Text-image Alignment for Diffusion-based Perception

Neehar Kondapaneni[1*] Markus Marks[1*] Manuel Knott[1,2*]
Rogério Guimarães[1] Pietro Perona[1]

[1]California Institute of Technology
[2]ETH Zurich, Swiss Data Science Center, Empa

Abstract

Diffusion models are generative models with impressive text-to-image synthesis capabilities and have spurred a new wave of creative methods for classical machine learning tasks. However, the best way to harness the perceptual knowledge of these generative models for visual tasks is still an open question. Specifically, it is unclear how to use the prompting interface when applying diffusion backbones to vision tasks. We find that automatically generated captions can improve text-image alignment and significantly enhance a model's cross-attention maps, leading to better perceptual performance. Our approach improves upon the current SOTA in diffusion-based semantic segmentation on ADE20K and the current overall SOTA in depth estimation on NYUv2. Furthermore, our method generalizes to the cross-domain setting; we use model personalization and caption modifications to align our model to the target domain and find improvements over unaligned baselines. Our object detection model, trained on Pascal VOC, achieves SOTA results on Watercolor2K. Our segmentation method, trained on Cityscapes, achieves SOTA results on Dark Zurich-val and Nighttime Driving.

Project page: vision.caltech.edu/TADP/

1. Introduction

Diffusion models have set the state-of-the-art for image generation [30, 33, 36, 49]. Recently, a few works have shown diffusion pre-trained backbones have a strong prior for scene understanding that allows them to perform well in advanced discriminative vision tasks, such as semantic segmentation and monocular depth estimation [16, 50]. Unlike contrastive vision language models (like CLIP) [21, 25, 29], generative models have a causal relationship with text, in which text guides image generation. In latent diffusion models, text prompts control the denoising U-Net [34], moving the image latent in a semantically meaningful direction [5].

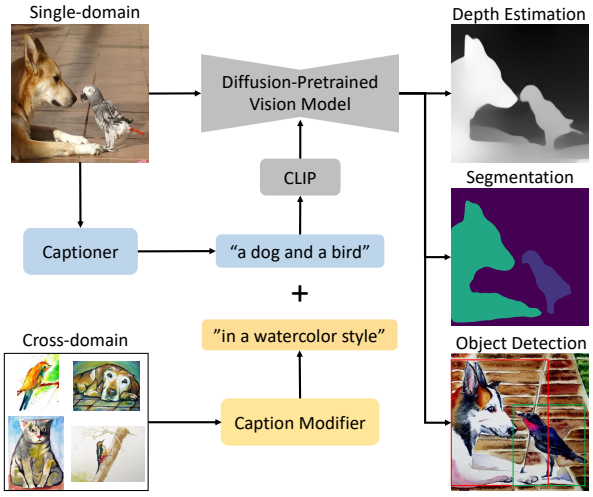


Figure 1. **Text-Aligned Diffusion Perception (TADP).** In TADP, image captions align the text prompts and images passed to diffusion-based vision models. In cross-domain tasks, target domain information is incorporated into the prompt to boost performance.

We explore this relationship and find that text-image alignment significantly improves the performance of diffusion-based perception. We then investigate text-target domain alignment in cross-domain vision tasks, finding that aligning the text with the target domain while training on the source domain can improve a model's target domain performance (Fig. 1).

We first study prompting for diffusion-based perceptual models and find that increasing text-image alignment improves semantic segmentation and depth estimation performance. We hypothesize that unaligned text prompts can introduce semantic shifts to the feature maps of the diffusion model [5] and that these shifts can make it more difficult for the task-specific head to solve the target task. Specifi-

*Equal contribution.

cally, we ask whether unaligned text prompts, such as averaging class-specific sentence embeddings ([29, 50]), hinder performance by interfering with feature maps through the cross-attention mechanism. Through ablation experiments on Pascal VOC2012 segmentation [13] and ADE20K [52], we find that off-target and missing class names degrade image segmentation quality. We find automated image captioning [24] achieves sufficient text-image alignment for perception. Our approach (along with latent representation scaling, see Sec. 4.1) improves performance for semantic segmentation on Pascal and ADE20k by 4.0 mIoU and 1.7 mIoU, respectively, and depth estimation on NYUv2 [40] by 0.2 RMSE (+8% relative) setting the new SOTA.

Next, we focus on cross-domain adaptation: can appropriate image captioning help visual perception when the model is trained in one domain and tested on a different domain? We find that training models on the source domain with the appropriate prompting strategy can lead to excellent unsupervised cross-domain performance on several benchmarks. We evaluate our cross-domain method on Pascal VOC [12, 13] to Watercolor2k (W2K) and Comic2k (C2K) [20] for object detection and Cityscapes (CS) [8] to Dark Zurich (DZ) [37] and Nighttime (ND) Driving [9] for semantic segmentation. We explore varying degrees of text-target domain alignment and find that improved alignment results in better performance. We also demonstrate using two diffusion personalization methods, Textual Inversion [15] and DreamBooth [35], for better target domain alignment and performance. We find that diffusion pre-training is sufficient to achieve SOTA (+5.8 mIoU on CS→DZ, +4.0 mIoU on CS→ND, +0.7 mIoU on VOC→W2k) or near SOTA results on all cross-domain datasets with no text-target domain alignment, and including our best text-target domain alignment method further improves +1.4 AP on Watercolor2k, +2.1 AP on Comic2k, and +3.3 mIoU on Nighttime Driving.

Overall, our contributions are as follows:

- We analyze the effects of text-image alignment on diffusion-pretrained vision models and use our insights to improve performance for semantic segmentation and depth estimation.
- We show that diffusion-based perception generalizes well across domains and that text-target domain alignment is important to improve performance.
- We show how to use diffusion personalization methods to close the domain gap between a source and target domain.

2. Related Work

2.1. Diffusion models for single-domain vision tasks

Diffusion models are trained to reverse a step-wise forward noising process. Once trained, they can generate highly realistic images from pure noise [30, 33, 36, 49]. To control image generation, diffusion models are trained with text prompts/captions that guide the diffusion process. These prompts are passed through a text encoder to generate text embeddings that are incorporated into the reverse diffusion process via cross-attention layers.

Recently, some works have explored using diffusion models for discriminative vision tasks. This can be done by either utilizing the diffusion model as a backbone for the task [16, 50] or through fine-tuning the diffusion model for a specific task and then using it to generate synthetic data for a downstream model [2, 47]. Our work falls into the first category; we use the diffusion model as a backbone for downstream vision tasks.

VPD [50] encodes images into latent representations and passes them through one step of the Stable Diffusion model. The cross-attention maps, multi-scale features, and output latent code are concatenated and passed to a task-specific head. Text prompts influence all these maps through the cross-attention mechanism, which guides the reverse diffusion process. The cross-attention maps are incorporated into the multi-scale feature maps and the output latent representation. The text guides the diffusion process and can accordingly shift the latent representation in semantic directions [1, 5, 15, 17]. The details of how VPD uses the prompting interface are described in Sec. 3. In short, VPD uses *unaligned* text prompts. In our work, we show how aligning the text to the image,by using a captioner, can significantly improve semantic segmentation and depth estimation performance.

2.2. Image captioning

CLIP [29] introduced a novel learning paradigm to align images with their captions. Shortly after, the LAION-5B dataset [39] was released with 5B image-text pairs; this dataset was used to train Stable Diffusion. We hypothesize that text-image alignment is important for diffusion-pretrained vision models. However, images used in advanced vision tasks (like segmentation and depth estimation) are not naturally paired with text captions. To obtain image-aligned captions, we use BLIP-2 [24], a model that inverts the CLIP latent space to generate captions for novel images.

2.3. Diffusion models for cross-domain vision tasks

A few works explore the cross-domain setting with diffusion models [2, 16]. Benigmim et al. [2] use a diffusion model to generate data for a downstream unsupervised domain adaptation (UDA) architecture. In [16], the diffusion backbone is frozen, and the segmentation head is trained with a consistency loss with category and scene prompts guiding the latent code towards target cross-domains. Similar to VPD, the category prompts consist of token embed-

dings for all classes present in the dataset, irrespective of their presence in any specific image. The consistency loss forces the model to predict the same output mask for all the different scene prompts, helping the segmentation head become invariant to the scene type. In contrast, our approach uses image captions to better align the text to the image and caption modifiers to further align the text to the target domain. We do not use a consistency loss. Instead, we train the diffusion model backbone and task head on the source domain data with and without incorporating the style of the target domain in the caption. We find that better alignment with the target domain (i.e. target domain information incorporated in the prompt) results in better cross-domain performance.

2.4. Cross-domain object detection

Cross-domain object detection can be divided into multiple subcategories, depending on what data / labels are at train / test time available. Unsupervised domain adaptation objection detection (UDAOD) tries to improve detection performance by training on unlabelled target domain data with approaches such as self-training [10, 41], adversarial distribution alignment [51] or generating pseudo labels for self-training [22]. Cross-domain weakly supervised object detection (CDWSOD) assumes the availability of image-level annotations at training time and utilizes pseudo labeling [20, 28], alignment [48] or correspondence mining [18]. Recently, [43] used CLIP [29] for Single Domain Generalization, which aims to generalize from a single domain to multiple unseen target domains. Our text-based method defines a new category of cross-domain object detection that tries to adapt from a single source to an unseen target domain by only having the broad semantic context of the target domain (e.g., foggy/night/comic/watercolor) as text input to our method. When we incorporate model personalization, our method can be considered a UDAOD method since we train a token based on unlabelled images from the target domain.

3. Methods

Stable Diffusion [33]. The text-to-image Stable Diffusion model is composed of four networks: an encoder $\mathcal{E}$, a conditional denoising autoencoder (a UNet in Stable Diffusion) ϵ_θ, a language encoder τ_θ (the CLIP text encoder in Stable Diffusion), and a decoder $\mathcal{D}$. $\mathcal{E}$ and $\mathcal{D}$ are trained before ϵ_θ, such that $\mathcal{D}(\mathcal{E}(x)) = \tilde{x} \approx x$. Training ϵ_θ is composed of a pre-defined forward process and a learned reverse process, the reverse process is learned using LAION-400M [38], a dataset of 400 million images ($x \in X$) and captions ($y \in Y$). In the forward process, an image x is encoded into a latent $z_0 = \mathcal{E}(x)$, and t steps of a forward noise process are executed to generate a noised latent z_t. Then, to learn the reverse process, the latent z_t is passed to the denoising autoencoder ϵ_θ, along with the time-step t and the image caption's representation $\mathcal{C} = \tau_\theta(y)$. τ_θ adds information about y to ϵ_θ using a cross-attention mechanism, in which the query is derived from the image, and the key and value are transformations of the caption representation. The model ϵ_θ is trained to predict the noise added to the latent in step t of the forward process:

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (1)$$

where $t \in \{0, ..., T\}$. During generation, a pure noise latent z_T and a user-specified prompt are passed through the denoising autoencoder ϵ_θ for T steps and decoded $\mathcal{D}(z_0)$ to generate an image-guided by the text prompt.

VPD [50]. uses ϵ_θ as a generatively pre-trained backbone for perceptual vision tasks. An image latent $z_0 = \mathcal{E}(x)$ and a text-derived conditioning $\mathcal{C}_{avg}$ are passed through the last step of the denoising process $\epsilon_\theta(z_0, 0, \mathcal{C})$. The cross-attention maps A, and the multi-scale feature maps F of the UNet are concatenated $V = A \oplus F$ and passed to a task-specific head H to generate a prediction $\hat{p} = H(V)$. The backbone ϵ_θ and head H are trained with a task-specific loss $\mathcal{L}_H(\hat{p}, p)$. In order to generate $\mathcal{C}_{avg}$ in VPD, a list of 80 sentence templates for each class of interest (such as "a <adjective> photo of a <class name>") are passed through the CLIP text encoder (similar to [29]). We use $\mathcal{B}$ to denote the set of class names in a dataset. For a specific class ($b \in \mathcal{B}$), the CLIP text encoder returns a $80 \times N \times D$ tensor, where N is the maximum number of tokens over all the templates and D is 768 (the dimension of each token embedding). Shorter sentences are padded with EOS tokens to fill out the max number of tokens. The first EOS token from each sentence template is averaged and used as the representative embedding for the class such that $\mathcal{C} \in \mathcal{R}^{|\mathcal{B}|\times 768}$. For semantic segmentation, all of the class embeddings, irrespective of presence in the image, are passed to the cross-attention layers. Only the class embedding of the room type is passed to the cross-attention layers for depth estimation.

3.1. Text-Aligned Diffusion Perception (TADP)

Our work proposes a novel method for prompting diffusion-pretrained perception models. Specifically, we explore different prompting methods $\mathcal{G}$ to generate $\mathcal{C}$. In the single-domain setting, we show the effectiveness of a method that uses BLIP-2 [24], an image captioning algorithm, to generate a caption as the conditioning for the model: $\mathcal{G}(x) = \tilde{y} \rightarrow \mathcal{C}$. We then extend our method to the cross-domain setting by incorporating target domain information to $\mathcal{C} = \mathcal{C} + \mathcal{M}(\mathcal{P})_s$, where $\mathcal{M}$ is a caption modifier that takes target domain information $\mathcal{P}$ as input and outputs a caption modification $\mathcal{M}(\mathcal{P})_s$ and a model modification $\mathcal{M}(\mathcal{P})_{\epsilon_\theta}$. In Sec. 4, we analyze the text-image interface of the diffusion model by varying the captioner $\mathcal{G}$ and caption modifier $\mathcal{M}$ in a systematic manner for three different

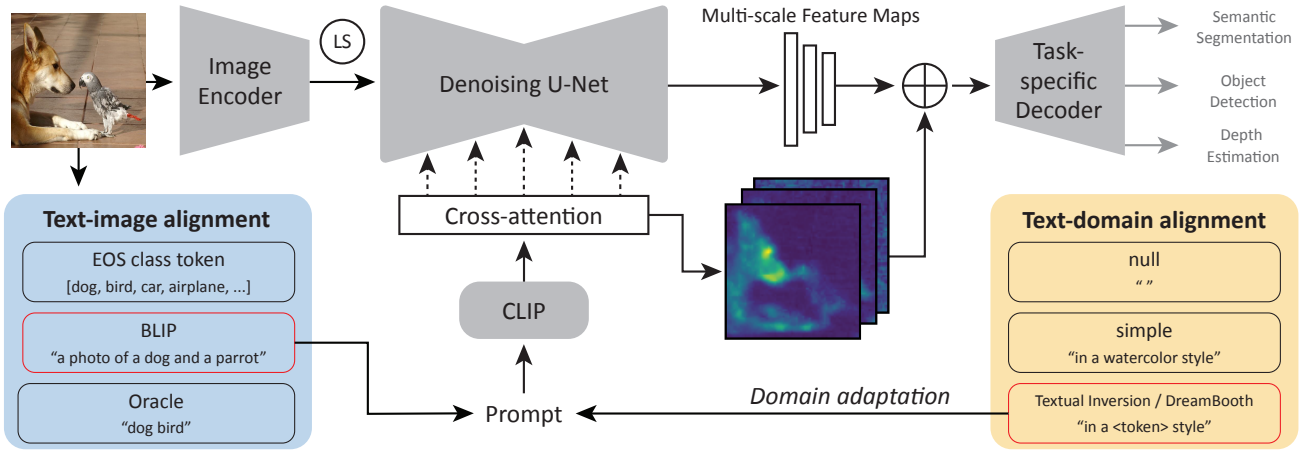


Figure 2. **Overview of TADP.** We encode images using the latent scaling (LS) described in Sec. 4.1. We experiment with several prompting strategies and evaluate their impact on downstream vision task performance. Our method concatenates the cross-attention maps and the multi-scale feature maps before passing them to the vision-specific decoder. In the blue box, we show three single-domain captioning strategies with differing levels of text-image alignment. We propose using BLIP [24] captioning to improve image-text alignment. We extend our analysis to the cross-domain setting (yellow box), exploring whether aligning the source domain text captions to the target domain may impact model performance by appending caption modifiers to image captions generated in the source domain and find model personalization modifiers (Textual Inversion/Dreambooth) work best.

Method	Avg	TA	LS	G	OT	mIoU^{ss}
VPD [50]	✓	✓			✓	82.34
VPD(LS)			✓		✓	83.06
Class Embs			✓		✓	82.72
Class Names			✓		✓	84.08
TADP-0			✓	✓		86.36
TADP-20			✓	✓		86.19
TADP-40			✓	✓		**87.11**
TADP(NO)-20			✓			86.35
TADP-Oracle			✓			89.85

Table 1. **Prompting for Pascal VOC2012 Segmentation.** We report the single scale validation mIoU for Pascal experiments. Avg: EOS token averaging, TA: Text Adapter, LS: Latent Scaling, G: Grammar, OT: Off-target information. For our method, we indicate the minimum length of the BLIP caption with TADP-X and nouns only with (NO).

vision tasks: semantic segmentation, object detection, and monocular depth estimation. Our method and experiments are presented in Fig. 2. For implementation details, refer to Sec. C.

4. Results

4.1. Latent scaling

Before exploring image-text alignment, we apply latent scaling to encoded images (Appendix G of Rombach et al. [33]). This normalizes the image latents to have a standard normal distribution. We pre-compute the scaling factor and fixate it to 0.18215. We find that latent scaling improves performance using VPD's original prompting scheme for segmentation and depth estimation (Fig. 3). Specifically, latent scaling improves ∼0.8% mIoU on Pascal, ∼0.3% mIoU on ADE20K, and a relative ∼5.5% RMSE on NYUv2 Depth (Fig. 3)

4.2. Single-domain alignment

Average EOS Tokens. We scrutinize the use of averaged EOS tokens for $\mathcal{C}$ (see Sec. 3). This method was introduced in CLIP [29] for generating a class-specific text representation for zero-shot classification and is also used in VPD to define representations for the classes in a dataset. We hypothesize the usage of the text prompting interface in this manner is mismatched with how diffusion models use text in image generation. While averaging is sensible when measuring cosine similarities in the CLIP latent space, it is unsuitable in diffusion models, where the text guides the diffusion process through cross-attention. In our qualitative analysis, we find that averaging degrades the cross-attention maps (Fig. 4). Instead of averaging, we first explore using CLIP to embed each class name independently and use the tokens corresponding to the actual word (not the EOS token) and pass this as input to the cross-attention layer:

$$\mathcal{G}_{\text{ClassEmbs}}(\mathcal{B}) = concat(\textit{CLIP}(b)|b \in \mathcal{B}) \rightarrow \mathcal{C}_{\text{ClassEmbs}} \quad (2)$$

Second, we explore a generic prompt, a string of class names separated by spaces:

$$\mathcal{G}_{\text{ClassNames}}(\mathcal{B}) = \{\text{`` ''} + b|b \in \mathcal{B}\} \rightarrow \mathcal{C}_{\text{ClassNames}} \quad (3)$$

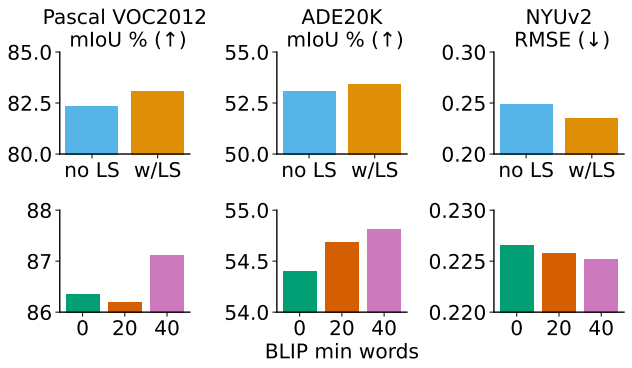


Figure 3. **Effects of Latent Scaling (LS) and BLIP caption minimum length.** We report mIoU for Pascal (left) and ADE20K (center) and RMSE for NYUv2 depth (right). (Top) Latent scaling improves performance on Pascal ∼0.8 mIoU (higher is better), ∼0.3 mIoU, and ∼5.5% relative RMSE (lower is better). (Bottom) We see a similar effect for BLIP minimum token length, with longer captions performing better, improving ∼0.8 mIoU on Pascal VOC2012, ∼0.9 mIoU on ADE20K, and ∼0.6% relative RMSE

These prompts are similar to the ones used for template-averaged embeddings w.r.t. overall text-image alignment but do not incorporate averaging. We evaluate these variations on Pascal VOC2012 segmentation. We find that $\mathcal{C}_{\text{ClassNames}}$ improves performance by 1.0 mIoU, but $\mathcal{C}_{\text{ClassEmbs}}$ reduces performance by 0.3 mIoU (see Tab. 1).

TADP. To align the diffusion model text input to the image, we use BLIP-2 [24] to generate captions for every image in our single-domain datasets (Pascal, ADE20K, and NYUv2).

$$\mathcal{G}_{\text{TADP}}(x) = \textit{BLIP-2}(x) \rightarrow \mathcal{C}_{\text{TADP}}(x) \tag{4}$$

BLIP-2 is trained to produce image-aligned text captions and is designed around the CLIP latent space. However, other vision-language algorithms that produce captions could also be used. We find that these text captions improve performance in all datasets and tasks (Tabs. 1, 2, 3). Performance improves on Pascal segmentation by ∼4% mIoU, ADE20K by ∼1.4% mIoU, and NYUv2 Depth by a relative RMSE improvement of 4%. We see stronger effects on the fast schedules for ADE20K with an improvement of ∼5 mIoU at (4k), ∼2.4 mIoU (8K). On NYUv2 Depth, we see a smaller gain on the fast schedule ∼2.4%. All numbers are reported relative to VPD with latent scaling.

We perform some ablations to analyze what aspects of the captions are important. We explore the minimum token number hyperparameter for BLIP-2 to explore if longer captions can produce more useful feature maps for the downstream task. We try a minimum token number of 0, 20, and 40 tokens (denoted as $\mathcal{C}_{\text{TADP-N}}$) and find small but consistent gains with longer captions, resulting on average 0.75% relative gain for 40 tokens vs. 0 tokens. (Fig. 3). Next, we ablate the Pascal $\mathcal{C}_{\text{TADP-20}}$ captions to understand what in the caption is necessary for the performance gains we observe. We use NLTK [4] to filter for the nouns in the captions. In the $\mathcal{C}_{\text{TADP(NO)-20}}$ nouns-only caption setting, we achieve 86.4% mIoU, similar to 86.2% mIoU with $\mathcal{C}_{\text{TADP-20}}$ (Tab. 1), suggesting nouns are sufficient.

Oracle. This insight about nouns leads us to ask if an oracle caption, in which all the object class names in an

Method	#Params	Crop	mIoU$^{\text{ss}}$	mIoU$^{\text{ms}}$
self-supervised pre-training				
EVA [14]	1.01B	896^2	61.2	61.5
InternImage [45]	1.08B	896^2	**62.5**	**62.9**
multi-modal pre-training				
CLIP-ViT-B [31]	105M	640^2	50.6	51.3
ViT-Adapter [7]	571M	896^2	61.2	61.5
BEiT-3 [46]	1.01B	896^2	**62.0**	62.8
ONE-PEACE [44]	1.52B	896^2	**62.0**	**63.0**
diffusion-based pre-training				
VPD$_{\text{A32}}$ [50]	862M	512^2	53.7	54.6
VPD(R)	862M	512^2	53.1	54.2
VPD(LS)	862M	512^2	53.7	54.4
TADP-40 (Ours)	862M	512^2	**54.8**	**55.9**
TADP-Oracle	862M	512^2	72.0	-

Table 2. **Semantic segmentation with different methods for ADE20k.** Our method (green) achieves SOTA within the diffusion-pretrained models category. The results of our oracle indicate the potential of diffusion-based models for future research as it is significantly higher than the overall SOTA (highlighted in yellow).

Method	RMSE↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	REL ↓	log10 ↓
default schedule						
SwinV2-L [26]	0.287	0.949	0.994	0.999	0.083	0.035
AiT [27]	0.275	0.954	0.994	0.999	0.076	0.033
ZoeDepth [3]	0.270	0.955	0.995	0.999	0.075	0.032
VPD [50]	0.254	0.964	0.995	0.999	0.069	0.030
VPD(R)	0.248	0.965	0.995	0.999	0.068	0.029
VPD(LS)	0.235	0.971	0.996	0.999	0.064	0.028
TADP-40	**0.225**	**0.976**	**0.997**	**0.999**	**0.062**	**0.027**
fast schedule, 1 epoch						
VPD	0.349	0.909	0.989	0.998	0.098	0.043
VPD(R)	0.340	0.910	0.987	0.997	0.100	0.042
VPD(LS)	0.332	0.926	0.992	0.998	0.097	0.041
TADP-0	**0.328**	0.935	**0.993**	**0.999**	0.082	0.038
TADP(TA)-0	0.332	**0.940**	0.993	**0.999**	**0.081**	**0.036**

Table 3. **Depth estimation in NYUv2.** We find latent scaling accounts for a relative gain of ∼ 5.5% on the RMSE metric. Additionally, image-text alignment improves ∼ 4% relative on the RMSE metric. A minimum caption length of 40 tokens performs the best.

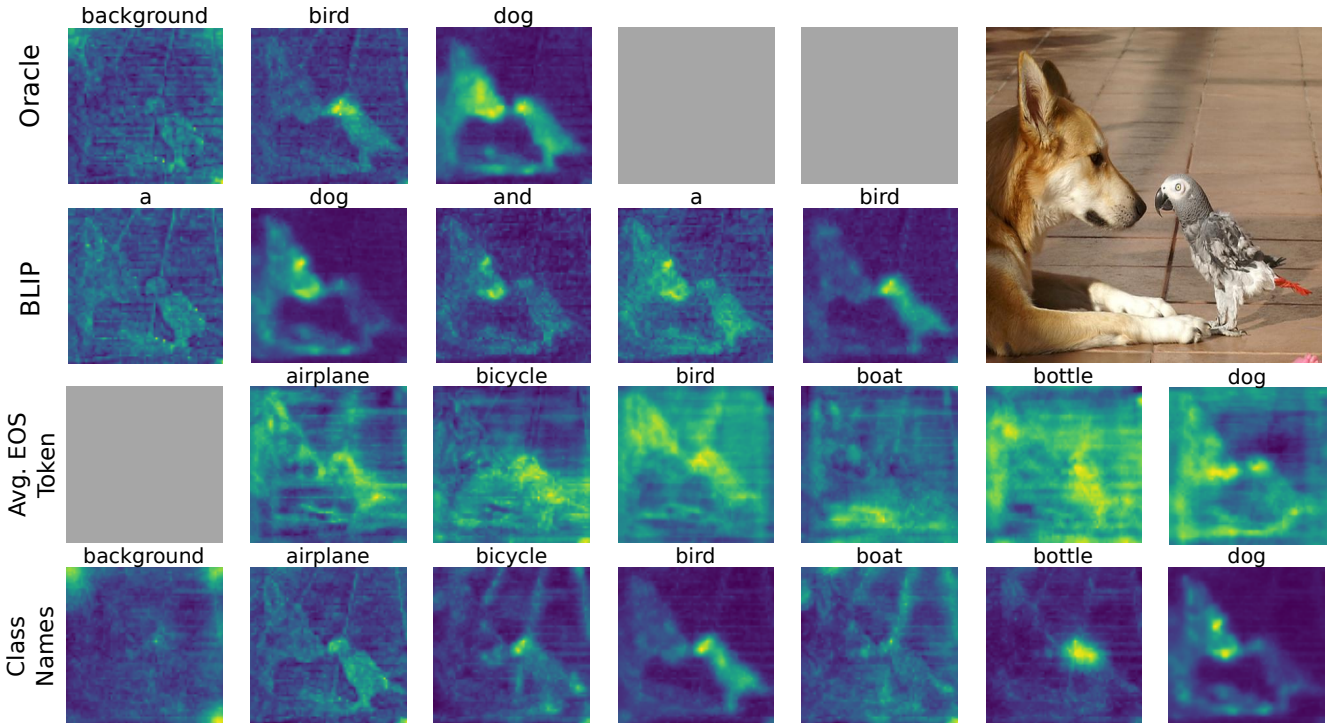


Figure 4. **Cross-attention maps for different types of prompting (before training).** We compare the cross-attention maps for four types of prompting: oracle, BLIP, averaged template EOS tokens, and class names as a space-separated string. The cross-attention maps for different heads at all different scales are upsampled to 64x64 and averaged. When comparing Averaged Template EOS and Class Names, we see (qualitatively) averaging degrades the quality of the cross-attention maps. Furthermore, we find that class names that are not present in the image can have highly localized attention maps (e.g., 'bottle').

image are provided as a caption, can improve performance further. We define $\mathcal{B}(x)$ as the set of class names present in image x.

$$\mathcal{G}_{\text{Oracle}}(x) = \{\text{`` '' } + b | b \in \mathcal{B}(x)\} \rightarrow \mathcal{C}_{\text{Oracle}}(x) \quad (5)$$

While this is not a realistic setting, it serves as an approximate upper bound on performance for our method on the segmentation task. We find a large improvement in performance in segmentation, achieving 89% mIoU on Pascal and 72.2% mIoU on ADE20K. For depth estimation, multi-class segmentation masks are only provided for a smaller subset of the images, so we cannot generate a comparable oracle. We perform ablations on the oracle captions to evaluate the model's sensitivity to alignment. For ADE20K, on the 4k iteration schedule, we modify the oracle captions by randomly adding and removing classes such that the recall and precision are at 0.5, 0.75, and 1.0 (independently) (Tab. S2). We find that both precision and recall have an effect, but recall is significantly more important. When recall is lower (0.50), improving precision has very little impact ($<$1% mIoU). However, as recall increases to 0.75 and 1.00, precision has progressively larger impacts ($\sim$3% mIoU and $\sim$7% mIoU). In contrast, recall has large impacts at every precision level: 0.5 - ($\sim$6% mIoU), 0.75 - ($\sim$9% mIoU), and 1.00 - ($\sim$13% mIoU). We find that BLIP-2 captioning performs similarly to a precision of 1.00 and a recall of 0.5 (Tab. 2).

4.3. Cross-domain alignment

Next, we ask if text-image alignment can benefit cross-domain tasks. In cross-domain, we train a model on a source domain and test it on a different target domain. There are two aspects of alignment in the cross-domain setting: the first is also present in single-domain, which is image-text alignment; the second is unique to the cross-domain setting, which is text-target domain alignment. The second is challenging because there is a large domain shift between the source and target domain. Our intuition is that while the model has no information on the target domain from the training images, an appropriate text prompt may carry some general information about the target domain. In our cross-domain experiments, we focus on the text-target domain alignment and use $\mathcal{G}_{\text{TADP}}$ for image-text alignment (following our insights from the single-domain setting).

Training. Our experiments in this setting are designed in the following manner: we train a diffusion model on

Method	Dark Zurich-val mIoU	ND mIoU
DAFormer [19]	–	54.1
Refign-DAFormer [6]	–	56.8
PTDiffSeg [16]	37.0	–
$\text{TADP}_{\text{null}}$	**42.8**	57.5
$\text{TADP}_{\text{simple}}$	39.1	56.9
$\text{TADP}_{\text{TextualInversion}}$	41.4	**60.8**
$\text{TADP}_{\text{DreamBooth}}$	38.9	60.4
$\text{TADP}_{\text{NearbyDomain}}$	41.9	56.9
$\text{TADP}_{\text{UnrelatedDomain}}$	42.3	55.1

Table 4. **Cross-domain semantic segmentation.** Cityscapes (CD) to Dark Zurich (DZ) val and Nighttime Driving (ND). We report the mIoU. Our method sets a new SOTA for DarkZurich and Nighttime Driving.

Method	Watercolor2k AP	Watercolor2k AP_{50}	Comic2k AP	Comic2k AP_{50}
Cross domain weakly supervised object detection				
PLGE [28]	–	56.5	–	41.7
ICCM [18]	–	57.4	–	37.1
H2FA R-CNN [48]	–	59.9	–	46.4
Unsupervised domain adaptation object detection				
ADDA [42]	–	49.8	–	23.8
MCAR [51]	–	56.0	–	33.5
UMT [10]	–	58.1	–	–
DASS-Detector (extra data) [41]	–	71.5	–	**64.2**
$\text{TADP}_{\text{null}}$	42.1	72.1	31.1	**57.4**
$\text{TADP}_{\text{simple}}$	**43.5**	**72.2**	31.9	56.6
$\text{TADP}_{\text{TextualInversion}}$	43.2	**72.2**	**33.2**	**57.4**
$\text{TADP}_{\text{DreamBooth}}$	43.2	**72.2**	32.9	56.9
$\text{TADP}_{\text{NearbyDomain}}$	42.0	71.5	31.8	56.4
$\text{TADP}_{\text{UnrelatedDomain}}$	42.2	71.9	32.0	55.9

Table 5. **Cross-domain object detection.** Pascal VOC to Watercolor2k and Comic2k. We report the AP and AP_{50}. Our method sets a new SOTA for Watercolor2K.

the source domain captions $\mathcal{C}_{\text{TADP}}(x)$. With these source domain captions, we experiment with four different caption modifications (each increasing in alignment to the target domain), a null $\mathcal{M}_{null}(\mathcal{P})$ caption modification where $\mathcal{M}_{null}(\mathcal{P})_s = \varnothing$ and $\mathcal{M}_{null}(\mathcal{P})_{\epsilon_\theta} = \varnothing$, a simple $\mathcal{M}_{simple}(\mathcal{P})$ caption modifier where $\mathcal{M}_{simple}(\mathcal{P})_s$ is a hand-crafted string describing the style of the target domain appended to the end and $\mathcal{M}_{simple}(\mathcal{P})_{\epsilon_\theta} = \varnothing$, a Textual Inversion [15] $\mathcal{M}_{TI}(\mathcal{P})$ caption modifier where the output $\mathcal{M}_{TI}(\mathcal{P})_s$ is a learned Textual Inversion token <*> and $\mathcal{M}_{TI}(\mathcal{P})_{\epsilon_\theta} = \varnothing$, and a DreamBooth [35] $\mathcal{M}_{DB}(\mathcal{P})$ caption modifier where $\mathcal{M}_{DB}(\mathcal{P})_s$ is a learned DreamBooth token <SKS> and $\mathcal{M}_{DB}(\mathcal{P})_{\epsilon_\theta}$ is a DreamBoothed diffusion backbone. We also include two additional control experiments. In the first, $\mathcal{M}_{ud}(\mathcal{P})$ an *unrelated* target domain style is appended to the end of the string. In the second, $\mathcal{M}_{nd}(\mathcal{P})$ a *nearby* but a different target domain style is appended to the caption. $\mathcal{M}_{TI}(\mathcal{P})$ and $\mathcal{M}_{DB}(\mathcal{P})$ require more information than the other methods, such that $\mathcal{P}$ represents a subset of unlabelled images from the target domain.

Testing. When testing the trained models on the target domain images, we want to use the same captioning modification for the test images as in the training setup. However, $\mathcal{G}_{\text{TADP}}$ introduces a confound since it naturally incorporates target domain information. For example, $\mathcal{G}_{\text{TADP}}(x)$ might produce the caption "a watercolor painting of a dog and a bird" for an image from the Watercolor20K dataset. Using the $\mathcal{M}_{simple}(\mathcal{P})$ captioning modification on this prompt would introduce redundant information and would not match the caption format used during training. In order to remove target domain information and get a plain caption that can be modified in the same manner as in the training data, we use GPT-3.5 to remove all mentions of the target domain shift. For example, after using GPT-3.5 to remove mentions of the watercolor style in the above sentence, we are left with "an image of a bird and a dog". With these *GPT-3.5 cleaned captions*, we can match the caption modifications used during training when evaluating test images. This caption-cleaning strategy allows us to control how target domain information is included in the test image captions.

4.3.1 Evaluation

We evaluate cross-domain transfer on several datasets. We train our model on Pascal VOC [12, 13] object detection and evaluate on Watercolor2K (W2K) [20] and Comic2K (C2K) [20]. We also train our model on the Cityscapes [8] dataset and evaluate on the Nighttime Driving (ND) [9] and Dark Zurich-val (DZ-val) [37] datasets. We show all results in Tab. 5.

Null caption modifier. The null captions have no target domain information. In this setting, the model is trained with captions with no target domain information and tested with GPT-3.5 cleaned target domain captions. We find diffusion pre-training to be extraordinarily powerful on its own, with just plain captions (no target domain information); the model already achieves SOTA on VOC→W2K with 72.1 AP_{50}, SOTA on CD→DZ-val with 42.8 mIoU and SOTA on CD→ND with 60.8 mIoU. AP. Our model performs better than the current SOTA [41] on VOC→W2K and worse on VOC→C2K (highlighted in yellow in Tab. 5). However, [41] uses a large extra training dataset from the target (comic) domain, so we highlight in bold our results

in Tab. 5 to show they outperform all other methods that use only images in C2K as examples from the target domain. Furthermore, these results are with a lightweight FPN [23] head, in contrast to other competitive methods like Refign [6], which uses a heavier decoder head. We use the plain captions as our baseline for the following sections; the deltas are given relative to the numbers in this setting.

Simple caption modifier. We then add target domain information to our captions by prepending the target domain's semantic shift to the generic captions. These caption modifiers are hand-crafted, for example, "a dog and a bird" becomes "a X style painting of a dog and a bird" (where X is watercolor for W2K and comic for C2K) and "a dark night photo of a dog and a bird" for DZ. We achieve 72.2 (+0.1) AP_{50} on W2K, 56.6 (-0.8) AP_{50} on C2K, 39.1 (-3.7) mIoU on DZ-val, and 56.9 (-0.6) mIoU on ND.

Textual Inversion caption modifier. Textual inversion [15] is a method that learns a target concept (an object or style) from a set of images and encodes it into the embedding of a token. We learn a novel token from target domain image samples to further increase image-text alignment (for details, see Sec. C.1). In this setting, the sentence template becomes "a <token> style painting of a dog and a bird". Our performance improves to 60.8 (+3.3) mIoU on ND and 72.2 (+0.1) AP_{50} on W2K. Our method performs worse on DZ-val 41.4 (-1.4) mIoU, and has no impact on C2K 57.4 (+0.0) AP_{50}. However, on W2K and C2K, our method improves under the AP metric to 43.2 (1.1) AP and 33.2 (+2.1) AP.

DreamBooth caption modifier. DreamBooth-ing [35] is a more compute-intensive method for achieving the same goal as textual inversion. Along with learning a new token, the stable-diffusion backbone itself is fine-tuned with a set of target domain images (for details, see Sec. C.1). We swap the stable diffusion backbone with the DreamBoothed backbone before training. We use the same template as in textual inversion. On W2K, we improve performance to 72.2 (+0.1) AP_{50} and 43.2 (+1.1) AP. On C2K, our performance is 56.9 (-0.6) AP_{50} and 32.9 (+1.8) AP. On DZ-val and ND we achieve 38.9 (-3.9) mIoU and 60.4 (+2.9) mIoU on ND.

Ablations We ablate our target domain alignment strategy by introducing unrelated and nearby target-domain style modifications. For example, this would be "a **dashcam** photo of a dog and a bird" (unrelated) and "a **constructivism** painting of a dog and a bird" (nearby) for the W2K and C2K datasets. "A **watercolor** painting of a car on the street" (unrelated) and "a **foggy** photo of a car on the street" for the ND and DZ-val datasets. We find these off-target domains reduce performance on all datasets. On C2K (AP_{50}), the nearby and unrelated domain modifier reduced performance to 56.4 (-1.0) and 55.9 (-1.5) AP_{50}. On W2K (AP_{50}), performance decreased to 71.5 (-0.6) and 71.9 (-0.2). On DZ-val, performance falls to 41.9 (-0.9) and 42.33 (-0.5) mIoU. Finally, on ND, performance falls to 56.9 (-0.6) and 55.1 (-2.4) mIoU.

5. Discussion

We present the first systematic exploration of the impact of image-text alignment on diffusion-based perception tasks. The method we propose for image-text alignment is general, fully automated, and can be applied to any diffusion-based perception model. We investigate whether similar principles apply in the cross-domain setting and find that alignment towards the target domain during training improves downstream cross-domain performance. In this work, we develop several insights into the nature of the text-image interface in diffusion models. The first finding is that EOS token averaging, a method borrowed from prior work related to CLIP, does not work as effectively with diffusion models. In addition, our oracle ablation experiments indicate that our diffusion pre-trained segmentation model is particularly sensitive to missing classes and less sensitive to off-target classes (reduced precision), and both have negative impacts. Our proposed method shows how using a captioner (BLIP-2), which has the benefit of being open vocabulary, high precision, and downstream task agnostic, improves performance significantly. Future work may explore closed vocabulary captioners that are more task-specific in order to get closer to oracle-level performance. Our results show that aligning text prompts to the image is important in identifying/generating good multi-scale feature maps for the downstream segmentation head. In addition, it implies that the multi-scale features and latent representations do not naturally identify semantic concepts without the guidance of the text in diffusion models. We also find that diffusion models can be used effectively for cross-domain tasks. Because of its strong generalization performance, our model performs well across domains using the null captioner. On average, we find that target domain alignment can help with cross-domain performance and misalignment leads to worse performance. Capturing information about the target domain in words alone can be difficult. For these cases we show, that model personalization through Textual Inversion or Dreambooth can bridge the gap without requiring labeled data. Future work could explore how to expand our framework to generalize to multiple unseen domains. Our work analyzes and successfully exploits the text-image interface in diffusion pre-trained models for the downstream vision tasks of semantic segmentation, monocular depth estimation, and object detection.

Acknowledgements

This work was supported by an ETH Zurich Doc.Mobility Fellowship. We thank Oisin Mac Aodha, Yisong Yue, and Mathieu Salzmann for their valuable inputs that helped improve this work.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models. *arXiv preprint arXiv:2303.18080*, 2023. 2

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *arXiv preprint arXiv:2302.12288*, 2023. 5

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009. 5

[5] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing Diffusion using Semantic Dimensions. *arXiv preprint arXiv:2301.12247*, 2023. 1, 2

[6] David Brüggemann, Christos Sakaridis, Prune Truong, and Luc Van Gool. Refign: Align and Refine for Adaptation of Semantic Segmentation to Adverse Conditions. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 7, 8

[7] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Y. Qiao. Vision Transformer Adapter for Dense Predictions. *arXiv preprint arXiv:2205.08534*, 2022. 5

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 2, 7

[9] Dengxin Dai and Luc Van Gool. Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824, 2018. 2, 7

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 3, 7

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014. 8

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2, 7

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012), 2012. 2, 7

[14] Yuxin Fang, Wen Wang, Binhui Xie, Quan-Sen Sun, Ledell Yu Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2022. 5

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 7, 8

[16] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting Diffusion Representations for Cross-Domain Semantic Segmentation. *arXiv preprint arXiv:2307.02138*, 2023. 1, 2, 7

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[18] Luwei Hou, Yu Zhang, Kui Fu, and Jia Li. Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9929–9938, 2021. 3, 7

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9914–9925, 2022. 7

[20] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5009, 2018. 2, 3, 7

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1

[22] Junguang Jiang, Baixu Chen, Jianmin Wang, and Mingsheng Long. Decoupled adaptation for cross-domain object detection. *arXiv preprint arXiv:2110.02578*, 2021. 3

[23] Alexander Kirillov, Ross B. Girshick, Kaiming He, and Piotr Dollár. Panoptic Feature Pyramid Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, 2019. 8

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 4, 5

[25] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. *arXiv preprint arXiv:2110.05208*, 2022. 1

[26] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11999–12009, 2021. 5

[27] Jia Ning, Chen Li, Zheng Zhang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in Tokens: Unifying Output Space of Visual Tasks via Soft Token. *arXiv preprint arXiv:2301.02229*, 2023. 5

[28] Shengxiong Ouyang, Xinglu Wang, Kejie Lyu, and Yingming Li. Pseudo-label generation-evaluation framework for cross domain weakly supervised object detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 724–728. IEEE, 2021. 3, 7

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya 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. *International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2, 3, 4

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[31] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-Guided Dense Prediction with Context-Aware Prompting. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18061–18070, 2022. 5

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 8

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 1, 2, 3, 4

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, Cham, 2015. 1

[35] 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. *arXiv preprint arXiv:2208.12242*, 2022. 2, 7, 8

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[37] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided Curriculum Model Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7373–7382, 2019. 2, 7

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. *European Conference on Computer Vision (ECCV)*, 2012. 2

[41] Barış Batuhan Topal, Deniz Yuret, and Tevfik Metin Sezgin. Domain-adaptive self-supervised pre-training for face & body detection in drawings. *arXiv preprint arXiv:2211.10641*, 2022. 3, 7

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 7

[43] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. CLIP the Gap: A Single Domain Generalization Approach for Object Detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3219–3229, 2023. 3

[44] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. *arXiv preprint arXiv:2305.11172*, 2023. 5

[45] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiao-hua Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Y. Qiao. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14408–14419, 2022. 5

[46] Wen Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiangbo Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. *2023 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 19175–19186, 2023. 5

[47] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models. *arXiv preprint arXiv:2303.11681*, 2023. 2

[48] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2022. 3, 7

[49] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 2

[50] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing Text-to-Image Diffusion Models for Visual Perception. *arXiv preprint arXiv:2303.02153*, 2023. 1, 2, 3, 4, 5, 8

[51] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 54–69. Springer, 2020. 3, 7

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 2

Text-image Alignment for Diffusion-based Perception Supplementary Materials

A. Off-target class name analysis

Qualitative image-to-image variation analysis. We present a qualitative and quantitative analysis on the effect of off-target class names added to the prompt. In Fig. S1, we use the stable diffusion image to image (img2img) variation pipeline (with the original Stable Diffusion 1.5 weights) to qualitatively analyze the effects of prompts with off-target classes. The img2img variation pipeline encodes a real image into a latent representation, adds a user-specified amount of noise to the latent representation, and de-noises it (according to a user-specified prompt) to generate a variation on the original image. The amount of noise added is dictated by a strength ratio indicating how much variation should occur. A higher ratio results in more added noise and more denoising steps, allowing a relatively higher impact of the new text prompt on the image. We find that $\mathcal{C}_{\text{ClassNames}}$ (see caption for details) results in variations that incorporate the off-target classes. This effect is most clear looking across the panels left to right in which objects belonging to off-target classes (an airplane and a train) become more prominent. These qualitative results imply that this kind of prompt modifies the latent representation to incorporate information about off-target classes, potentially making the downstream task more difficult. In contrast, using the BLIP prompt changes the image, but the semantics (position of objects, classes present) of the image variation are significantly closer to the original. These results suggest a mechanism for how off-target classes may impact our vision models. We quantitatively measure this effect using a fully trained Oracle model in the following section.

Quantitative effect of $\mathcal{C}_{\text{ClassNames}}$ on Oracle model. To quantify the impact of the off-target classes on the downstream vision task, we measure the averaged pixel-wise scores (normalized via Softmax) per class when passing the $\mathcal{C}_{\text{ClassNames}}$ to the Oracle segmentation model for Pascal VOC 2012 (Fig. S2). We compare this to the original oracle prompt. We find that including the off-target prompts significantly increases the probability of a pixel being misclassified as one of the semantically nearby off-target classes. For example, if the original image contains a cow, including the words dog and sheep, it significantly raises the probability of misclassifying the pixels belonging to the cow as pixels belonging to a dog or a sheep. These results indicate that the task-specific head picks up the effect of off-target classes and is incorporated into the output.

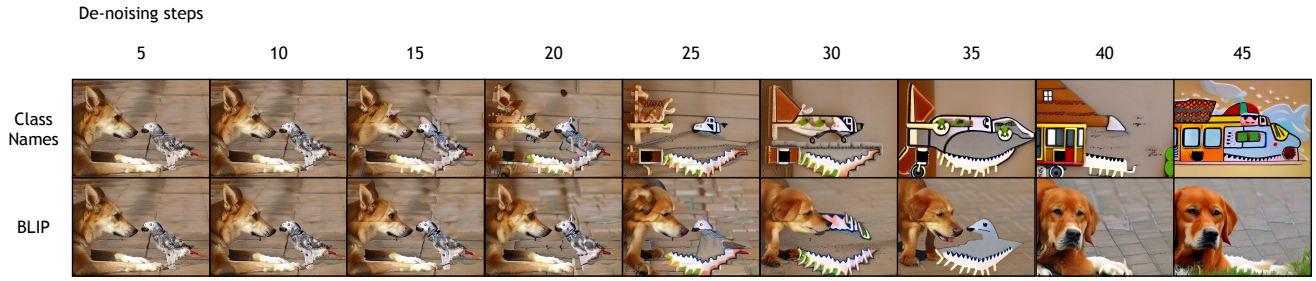
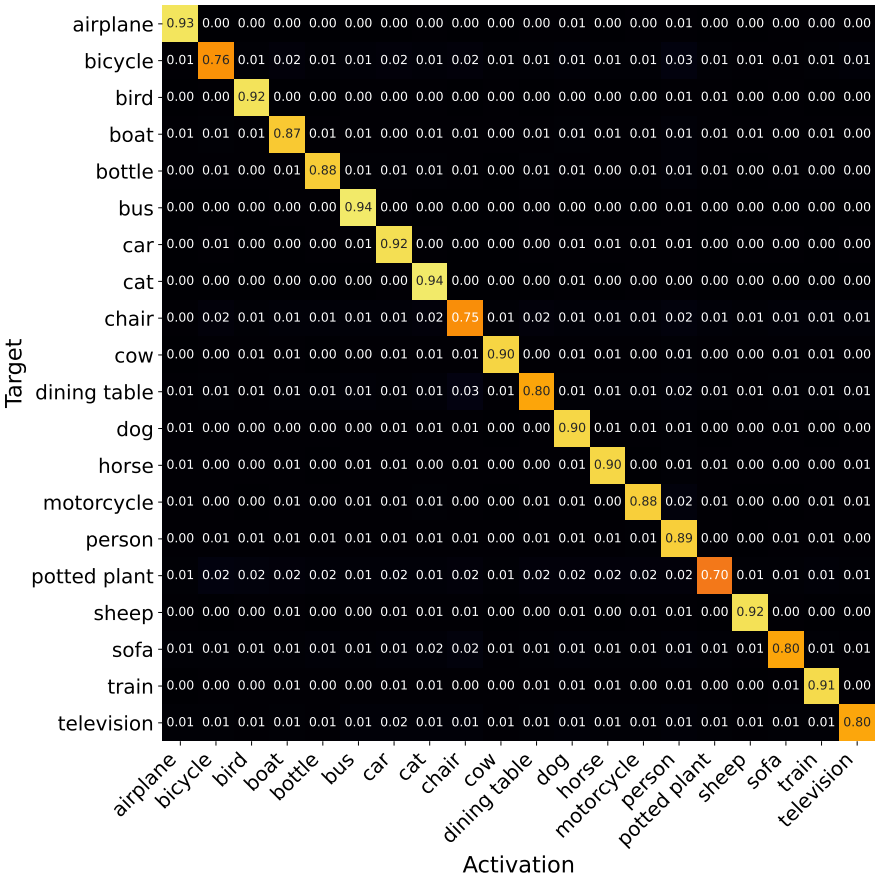


Figure S1. **Qualitative image-to-image variation.** An untrained stable diffusion model is passed an image to perform image-to-image variation. The number of denoising steps conducted increases from left to right (5 to 45 out of a total of 50). On the top row, we pass all the class names in Pascal VOC 2012: *"background airplane bicycle bird boat bottle bus car cat chair cow dining table dog horse motorcycle person potted plant sheep sofa train television"*. In the bottom row we pass the BLIP caption *"a bird and a dog"*.



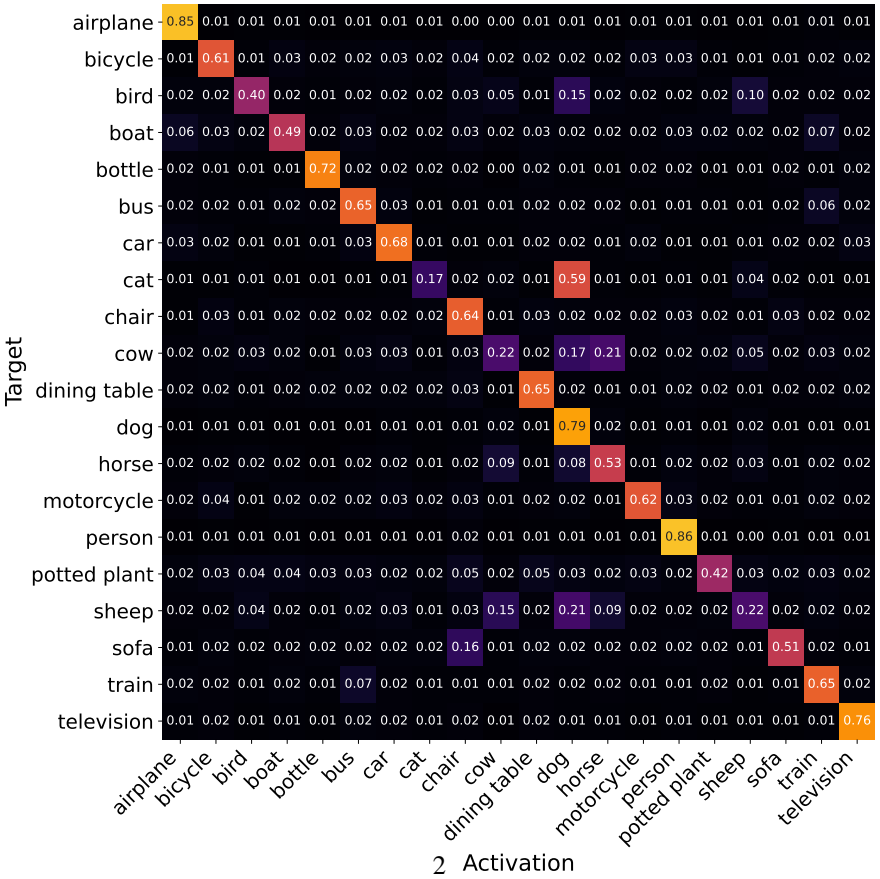


Figure S2. Normalized scores averaged over pixels on Pascal VOC 2012 for an oracle-trained model when receiving either present class names (top) or all class names (bottom).

B. Additional Results

Method	4K Iters		8K Iters	
	mIoU^{ss}	mIoU^{ms}	mIoU^{ss}	mIoU^{ms}
VPD_{A32} [50]	43.1	44.2	48.7	49.5
VPD(R)	42.6	43.6	49.2	50.4
VPD(LS)	45.0	45.8	50.5	51.1
TADP-20 (Ours)	**50.2**	**50.9**	**52.8**	**54.1**
TADP(TA)-20 (Ours)	49.9	50.7	52.7	53.4

Table S1. **Semantic segmentation fast schedule.** Our method has a large advantage over prior work on the fast schedule with significantly better performance in both the single-scale and multi-scale evaluations for 4k and 8k iterations.

		Recall		
		0.50	**0.75**	**1.00**
Precision	**0.50**	49.53	52.00	55.22
	0.75	49.17	51.46	58.62
	1.00	50.20	54.82	63.29

Table S2. **ADE20K - Oracle Precision-Recall Ablations** We modify the oracle captions by randomly adding or removing classes such that the precision and recall are 0.50, 0.75, or 1.00. We train models on ADE20K on a fast schedule (4K) using these captions. The 4k iteration oracle equivalent is highlighted in blue.

B.1. Qualitative Examples

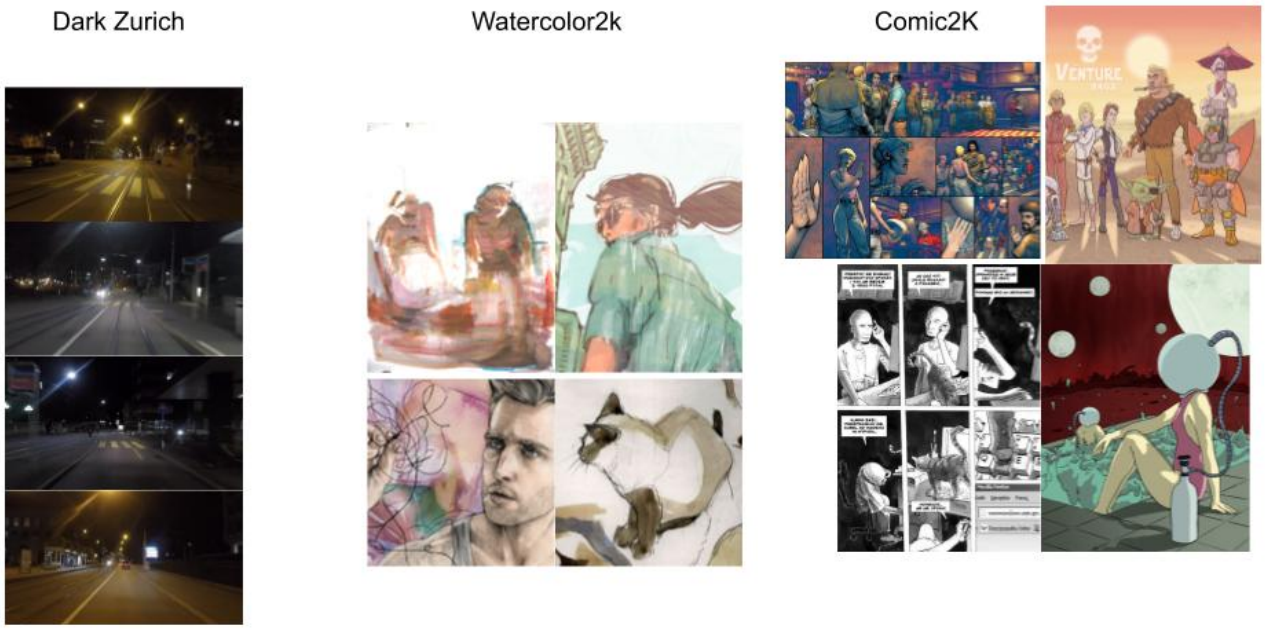


Figure S3. **Ground truth examples of the tokenized datasets**

Figure S4. **Textual inversion and Dreambooth tokens of Cityscapes to Dark Zurich**

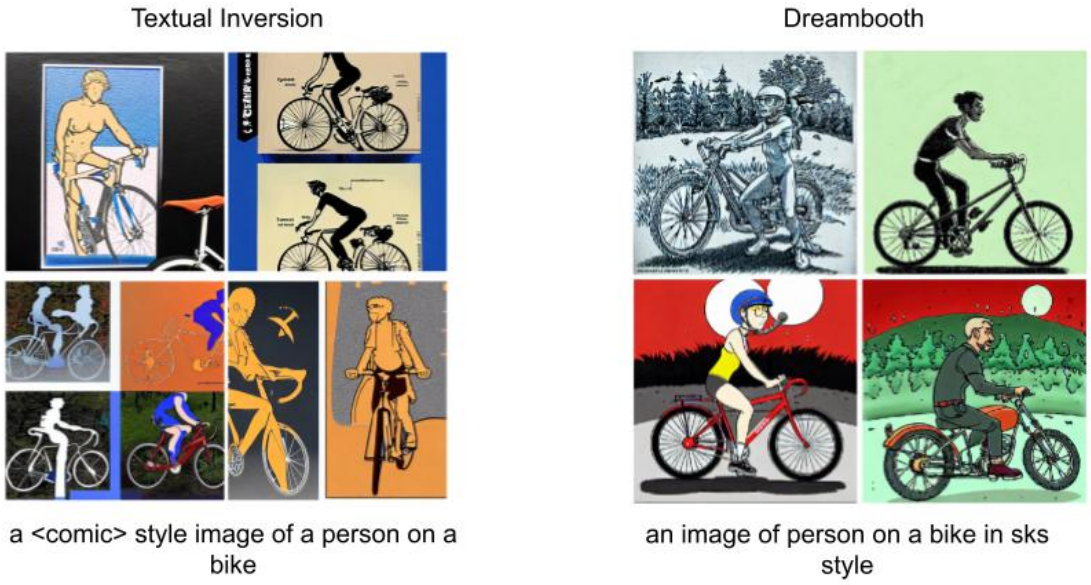


Figure S5. **Textual inversion and Dreambooth tokens of VOC to Comic**

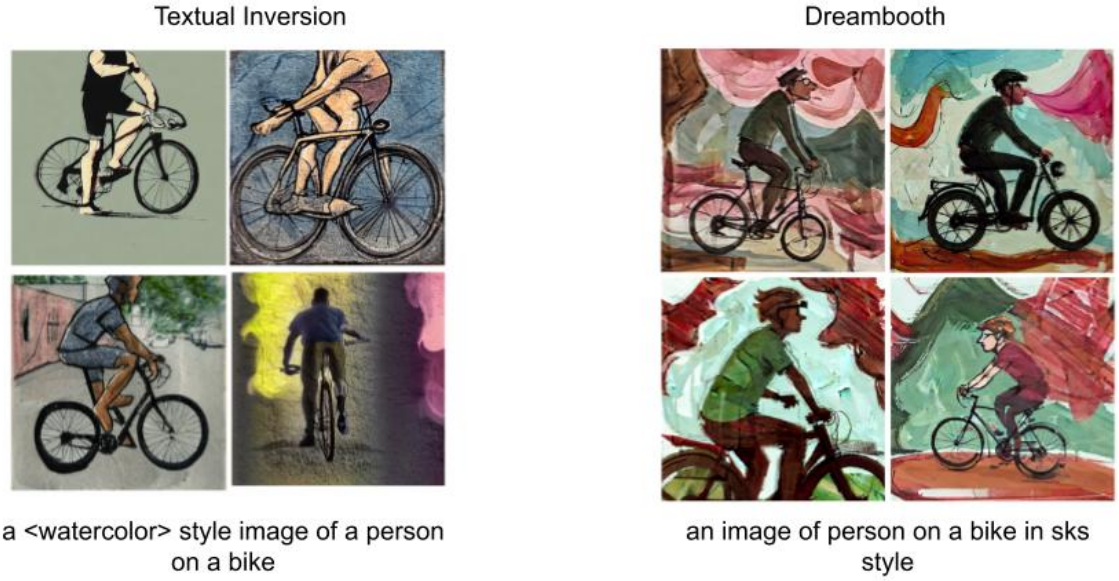


Figure S6. **Textual inversion and Dreambooth tokens of VOC to Watercolor**

Figure S7. **Predictions (top) and Ground Truth (bottom) visualizations for Pascal VOC2012.**

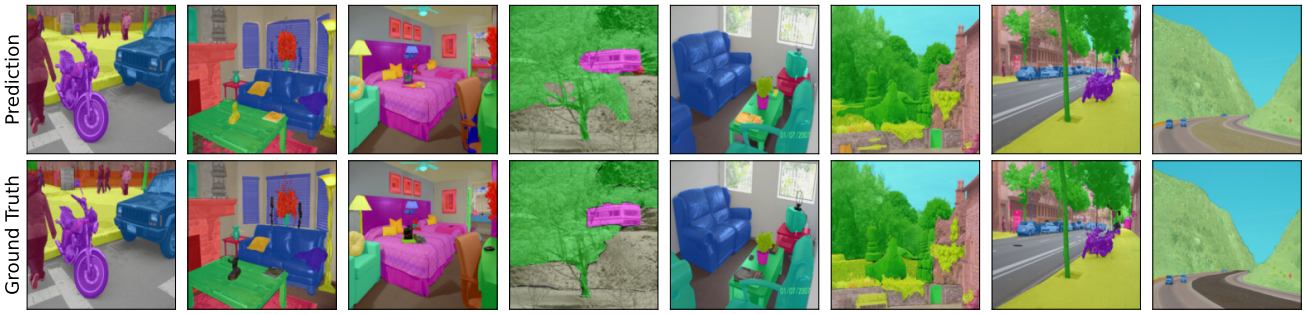


Figure S8. **Predictions (top) and Ground Truth (bottom) visualizations for ADE20K.**

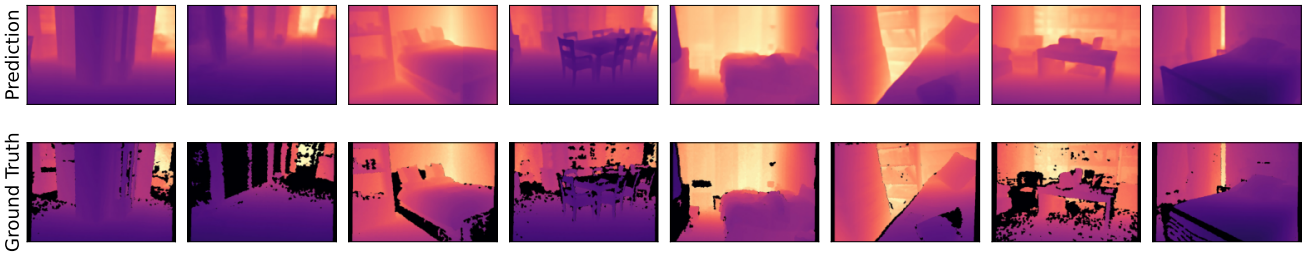


Figure S9. **Predictions (top) and Ground Truth (bottom) visualizations for NYUv2 Depth.**

C. Implementation Details

In order to isolate the effects of our text-image alignment method, we ensure our model setup follows prior work precisely. Following VPD [50], we jointly train the task-specific head and the diffusion backbone. The learning rate of the backbone is set to 1/10 the learning rate of the head to preserve the benefits of pre-training better. We describe the different tasks by describing H and $\mathcal{L}_H$. We use an FPN [23] head with a cross-entropy loss for segmentation. We use the same convolutional head used in VPD for monocular depth estimation with a Scale-Invariant loss [11]. For object detection, we use a Faster-RCNN head with the standard Faster-RCNN loss [32][1]. Further details of the training setup can be found in Tab. C.1 and Sec. C. In our single-domain tables, we include our reproduction of VPD, denoted with a (R). We compute our relative gains with our reproduced numbers, with the same seed for all experiments.

Table C.1. Single-Domain Hyperparameters

Hyperparameter	**Value**
Learning Rate	0.00008
Batch Size	2
Optimizer	AdamW
Weight Decay	0.005
Warmup Iters	1500
Warmup Ratio	$1e-6$
Unet Learning Rate Scale	0.01
Training Steps	80000

(a) ADE20k - full schedule

Hyperparameter	**Value**
Learning Rate	0.00016
Batch Size	2
Optimizer	AdamW
Weight Decay	0.005
Warmup Iters	150
Warmup Ratio	$1e-6$
Unet Learning Rate Scale	0.01
Training Steps	8000

(b) ADE20k - fast schedule 8k

Hyperparameter	**Value**
Learning Rate	0.00016
Batch Size	2
Optimizer	AdamW
Weight Decay	0.005
Warmup Iters	75
Warmup Ratio	$1e-6$
Unet Learning Rate Scale	0.01
Training Steps	4000

(c) ADE20k - fast schedule 4k

Hyperparameter	**Value**
Learning Rate	$5e-4$
Batch Size	3
Optimizer	AdamW
Weight Decay	0.1
Layer Decay	0.9
Epochs	25
Drop Path Rate	0.9

(d) NYUv2

Hyperparameter	**Value**
Learning Rate	0.00001
Batch Size	2
Gradient Accumulation	4
Epochs	15
Optimizer	AdamW
Weight Decay	0.01

(e) Pascal VOC

[1] Object detection was not explored in VPD.

Hyperparameter	Value
Learning Rate	0.00008
Batch Size	2
Optimizer	AdamW
Weight Decay	0.005
Warmup Iters	1500
Warmup Ratio	$1e-6$
Unet Learning Rate Scale	0.01
Training Steps	40000

(a) Cross-Domain Hyperparameters

(b) Cityscapes $\rightarrow$ Dark Zurich & NightTime Driving

Hyperparameter	Value
Learning Rate	0.00001
Batch Size	2
Epochs	100
Optimizer	AdamW
Weight Decay	0.01
Learning Rate Schedule	Lambda

(c) Pascal VOC $\rightarrow$ Watercolor & Comic

Hyperparameter	Value
Prior Preservation Cls Images	200
Learning Rate	$5e-6$
Training Steps	1000

(d) Dreambooth Hyperparameters

Hyperparameter	Value
Steps	3000
Learning Rate	$5.0e-04$
Batch Size	1
Gradient Accumulation	4

(e) Textual Inversion Hyperparameters

C.1. Model personalization

For textual inversion, we use 500 images from DZ-train and 5 images for W2K and C2K and train all tokens for 1000 steps. We use a constant learning rate scheduler with a learning rate of $5e-4$ and no warmup. For Dreambooth, we use the same images as in textual inversion but train the model for 500 steps (DZ) steps or 1000 steps (W2K and C2K). We use a learning rate of $2e-6$ with a constant learning rate scheduler and no warmup. We use no prior preservation loss.