multiple images, such as correspondence and appearance similarity. Furthermore, by comparing the readout estimates to a user-defined target, and back-propagating the gradient through the readout head, these estimates can be used to guide the sampling process. Compared to prior methods for conditional generation, Readout Guidance requires significantly fewer added parameters and training samples, and offers a convenient and simple recipe for reproducing different forms of conditional control under a single framework, with a single architecture and sampling procedure. We showcase these benefits in the applications of drag-based manipulation, identity-consistent generation, and spatially aligned control. Project page:* https://readout-guidance.github.io.

1. Introduction

Diffusion models have shown enormous potential in accurately modeling the space of natural images. However, one remaining open challenge that is critical to many applications is enabling arbitrary user control over their outputs. Existing solutions for enabling custom user control typically involve substantial model training on large annotated datasets, a process that is cumbersome and often infeasible for the average user. In this work, we provide an alternative solution for enabling user control that combines two ideas; first, that diffusion models contain rich internal representations that are useful for extracting relevant image properties, and second, that these extracted image properties can be used to guide the generation process towards desired user constraints. We call the combined approach Readout Guidance, because it makes use of small auxiliary readout heads that can be easily trained on top of a frozen diffusion model.

Taking as input the set of intermediate diffusion features, these readout heads can be trained to extract arbitrary properties about the image being generated (Figure 1, left). These can include image-space properties, such as human pose, depth maps, and edges; but can also be higher-order properties that relate two or more images, such as appearance similarity, correspondence, or shared identity. These readout heads consist of very few parameters, meaning that they can be trained on a single consumer GPU in a matter of hours—and since they bootstrap the already-rich diffusion features, they only need as few as 100 training examples. Beyond their utility in efficiently extracting properties about the generated images, the readout heads offer a useful mechanism for *guiding* the sampling process (Figure 1, right). At each step of the sampling process, properties can be extracted from the readout heads and compared to user-defined targets. In a similar fashion to classifier guidance [13], this comparison can be used as a guidance signal that encourages the generated image to match the target constraints.

We show that Readout Guidance can implement a number of popular forms of user control—all within the same simple guidance framework. In particular, we demonstrate state-of-the-art performance on the task of drag-based image manipulation, an application that has previously required bespoke architectural modifications and additional per-example fine-tuning. We also showcase our method on the task of identity-consistent image generation, in which outputs can be guided to contain the same person as a reference image. Finally, our method can also be used for spatially-aligned controls, such as depth-guided or pose-guided generation, as popularized by ControlNet [70] and T2IAdapter [40]. Notably, when compared to ControlNet [70], our method requires significantly less training data (as few as 100 supervised pairs vs. 200k), much less training time (a few hours vs. more than a week), and fewer added parameters (49MB vs. 1.4GB) [69].

2. Related Work

Conditional Diffusion Models. It has become increasingly popular to fine-tune text-to-image diffusion models to condition generation on signals beyond text, including camera pose [35], a reference identity [51], a reference image [6, 32, 38], a depth map [3], and more [23]. Due to the high cost of training diffusion models, many methods propose to keep the base model frozen and instead train an additional network that takes in the control signal and modulates the intermediate diffusion features accordingly. These types of models, including ControlNets [70], Adapters [40], and LoRAs [22], require less compute to train, fewer training samples, and can be combined in various ways because they build on the same frozen base model. Our method is orthogonal and complementary to this kind of adapter tuning, since our method aims to *guide* the sampling process based on the diffusion model's features, rather than modulate it, and can therefore be applied to any base model with any additional adapter. Our approach is more similar to classifier guidance, which is shown in Dhariwal & Nichol [13] to not only enable conditional generation from unconditional models, but also reinforce the capabilities of a conditional model. In our experiments, we similarly demonstrate that our method can be applied to these adapter-based models [40, 70] to further improve their control capabilities.

Sampling-Time Guidance. Because diffusion models synthesize images through an iterative sampling process rather than a single forward pass, one can guide the sampling process in a particular direction without modifying the base model. Such guidance was first achieved with gradients from an ImageNet [12] classifier trained on noisy images from the diffusion forward process [13]. Follow-up works explored alternative guidance approaches including the difference of score estimates between a conditional and unconditional model [21], off-the-shelf models that operate on the predicted clean image [4, 60], and hand-designed training-free functions that operate on diffusion features [10, 15, 33, 43]. In our work, we focus on *learning* this guidance function from diffusion features. Depth-Aware Guidance [29] first explored a similar idea for depth refinement, Sketch-Guided Diffusion [59] for edge guidance, and Mid-U Guidance [61] for aesthetic guidance. We expand this line of work into a more general form that can be used for non spatially-aligned applications such as drag-based manipulation.

Diffusion Model Representations. There exist works that demonstrate the rich and expressive nature of diffusion features beyond image synthesis, and how they can be readily applied to tasks like segmentation [5, 63, 66, 71], depth estimation [63, 71], human pose prediction [63], and semantic correspondence [20, 36, 55, 68]. These methods typically train additional decoders to extract an application-specific representation from the features of a particular, single diffusion timestep. To use these signals for guidance, we build on this work to instead extract an evolving prediction of these image properties at *each* step of the *entire* diffusion process.

Drag-based Manipulation. The concurrent works DragDiffusion [53] and DragonDiffusion [39] explore drag-based editing [42] with diffusion models. In this task, the user provides an image and a sparse set of point correspondences and the method produces an edited image that respects the specified appearance and deformation. These works enforce their appearance constraint using custom sampling-time processes which include single-image finetuning [53] or shared attention features [39] (following work in video stylization [7, 18, 28, 62]). The deformation constraints are enforced by encouraging high similarity between the corresponding points in the feature maps extracted from the diffusion model, as proposed in diffusion-based semantic correspondence methods [36, 55]. In contrast, our method learns the appearance similarity and correspondence constraints automatically from video data and applies both through the same sampling-time guidance procedure without any hand-designed operators or per-example fine-tuning.

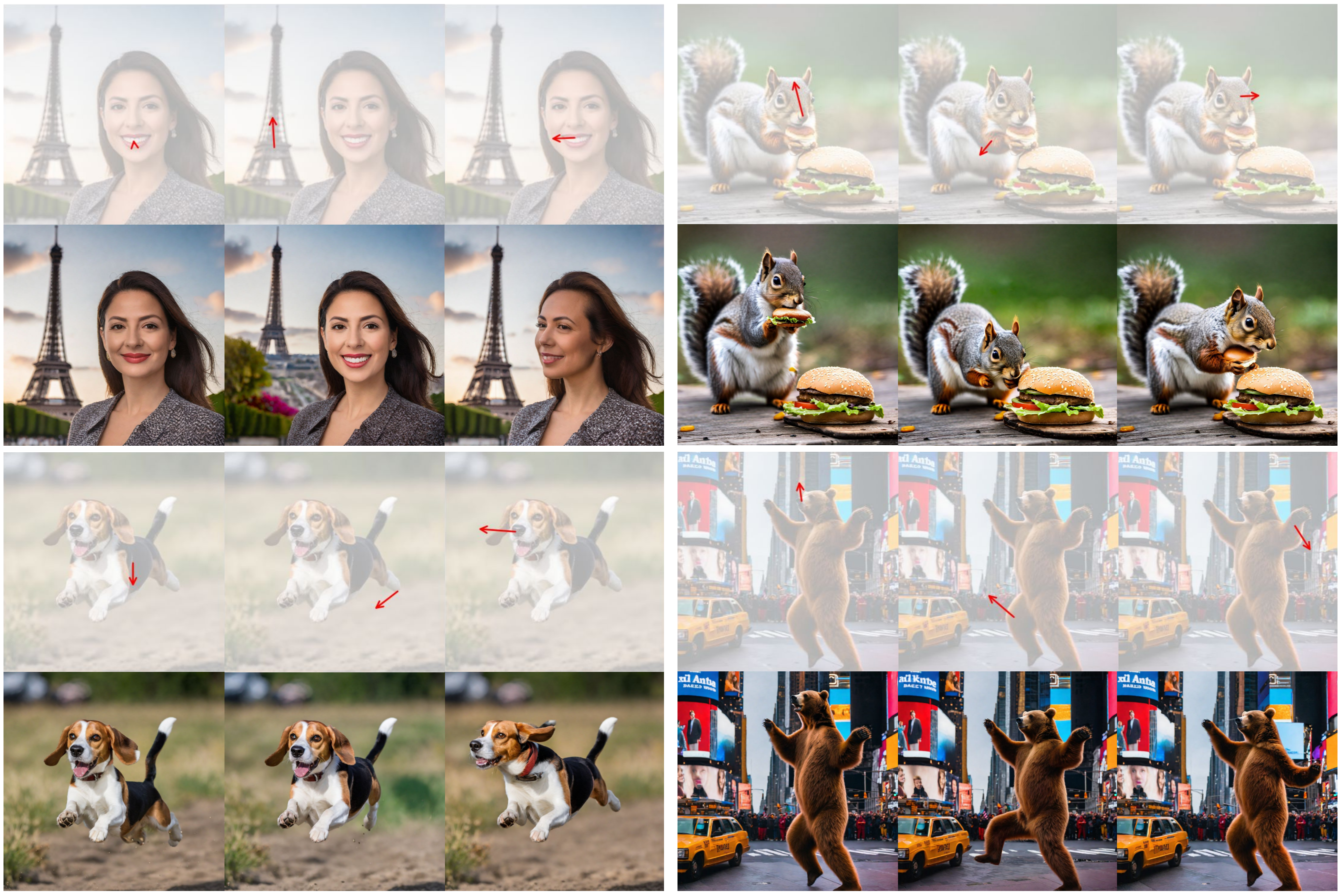


Figure 2. **Drag Based Manipulation (Generated Images)**: We show generated images with a single user correspondence constraint (with overlay) followed by the Readout Guidance generated result. Please see the Supplemental for the associated text prompts.

3. Readout Guidance

In this section, we describe the process of Readout Guidance: first describing the use of readout heads in the sampling process, then providing a general recipe for training a readout head on a custom dataset. In Sec. 4, we describe a number of example readout heads, showing that they can model spatially aligned properties like depth and human pose, as well as pairwise relative properties, such as correspondence and appearance similarity. We later showcase these readout heads in a number of conditional control applications, such as drag-based manipulation (Figure 2, 5), appearance preservation (Figure 6, 7), and spatially aligned control (Figure 4).

3.1. Background

Classifier guidance is a sampling procedure that enables conditional synthesis in an unconditional diffusion model. For the deterministic DDIM sampler [54], Dhariwal & Nichol [13] derive the update rule:

$$\hat{\epsilon}_t \leftarrow \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\nabla_{x_t}\log p_\phi(y|x_t)$$

where $\epsilon_\theta(\cdot)$ is the diffusion model, $p_\phi(y|\cdot)$ is the classifier, and $\sqrt{1-\bar{\alpha}_t}$ is a timestep-conditional scaling factor. One can think of this process as the classifier "nudging" the generation at each step such that it remains on the natural image manifold but moves towards to the classifier's criteria. In our method, we bootstrap pre-trained diffusion features from the frozen U-Net decoder to train a small readout head on top of the diffusion model. This allows us to design simple heads that are the same across entire classes of tasks (i.e., we use one architecture for all relative properties and another for all spatially-aligned properties) and only differ in the input data and loss.

3.2. A Revised Guidance Recipe

From Classification to Regression. In class and text guidance, the output domain y of the guidance function is often a fixed set of classes (e.g., object categories or a binary decision of "does the caption match" vs. "does the caption not match"). We replace classifiers with regressors that predict continuous values, where instead of maximizing a log probability, a distance function is minimized. As such, we write

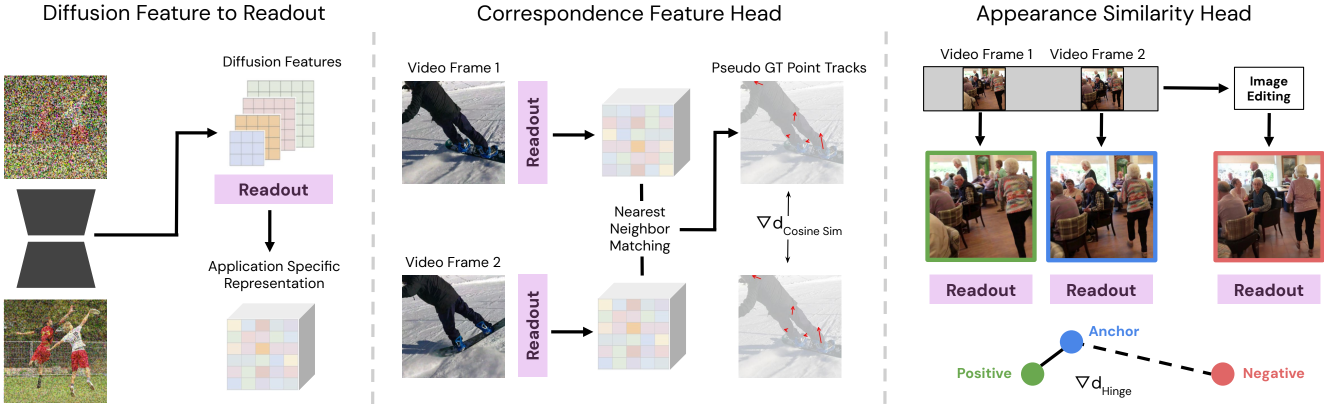


Figure 3. **Readout Head Training**: (left) Readout heads convert frozen diffusion features into representations useful for a diverse set of tasks, including predicting (middle) correspondences between a source and target image and (right) an appearance similarity feature between an anchor and positive / negative images.

our guidance function as a distance between the reference r and our predicted readout $\hat{r} = f_\psi(x_t)$.

$$\hat{\epsilon}_t \leftarrow \epsilon_\theta(x_t) + w \cdot \nabla_{x_t} d(r, f_\psi(x_t))$$

where $\epsilon_\theta(\cdot)$ is the diffusion model, w is the guidance weight, and $f_\psi(\cdot)$ is our learned readout head.

Relative Constraints. In the case of readout heads that relate multiple images, the reference is not a user input but rather derived from a separate diffusion process with a reference noisy image z_t:

$$\hat{\epsilon}_t \leftarrow \epsilon_\theta(x_t) + w \cdot \nabla_{x_t} d(f_\psi(z_t), f_\psi(x_t))$$

Our distance function $d(\cdot, \cdot)$ does not necessarily have to be computed between spatially aligned positions on the readouts – it can encode arbitrary spatial relationships, such as corresponding points or an object to be re-positioned. This flexibility in the loss formulation represents the generality of our method over ControlNet [70], which modulates features via a *per-pixel sum* between the control and diffusion model features, thereby making the method difficult to adapt to spatially unaligned constraints.

Combining Guidance Functions. One can also combine our readout guidance with other common forms of guidance, such as classifier-free text guidance [21]. It is also possible to guide with multiple readout heads $\psi_1, ..., \psi_n$ simultaneously. For example, we guide with both an appearance similarity and correspondence feature head for drag-based manipulation (Figure 2).

4. Readout Heads

4.1. Common Architecture

Our readout heads extend the architecture of Diffusion Hyperfeatures [36]. The heads extract features from the frozen decoder layers of the U-Net, reshape them with learned projection layers and bilinear upsampling, and then learn mixing weights to compute a weighted average of the layer features. Unlike fine-tuned conditional models [40, 70], our method does not require a large captioned dataset for training. We find that our method works well with extracted features from the unconditional branch, using the empty string “” as the prompt for all images. Given that we are also using these heads for guidance at sampling time, unlike the multi-timestep aggregation in Diffusion Hyperfeatures, we instead aggregate only over network layers. This allows us to query readout estimates at each sampling timestep, so that they can be used for guidance. Thus, we add timestep conditioning to the readout heads, and train them on features resulting from noisy images seen during the diffusion forward process. We define two types of heads, one for spatially-aligned properties, such as depth and pose, and one for “relative” properties, where the goal is to enforce semantics of one frame *relative* to another reference frame. For spatially-aligned properties, we add three convolutional layers to convert the feature map into an RGB readout, and for relative properties like correspondence and appearance similarity, we compute a distance metric directly on the feature map produced by the readout head. Figure 3 depicts the training procedure of our relative heads, and a more detailed architecture diagram can be found in the Supplemental.

4.2. Spatially Aligned Heads

Pose, Edge, Depth Head. We train our spatial heads on images from PascalVOC [16]. We use a similar setup to ControlNet [70], in which pseudo labels for the control are computed by off-the-shelf models and standardized into RGB images; we use OpenPose [8] to extract human pose, MiDaS [48, 49] for depth, and HED [65] for edges. For a given image, the spatial head extracts an aligned readout image. We then compute a per-pixel mean squared error loss against the ground truth pseudo label. For the pose and

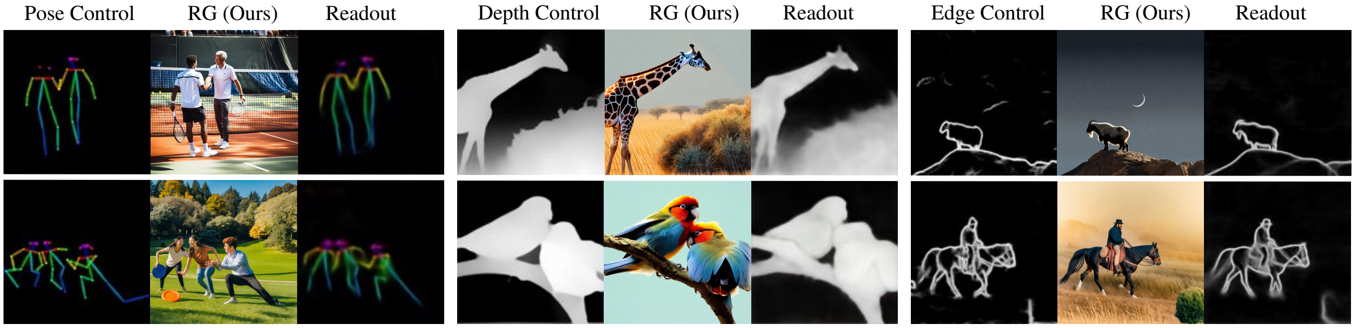


Figure 4. **Spatially Aligned Controls (Generated Images)**: In each example, we show the input user control, provided as a pose, depth map, or edge map derived from a different image (not shown), as well as our generated result, and a visualization of the readout head output for the generated image. We show more results in the Supplemental.

edge heads we re-weight the loss such that the contribution of the non-null and null regions are equal, as both controls are inherently sparse and include large portions of black background. In Figure 4 we depict example generations guided to match a user control by these heads as well as their associated readouts.

4.3. Relative Heads

Correspondence Feature Head. The goal of the correspondence feature head is to learn features that can be used to enforce correspondence constraints to the generated image. We train this readout head on video frames on DAVIS, following the approach of Diffusion Hyperfeatures [36], using image pairs with labeled point correspondences and training a network such that the feature distance between corresponding points is minimized, *i.e.*, the target point feature is the nearest neighbor for a given source point feature. We compute pseudo-labels using a point tracking algorithm [26] to track a grid of query points across the entire video. We randomly select two frames from the same video and a subset of the tracked points that are visible in both frames. For each image in the pair, the correspondence feature head constructs an associated feature map. We then use a contrastive loss – symmetric cross entropy [46] – to maximize the cosine similarity of matched source and target points.

Appearance Similarity Head. The goal of the appearance similarity head is to encourage the generation towards one where the appearance is similar to a reference image, but the image composition is unconstrained. That is, we want to penalize changes related to color, texture, and identity, but not changes related to object pose or camera angle. To do this, we define a loss based on *videos*, where we pick an anchor frame randomly selected from a video, and use a triplet loss to enforce that its features are more similar to those from another random frame from the *same video* than they are to an image with the *same structural layout* as the anchor, but with a *different* appearance. We achieve this by using an image editing method (SDEdit [14, 37]) that noises and denoises the anchor frame to produce variations that are similar in structure and semantic content, but have variably perturbed textural content. All three images (anchor, positive, negative) are processed independently by the appearance head, which constructs corresponding feature maps. For each pair of feature maps, we compute a per-pixel cosine distance and spatially pool these distances into a single scalar. We supervise the anchor to be closer in distance to the positive using a hinge loss with a margin of 0.5 [9, 17]. We train entirely on videos from DAVIS [45].

Identity Head. We also train a variant of the appearance similarity head with an identical architecture and training scheme but specific to preserving the identity of people. We train on tightly cropped images of faces from the CelebA-HQ dataset [27], where we use the same image editing technique [37] to produce hard negatives containing similar faces of different identities.

5. Results

Experimental Details. We implement readout heads for both Stable Diffusion v1-5 [50] (SDv1-5) and Stable Diffusion XL [44] (SDXL), such that we use the same model as the baseline for every comparison. We sample with image resolutions of 512 and 1024, resulting in latent resolutions of 64 and 128 respectively. Although these are both latent diffusion models [50], we find that our method produces high quality readouts and guidance operating directly on latent U-Net features, without the need for a latent decoder. In all our experiments, both when training our heads and sampling, we use a single Nvidia A100 40G GPU, and training a readout head takes at most three hours. We list additional information about the compute requirements and sampling hyperparameters for our method in the Supplemental.

Drag-Based Manipulation. Given a reference image, as well as a sparse set of points representing a desired deformation, the model is tasked with producing an edited image that respects both the input image appearance and the desired transformation. In Figure 5 we compare against the concurrent work DragDiffusion [53], which finetunes a

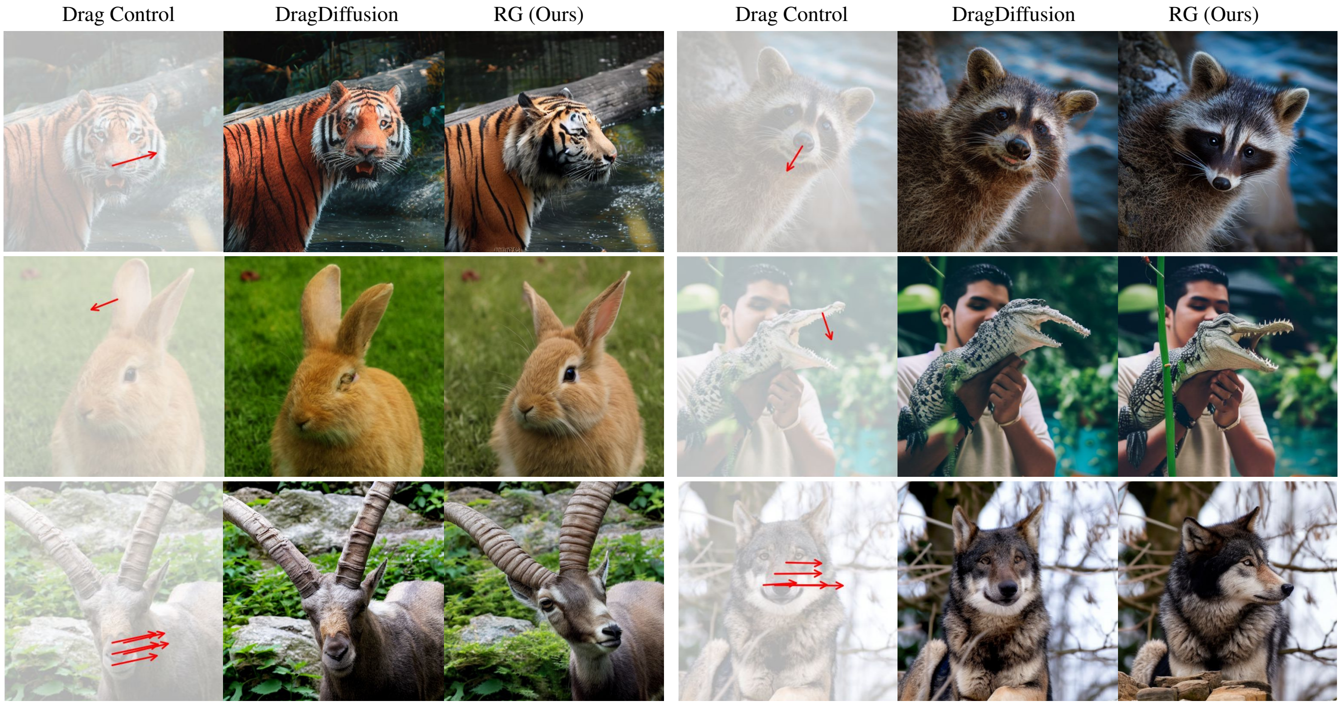


Figure 5. **Drag Based Manipulation (Real Images)**: The appearance similarity and correspondence feature head can operate on real images when seeding the reference features with those from DDIM inversion [54]. We compare against the concurrent work DragDiffusion [42]. Note that DragDiffusion requires an additional user input mask whereas our method does not.

LoRA [22] for each reference image and at sampling time optimizes such that the correspondences between the raw feature maps [55] match the points. For our method, we guide with our appearance similarity and correspondence feature heads across the entire diffusion process. This enables large out-of-plane motions, as seen in the first row where our method is able to fully rotate the tiger's head, whereas DragDiffusion slightly translates its nose upwards. These types of 3D transformations tend to be particularly challenging for prior work, likely because raw features encode some amount of orientation and perspective (e.g., if the nose is forward vs. side facing). We find that our correspondence feature head produces more orientation-agnostic features, likely due to our video training data, which allows us to correctly deform shapes. In addition, our appearance similarity guidance helps to keep the background consistent while editing the foreground object, which means that as opposed to DragDiffusion, our method does not require an additional input foreground mask.

Appearance Preservation. An important property of drag-based manipulation approaches is the ability to preserve subject identity or appearance. This task is related to existing work in model personalization [11, 15, 51, 52, 56], which typically requires model fine-tuning to learn the appearance of a given subject from multiple input observations. Our appearance heads can be used for a similar application, but without subject-specific fine-tuning on a reference image. Instead, we only apply guidance against a reference readout at sampling time, allowing our method to transform an image from a random noise seed to have consistent appearance with a reference image. In Figure 6 we ablate the effect of changing the guidance weight of our appearance similarity head for the same random seed. Since the notion of consistent appearance is somewhat ill-defined, by varying the weight of our guidance one can explore different definitions of this property, from only shared color and texture (columns 2, 3) to shared object identity and proportion (columns 4, 5).

Identity Consistency. The idea of appearance preservation can be further applied to specialized domains, for example people, where generative models often struggle to maintain identity across multiple generations. We showcase this potential application in Figure 7, where we show that our identity head can be used to encourage generated images to contain the same person as the reference image, effectively enabling the insertion of a consistent identity into different contexts. Additional details on this guidance procedure are provided in the Supplemental.

Spatially Aligned Control. We also demonstrate that our spatial heads can be used to handle commonly-used control signals, including pose, depth, and edge inputs. We take images from the unseen MSCOCO [34] validation set and extract their associated controls from pretrained models [8, 49, 65]. Figure 4 shows qualitative examples of our synthesized images and their associated readouts used during

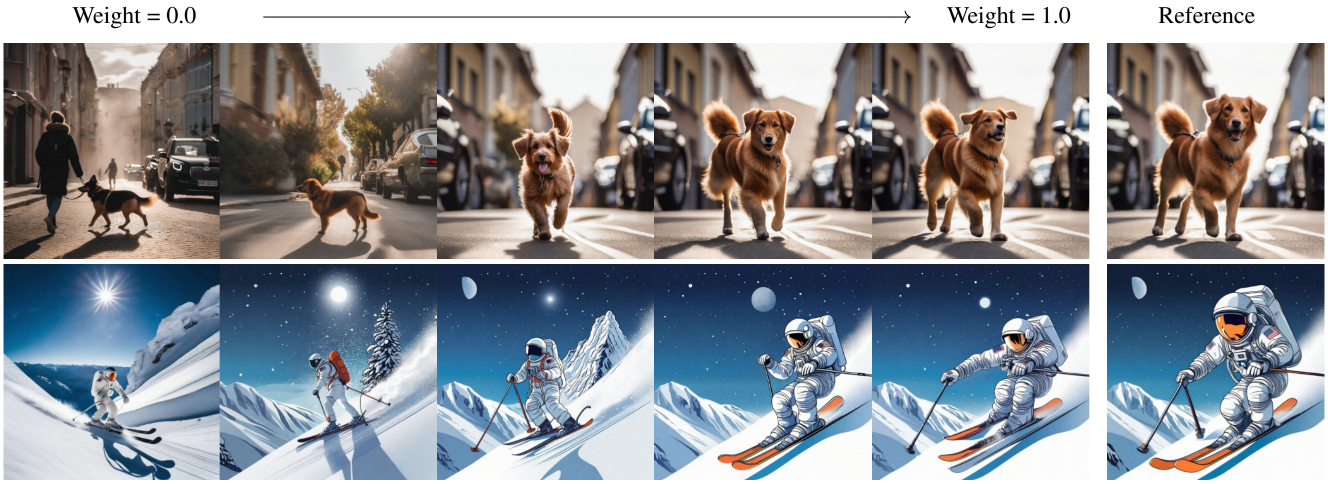


Figure 6. **Image Variations (Generated Images)**: The appearance similarity head can be used to create variations of a reference generated image. We show a random seed with no control (weight = 0.0) and visualize how the generation changes as we increase the strength of our Readout Guidance. Note that due to our definition of appearance similarity, identity is largely preserved while pose can change.

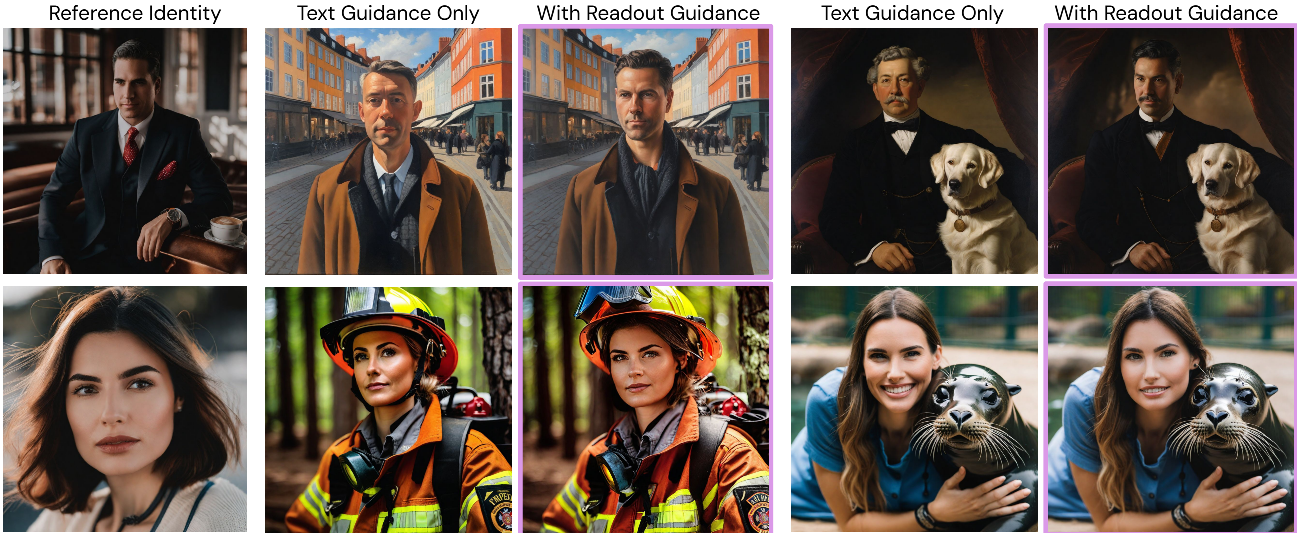


Figure 7. **Identity Guidance (Generated Images)**: Given a reference image (left), a specialized appearance similarity head can be used to guide a generated image with a different prompt to match the identity of the person in the reference image.

guidance. In Figure 8, we demonstrate that our guidance method is complementary to existing methods for conditional control. In fact, we can use the same guidance head trained on diffusion features from the vanilla base model on a model augmented with ControlNet or T2IAdapter. When combining approaches, we note that while ControlNet [70] and T2I-Adapter [40] work well for the task of pose control, they sometimes exhibit last-mile artifacts where the synthesized pose is slightly incorrect, and our guidance is able to refine these mistakes in those cases. Figure 9 also demonstrates that our method is competitive even when trained on limited data—as few as 100 training examples. We ablate training our pose readout head on 100, 1k, and all 8.5k images from PascalVOC [16] and demonstrate that all variants can handle challenging cases such as occlusions and multi-person control. Figure 10 provides a quantitative comparison of the input poses against OpenPose's [8] pose prediction of our generated image by computing the percentage of correct keypoints (PCK) below a certain error threshold. We normalize the PCK threshold by the size of the predicted pose, setting it to a constant scalar α multiplied by the size of the predicted pose bounding box. We observe that training on more data generally improves quality, where our best performing method is the one trained on all 8.5k images from PascalVOC [16]. Interestingly, as seen in Figure 10 our method trained on this relatively small dataset performs

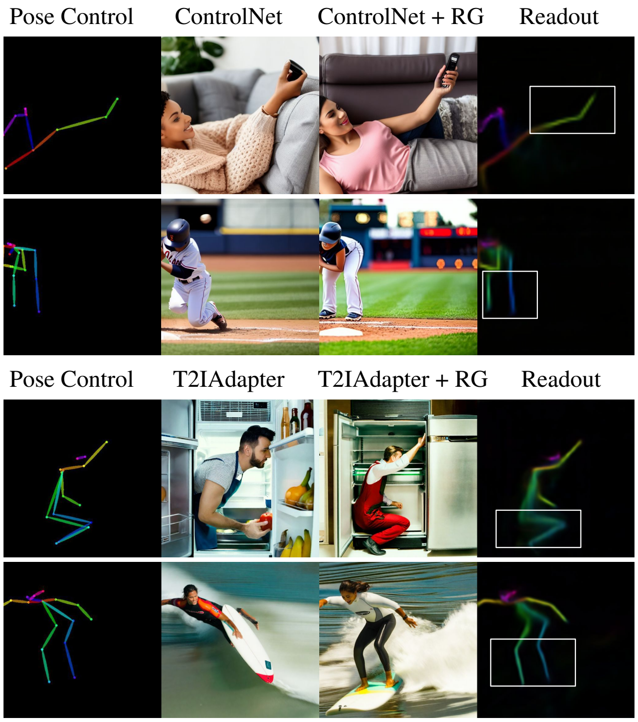


Figure 8. **Control Refinement**: Using Readout Guidance in combination with ControlNet [70] (top) and T2IAdapter [40] (bottom) can correct mistakes for difficult pose-guided generation cases shown here. We highlight these areas with overlaid white boxes.

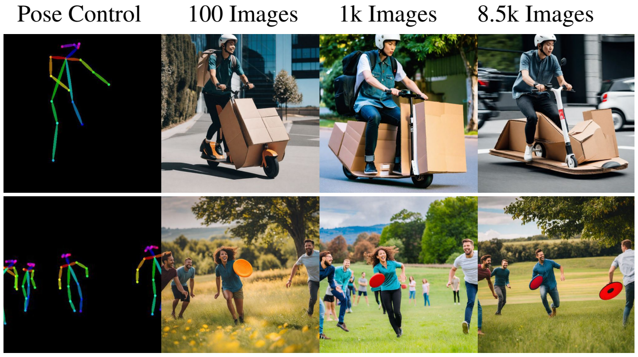


Figure 9. **Limited Data**: Readout heads are lightweight enough to apply in domains where there is little supervised data to train from. As few as 100 examples are needed to train a pose control model.

comparably to T2I-Adapter [40], which was trained on 3M images from LAION-Aesthetics V2 [57]. Moreover, combining T2I-Adapter + Readout Guidance significantly improves the performance of the base conditional model by 30% PCK ($\alpha = 0.05$), or 2.3x in PCK.

5.1. Limitations

Our method has a few limitations commonly associated with sampling-time guidance: (1) our method generally requires more memory and runtime during sampling to compute a gradient through the intermediate features via backpropagation, even when using the same number of sampling steps as the baselines; (2) our method can sometimes produce improbable images, in the form of cartoonish or unrealistic imagery satisfying the readout constraints. We provide additional discussion in the Supplemental.

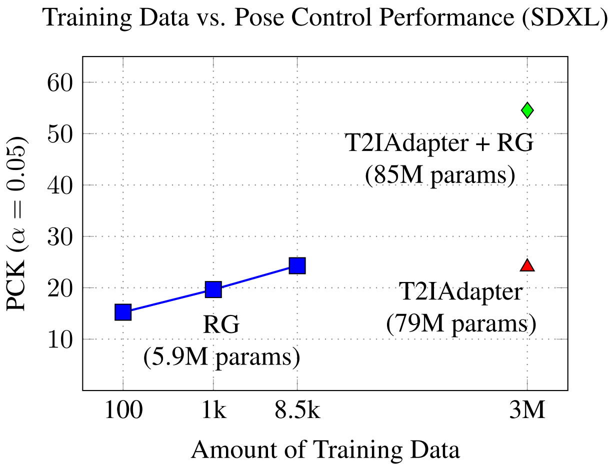


Figure 10. **Readout Guidance performs comparably to a conditional model [57] with 350x less training data**: We compute the percentage of correct keypoints (PCK) between an input pose and the pose of the synthesized image on 100 random images with humans from MSCOCO [34]. See additional results in the Supplemental.

6. Conclusion

We have introduced Readout Guidance, an approach to control the sampling of pre-trained diffusion models using readout heads trained on the diffusion features. Our method requires only small scale training data, making it easy to add to existing models. Because our method is a guidance technique, it can be used with both vanilla diffusion models as well as fine-tuned conditional models [40, 70]. We show that the general nature of Readout Guidance allows our approach to be used for a diverse set of controls, including point correspondences and appearance similarity.

7. Acknowledgements

We would like to thank Stephanie Fu, Eric Wallace, Yossi Gandelsman, Dave Epstein, Ben Poole, Thomas Iljic, Brent Yi, Kevin Black, Jessica Dai, Ethan Weber, Rundi Wu, Xiaojuan Wang, Luming Tang, Ruiqi Gao, Jason Baldridge, Zhengqi Li, and Angjoo Kanazawa for helpful discussions and feedback.

References

[1] https://www.pexels.com/video/an-alpine-ibex-with-horns-8217800/. 6

[2] https://www.pexels.com/video/wolf-looking-around-10727436/. 6

[3] Stability AI. Stable Diffusion Depth. https://huggingface.co/stabilityai/stable-diffusion-2-depth, 2022. 2

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal Guidance for Diffusion Models. In *CVPR*, pages 843–852, 2023. 2

[5] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*, 2022. 2

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*, pages 18392–18402, 2023. 2

[7] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *ICCV*, 2023. 2

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4, 6, 7, 5

[9] Gal Chechik, Varun Sharma, Uri Shalit, , and Samy Bengio. Large Scale Online Learning of Image Similarity Through Ranking. In *JMLR*, 2010. 5

[10] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-Free Layout Control with Cross-Attention Guidance. *arXiv preprint arXiv:2304.03373*, 2023. 2

[11] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven Text-to-Image Generation via Apprenticeship Learning. *arXiv preprint arXiv:2304.00186*, 2023. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255. Ieee, 2009. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, pages 8780–8794, 2021. 2, 3, 1

[14] Diffusers. Image-to-Image. https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/img2img, 2023. 5

[15] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion Self-Guidance for Controllable Image Generation. In *NeurIPS*, 2023. 2, 6

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 4, 7, 5

[17] Stephanie Fu*, Netanel Tamir*, Shobhita Sundaram*, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *NeurIPS*, 2023. 5

[18] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent Diffusion Features for Consistent Video Editing, 2023. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016. 2

[20] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised Semantic Correspondence Using Stable Diffusion. In *NeurIPS*, 2023. 2

[21] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 4, 1

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2022. 2, 6

[23] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and Controllable Image Synthesis with Composable Conditions. 2023. 2

[24] Huggingface. ResNetBlock2D. https://github.com/huggingface/diffusers/blob/7457aa67cb5c75132c38507080697b7cc7c4d9e6/src/diffusers/models/resnet.py#L746, 2023. 2

[25] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a Better Evaluation Metric for Image Generation. *arXiv:401.09603*, 2023. 4, 5

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is Better to Track Together. *arXiv:2307.07635*, 2023. 5

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 5, 3

[28] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In *ICCV*, 2023. 2

[29] Gyeongnyeon Kim, Wooseok Jang, Gyuseong Lee, Susung Hong, Junyoung Seo, and Seungryong Kim. DAG: Depth-Aware Guidance with Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv: Arxiv-2212.08861*, 2022. 2

[30] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. 2

[31] Tencent AI Lab. IP Adapter SDXL Plus Face Demo. https://github.com/tencent-ailab/IP-Adapter/blob/main/ip_adapter_sdxl_plus-face_demo.ipynb, 2023. 4

[32] Lambda Labs. Stable Diffusion Image Variations. https://huggingface.co/lambdalabs/sd-image-variations-diffusers, 2022. 2

[33] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655*, 2023. 2

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755. Springer, 2014. 6, 8, 4

[35] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot One Image to 3D Object. In *ICCV*, pages 9298–9309, 2023. 2

[36] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *NeurIPS*, 2023. 2, 4, 5

[37] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*, 2022. 5

[38] Runway ML. Stable Diffusion Inpainting. https://huggingface.co/runwayml/stable-diffusion-inpainting, 2022. 2

[39] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. DragonDiffusion: Enabling Drag-style Manipulation on Diffusion Models. *arXiv preprint arXiv:2307.02421*, 2023. 2

[40] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.08453*, 2023. 2, 4, 7, 8, 5, 15

[41] Intelligent Systems Lab Org. MidasNet. https://github.com/isl-org/MiDaS/blob/bdc4ed64c095e026dc0a2f17cabb14d58263decb/midas/midas_net.py#L37C11-L44C10, 2020. 3

[42] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 6

[43] Dong Huk Park*, Grace Luo*, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-Guided Diffusion with Inside-Outside Attention. In *WACV*, 2024. 2

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis, 2023. 5

[45] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *CoRR*, abs/1704.00675, 2017. 5

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 5

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 4

[48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. *ICCV*, 2021. 4

[49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 4, 6, 3

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, pages 10684–10695, 2022. 1, 5

[51] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *CVPR*, pages 22500–22510, 2023. 2, 6

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. HyperDreamBooth: HyperNetworks for Fast Personalization of Text-to-Image Models, 2023. 6

[53] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. *arXiv preprint arXiv:2306.14435*, 2023. 2, 5

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 3, 6, 1, 7

[55] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. In *NeurIPS*, 2023. 2, 6

[56] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. RealFill: Reference-Driven Generation for Authentic Image Completion. *arXiv preprint arXiv:2309.16668*, 2023. 6

[57] TencentARC. T2I-Adapter-SDXL - Openpose. https://huggingface.co/TencentARC/t2i-adapter-openpose-sdxl-1.0, 2023. 8

[58] timesler. facenet-pytorch. https://github.com/timesler/facenet-pytorch, 2019. 3

[59] Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. Sketch-Guided Text-to-Image Diffusion Models. In *SIGGRAPH*, 2023. 2

[60] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-End Diffusion Latent Optimization Improves Classifier Guidance. In *ICCV*, 2023. 2

[61] Jonathan Whitaker. Mid-U Guidance: Fast Classifier Guidance for Latent Diffusion Models. https://wandb.ai/johnowhitaker/midu-guidance/reports/Mid-U-Guidance-Fast-Classifier-Guidance-for-Latent-Diffusion-Models--VmlldzozMjg0NzA1#introduction-to-classifier-guidance, 2023. 2

[62] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-Shot Tuning of

Image Diffusion Models for Text-to-Video Generation. In *ICCV*, pages 7623–7633, 2023. 2

[63] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models. In *NeurIPS*, 2023. 2

[64] Jia Xiang and Gengming Zhu. Joint Face Detection and Facial Expression Recognition with MTCNN. In *ICISCE*, pages 424–427. IEEE, 2017. 3

[65] Saining Xie and Zhuowen Tu. Holistically-Nested Edge Detection. In *ICCV*, pages 1395–1403, 2015. 4, 6

[66] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *CVPR*, 2023. 2

[67] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. 2023. 4, 5

[68] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. In *NeurIPS*, 2023. 2

[69] Lvmin Zhang. Controlnet - Human Pose Version. https://huggingface.co/lllyasviel/sd-controlnet-openpose, 2023. 2

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 2, 4, 7, 8, 3, 5, 14

[71] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing Text-to-Image Diffusion Models for Visual Perception. In *ICCV*, 2023. 2