LD-ZNet: A Latent Diffusion Approach for Text-Based Image Segmentation

Koutilya PNVR[†*] Bharat Singh[‡*] Pallabi Ghosh[§] Behjat Siddiquie[§] David Jacobs[†]

University of Maryland College Park[†] Vchar.ai[‡] Amazon[§]

{koutilya, djacobs}@umiacs.umd.edu bharat@vchar.ai {gpallabi, behjats}@amazon.com

Abstract

Large-scale pre-training tasks like image classification, captioning, or self-supervised techniques do not incentivize learning the semantic boundaries of objects. However, recent generative foundation models built using text-based latent diffusion techniques may learn semantic boundaries. This is because they have to synthesize intricate details about all objects in an image based on a text description. Therefore, we present a technique for segmenting real and AI-generated images using latent diffusion models (LDMs) trained on internet-scale datasets. First, we show that the latent space of LDMs (z-space) is a better input representation compared to other feature representations like RGB images or CLIP encodings for text-based image segmentation. By training the segmentation models on the latent z-space, which creates a compressed representation across several domains like different forms of art, cartoons, illustrations, and photographs, we are also able to bridge the domain gap between real and AI-generated images. We show that the internal features of LDMs contain rich semantic information and present a technique in the form of LD-ZNet to further boost the performance of text-based segmentation. Overall, we show up to 6% improvement over standard baselines for text-to-image segmentation on natural images. For AI-generated imagery, we show close to 20% improvement compared to state-of-the-art techniques. The project is available at https://koutilya-pnvr.github.io/LD-ZNet/.

1. Introduction

Teaching neural networks to accurately find the boundaries of objects is hard and annotation of boundaries at internet scale is impractical. Also, most self-supervised or weakly supervised problems do not incentivize learning boundaries. For example, training on classification or captioning allows models to learn the most discriminative parts of the image without focusing on boundaries [42, 60]. Our insight is that Latent Diffusion Models (LDMs) [38], which can be trained without object level supervision at internet scale, must attend to object boundaries, and so we hypothesize that they can learn features which would be useful for open world image segmentation. We support this hypothesis by showing that LDMs can improve performance on this task by up to 6%, compared to standard baselines and these gains are further amplified when LDM based segmentation models are applied on AI generated images.

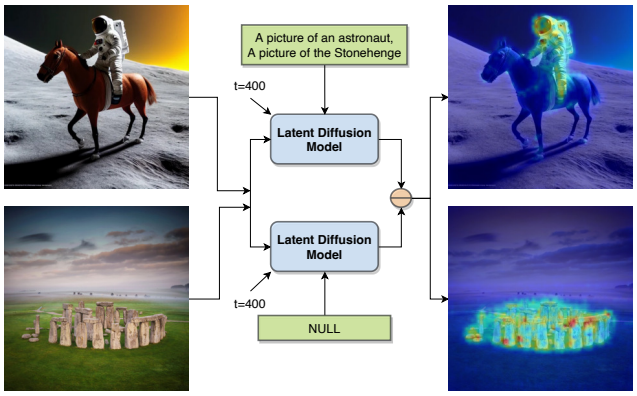


Figure 1: Coarse segmentation results from an LDM for two distinct images, demonstrating the encoding of fine-grained object-level semantic information within the model's internal features.

To test the aforementioned hypothesis about the presence of object-level semantic information inside a pre-trained LDM, we conduct a simple experiment. We compute the pixel-wise norm between the unconditional and text-conditional noise estimates from a pretrained LDM as part of the reverse diffusion process. This computation identifies the spatial locations that need to be modified for the noised input to align better with the corresponding text condition. Hence, the magnitude of the pixel-wise norm depicts regions that identify the text prompt. As shown in the Figure 1, the pixel-wise norm represents a coarse segmentation of the subject although the LDM is not trained on this task. This clearly demonstrates that these large scale LDMs can not only generate visually pleasing images, but their internal representations encode fine-grained semantic information, that can be useful for tasks like segmentation.

Recently, text-based image segmentation has gained

*This work was done when Koutilya and Bharat were at Amazon.

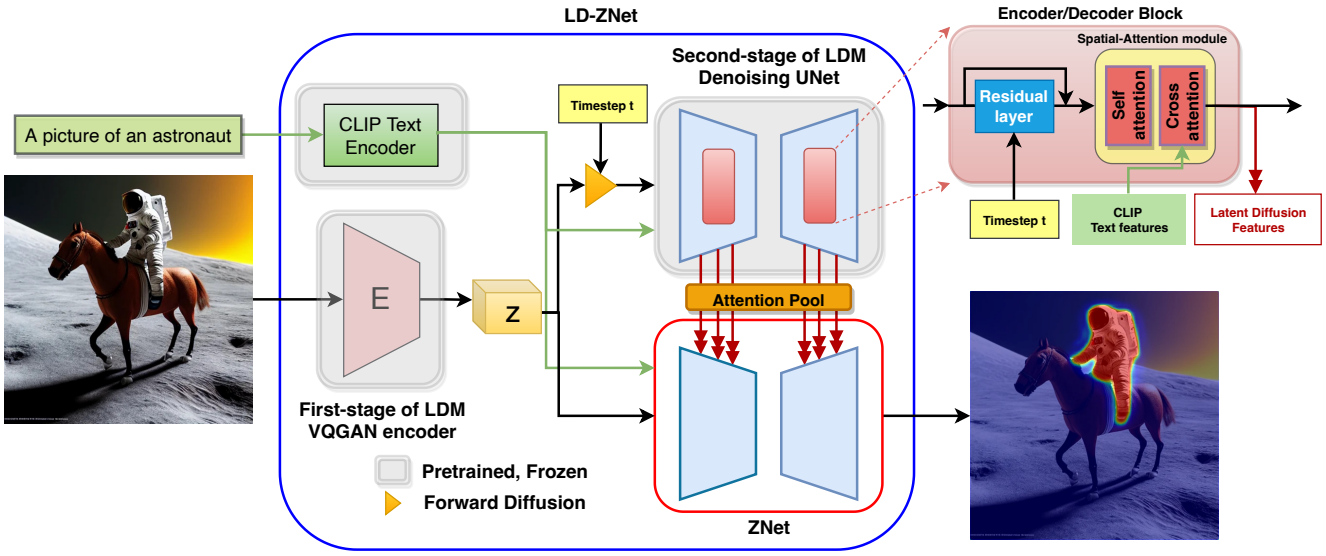


Figure 2: Overview of the proposed ZNet and LD-ZNet architectures. We propose to use the compressed latent representation z as input for our segmentation network ZNet. Next, we propose LD-ZNet, which incorporates the latent diffusion features at various intermediate blocks from the LDM's denoising UNet, into ZNet.

traction for creating and editing AI generated content (like AI art, illustrations, cartoons etc.) in image inpainting workflows [1] as it provides a conversational interface. Since the latent space z [11], extracted by a VQGAN is trained on several domains like art, cartoons, illustrations and real photographs, we posit that it is a more robust input representation for text-based segmentation on AI-generated images. Furthermore, the internal layers of the LDM are responsible for generating the structure of the image and hence contain rich semantic information about objects. Soft masks from these layers have also been used as a latent input in recent work on image editing [15, 2]. Since this information is already present while generating the image, we propose an architecture in the form of LD-ZNet (shown in Figure 2) to decode it for obtaining the semantic boundaries of objects generated in the scene. Not only does our architecture benefit segmentation of objects in AI generated images, but it also improves performance over natural images. Overall our contributions are as follows:

- We propose a text-based segmentation architecture, ZNet that operates on the compressed latent space of the LDM (z).
- Next, we study the internal representations at different stages of pretrained LDMs and show that they are useful for text-based image segmentation.
- Finally, we propose a novel approach named LD-ZNet to incorporate the visual-linguistic latent diffusion features from a pretrained LDM and show improvements across several metrics and domains for text-based image segmentation.

[1] https://github.com/brycedrennan/imaginAIry, https://github.com/AUTOMATIC1111/stable-diffusion-webui

2. Related work

2.1. Text-based image segmentation

Text-based image segmentation is the general task of segmenting specific regions in an image, based on a text prompt. This is different from the referring expression segmentation (RES) task, which aims to extract instance-level segmentation of different objects through distinctive referring expressions. While RES helps applications in robotics that require localization of a *single* object in an image, text-based segmentation benefits image editing applications by being able to also segment 1) "stuff" categories (clouds/ocean/beach *etc*.) and 2) multiple instances of an object category applicable to the text prompt. However, both these tasks have some shared literature in terms of approaches. Preliminary works [16, 25, 43, 24, 56] focused on the multi-modal feature fusion between the language and visual representations obtained from recurrent networks (such as LSTM) and CNNs respectively. The subsequent set of works [28, 59, 50, 54] included variations of multi-modal training, attention and cross-attention networks etc. Recently, [50, 26] used CLIP [34] to extract visual linguistic features of the image and the reference text separately. These features were then combined using a transformer based decoder to predict a binary mask. Alternately, [18, 62], proposed vision-language pretraining on other text-based visual recognition tasks (object detection and phrase grounding) and later finetuned for the segmentation task. The concurrent works segment-

anything (SAM) [22] and segment-everything-everywhere-all-at-once (SEEM) [67] allow interactive segmentation via point clicks, bounding boxes and text inputs *etc*. demonstrating good zero-shot performance. Different from all these works, we show the significance of using the latent space and the internal features from a pretrained latent diffusion model [38] for improving the more generic text-based image segmentation task.

2.2. Text-to-Image synthesis

Text-to-Image synthesis has initially been explored using GANs [53, 66, 45, 61, 55, 65] on publicly available image captioning datasets. Another line of work is by using autoregressive models [36, 9, 13] via a two stage approach. The first stage is a vector quantized autoencoder such as a VQVAE [48, 37] or a VQGAN [11] with an image reconstruction objective to convert an image into a shorter sequence of discrete tokens. This low dimensional latent space enables the training of compute intensive autoregressive models even for high resolution text-to-image synthesis. With the recent advancements in Diffusion Models (DM) [32, 8], both in unconditional and class conditional settings, they have started gaining more traction compared to GANs. Their success in the text-to-image tasks [40, 35] made them even more popular. However, the prior diffusion models worked in the high-dimensional image space that made training and inference computationally intensive. Subsequently, latent space representations [31, 14, 44, 38] were proposed for high resolution text-to-image synthesis to reduce the heavy compute demands. More specifically, the latent diffusion model (LDM) [38] mitigates this problem by relying on a perceptually compressed latent space produced by a powerful autoencoder from the first stage. Moreover, they employ a convolutional backed UNet [39] as the denoising architecture, allowing for different sized latent spaces as input. Recently this architecture is trained on large scale text-image data [41] from the internet and released as Stable-diffusion[2], which exhibited photo-realistic image generations. Subsequently, several language guided image editing applications such as inpainting [5, 27, 52], text-guided image editing [3, 2] became more popular and the usage for text-based image segmentation has surged, especially for AI generated images. We propose a solution for text-based image segmentation by leveraging the features which are already present as part of the synthesis process.

2.3. Semantics in generative models

Semantics in generative models such as GANs have been studied for binary segmentation [49, 29] as well as multi-class segmentation [64, 46, 33] where the intermediate features have been shown to contain semantic information for these tasks. Moreover, [23] highlighted the practical advantages of these representations, such as out-of-distribution robustness. However, prior generative models (GANs *etc*.) as representation learners have received less attention compared to alternative unsupervised methods [4], because of the training difficulties on complex, diverse and large scale datasets. Diffusion models [32], on the other hand are another class of powerful generative models that recently outperformed GANs on image synthesis [8] and are able to train on large datasets such as Imagenet [7] or LAION [41]. In [1], the authors demonstrated that the internal features of a pre-trained diffusion model were effective at the semantic segmentation task. However, this type of analysis [64, 1] has mostly been done in limited settings like few shot learning [12] or limited domains like faces [19], horses [57] or cars [57]. Different from these works, we analyze the visual-linguistic semantic information present in the internal features of a text-to-image LDM [38] for text based image segmentation, which is an open world visual recognition task. Furthermore, we leverage these LDM features and show performance improvements when training with full datasets instead of few-shot settings.

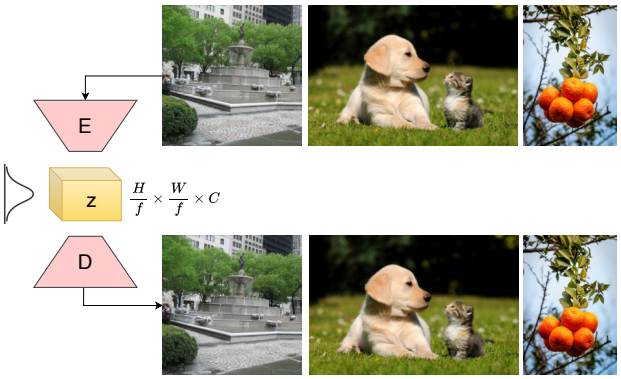


Figure 3: Reconstructions from the first stage of the LDM. Given an input image, the latent representation z generated by the encoder, can be used to reconstruct images that are perceptually indistinguishable from the inputs. The high quality of these reconstructions suggests that the latent representation z, preserves most of the semantic information present in the input images.

3. LDMs for Text-Based Segmentation

The text-to-image latent diffusion architecture introduced in [38] consists of two stages: 1) An auto-encoder based VQGAN [11] that extracts a compressed latent representation (z) for a given image 2) A diffusion UNet that is trained to denoise the noisy z created in the forward diffusion process, conditioned on the text features. These text features are obtained from a pretrained frozen CLIP text encoder [34] and is conditioned at multiple layers of the UNet via cross-attention.

In this paper, we show performance improvements on the

[2] https://github.com/CompVis/stable-diffusion

text-based segmentation task in two steps. Firstly, we analyze the compressed latent space (z) from the first-stage and propose an approach named ZNet that uses z as the visual input to estimate segmentation mask when conditioned on a text prompt. Secondly, we study the internal representations from the second stage of the stable-diffusion LDM for visual-linguistic semantic information and propose a way to utilize them inside ZNet for further improvements in the segmentation task. We name this approach as LD-ZNet.

3.1. ZNet: Leveraging Latent Space Features

We observe that the latent space (z) from the first-stage of the LDM is a compressed representation of the image that preserves semantic information, as depicted in Figure 3. The VQGAN in the first-stage achieves such semantic-preserving compression with the help of large scale training data as well as a combination of losses - perceptual loss [63], a patch-based [17] adversarial objective [10, 11, 58], and a KL-regularization loss.

In our experiments, we observe that this compressed latent representation z is more robust compared to the original image in terms of their association with the text prompts. We believe this is because z is a $\frac{H}{8} \times \frac{W}{8} \times 4$ dimensional feature with 48 $\times$ fewer elements compared to the original image, while preserving the semantic information. Several prior works [47, 21, 6], show that compression techniques like PCA, which create information preserving lower dimensional representations generalize better. Therefore, we propose using the z representation along with the frozen CLIP text features [34] as an input to our segmentation network. Furthermore, because the VQGAN is trained across several domains like art, cartoons, illustrations, portraits, etc., it learns a robust and compact representation which generalizes better across domains, as can be seen in our experiments on AI generated images. We call this approach ZNet. The architecture of ZNet is shown in the bottom box of Figure 2, and is the same as the denoising UNet module of the LDM. We therefore initialize it with pretrained weights of the second-stage of the LDM.

3.2. LD-ZNet: Leveraging Diffusion Features

Given a text prompt and a timestep t, the second-stage of the LDM is trained to denoise z_t - a noisy version of the latent representation z obtained via forward diffusion process for t timesteps. A UNet architecture is used whose encoder/decoder elements are shown in Figure 2 (top right). A typical encoder/decoder block contains a residual layer followed by a spatial-attention module that internally has self-attention and then cross-attention with the text features. We analyze the semantic information in the internal visual-linguistic representations developed at different blocks of encoder and decoder right after these spatial-attention modules. We also propose a way to utilize these latent diffusion features using cross-attention into the ZNet segmentation network and we call the final model as LD-ZNet.

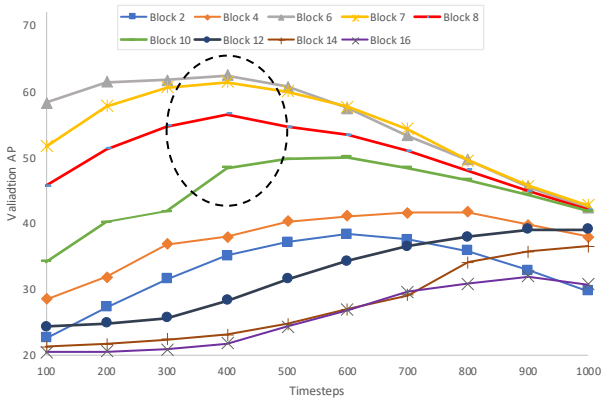


Figure 4: Semantic information present in the LDM features at various blocks and timesteps for the referring image segmentation task. AP is measured on a small validation subset of the PhraseCut dataset.

3.2.1 Visual-Linguistic Information in LDM Features

We evaluate the semantic information present in the pre-trained LDM at various blocks and timesteps for the text-based image segmentation task. In this experiment, we consider the latent diffusion features right after the spatial-attention layers 1-16 spanning across all the encoder and decoder blocks present in the UNet. At each block, we analyze the features for every 100^{th} timestep in the range $[100, 1000]$. We use a small subset of the training and validation sets from the Phrasecut dataset and train a simple decoder on top of these features to predict the associated binary mask. Specifically, given an image I and timestep t, we first extract its latent representation z from the first stage of LDM and add noise from the forward diffusion to obtain z_t for a timestep t. Next we extract the frozen CLIP text features for the text prompt and input both of them into the denoising UNet of the LDM to extract the internal visual-linguistic features at all the blocks for that timestep. We use these representations to train the corresponding decoders until convergence. Finally, we evaluate the AP metric on a small subset of the validation dataset. The performance of features from different blocks and timesteps is shown in Figure 4.

Similar to [1], we observe that the middle blocks {6,7,8,9,10} of the UNet contain more semantic information compared to either the early blocks of the encoder or the later blocks of the decoder. We also observe that the timesteps 300-500 contain the maximum visual-linguistic semantic information compared to other timesteps, for these middle blocks. This is in contrast to the findings of [1] that report the timesteps {50, 150, 250} to contain the

most useful information when evaluated on an unconditional DDPM model for the few shot semantic segmentation task for horses [57] and faces[19]. We believe that the reason for this difference is because, in our case, the image synthesis is guided by text, leading to the emergence of semantic information earlier in the reverse diffusion process (t=1000→0), in contrast to unconditional image synthesis.

3.2.2 LD-ZNet Architecture

We propose using the aforementioned visual-linguistic representations at multiple spatial-attention modules of the pretrained LDM into the ZNet as shown in Figure 2. These latent diffusion features are injected into the ZNet via a cross-attention mechanism at the corresponding spatial-attention modules as shown in Figure 5. This allows for an interaction between the visual-linguistic representations from the ZNet and the LDM. Specifically, we pass the latent diffusion features through an *attention pool* layer that not only acts as a learnable layer to match the range of the features participating in the cross-attention, but also adds a positional encoding to the pixels in the LDM representations. The outputs from the attention pool are now positional-encoded visual-linguistic representations that enable the proposed cross-attention mechanism to attend to the corresponding pixels from the ZNet features. ZNet when augmented with these latent diffusion features from the LDM (through cross-attention) is referred to as LD-ZNet.

Following the semantic analysis of latent diffusion features (Sec. 3.2.1), we incorporate the internal features from blocks {6,7,8,9,10} of the LDM into the corresponding blocks of ZNet, in order to make use of the maximum semantic and diverse visual-linguistic information from the LDM. For AI generated images, these blocks are anyways responsible to generate the final image and using LD-ZNet, we are able to tap into this information which can be used for segmenting objects in the scene.

4. Experiments

Implementation details: In this paper, we use the stable-diffusion v1.4 checkpoint as our LDM that internally uses the frozen ViT-L/14 CLIP text encoder [34]. We implement the above described ZNet and LD-ZNet in pytorch inside the stable-diffusion library. We also initialize our networks with the weights from the LDM wherever possible, while initializing the remaining parameters from a normal distribution. We train ZNet and LD-ZNet on 8 NVIDIA A100 gpus with a batch size of 4 using the Adam optimizer and a base learning rate of $5e^{-7}$ per mini-batch sample, per gpu. For all our experiments, we keep the text encoder frozen and use an image resolution of 384 for a fair comparison with the previous works.

Datasets: We use Phrasecut [51], which is currently the

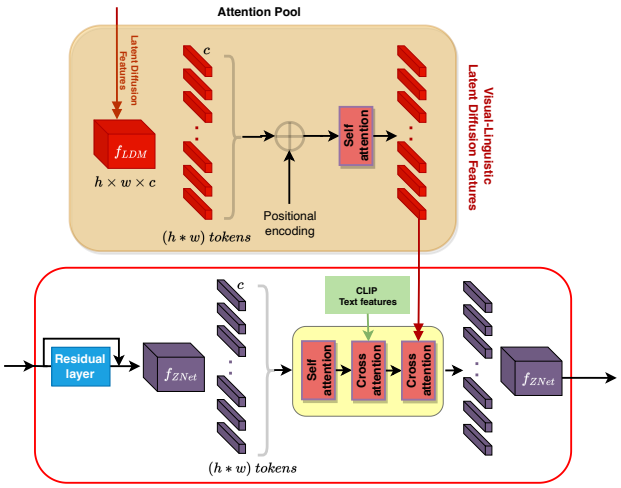


Figure 5: We propose to incorporate the visual-linguistic representations from LDM obtained at the spatial-attention modules via a cross-attention mechanism into the corresponding spatial-attention modules of the ZNet through an *attention pool* layer.

largest dataset for the *text-based image segmentation* task, with nearly 340K phrases along with corresponding segmentation masks that not only permit annotations for stuff classes but also accommodate multiple instances. Following [34], we randomly augment the phrases from a fixed set of prefixes. For the images, we randomly crop a square around the object of interest with maximum area, ensuring that the object remains at least partially visible. We avoid negative samples to remove ambiguity in the LDM features for non-existent objects.

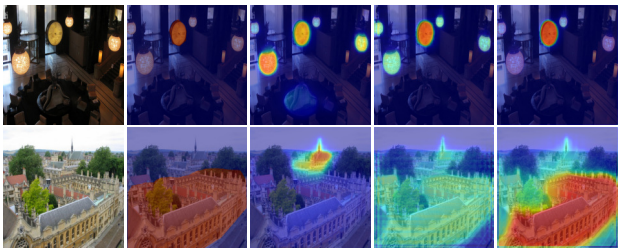
We create a dataset consisting of AI-generated images which we name **AIGI** dataset, to showcase the usefulness of our approach for text-based segmentation on a different domain. We use 100 AI-generated images from *lexica.art* and manually annotated multiple regions for 214 text-prompts relevant to these images.

We also use the popular referring expression segmentation datasets namely RefCOCO [20], RefCOCO+ [20] and G-Ref [30] to demonstrate the generalization abilities of ZNet and LD-ZNet. In RefCOCO, each image contains two or more objects and each expression has an average length of 3.6 words. RefCOCO+ is derived from RefCOCO by excluding certain absolute-location words and focuses on purely appearance based descriptions. For example it uses "the man in the yellow polka-dotted shirt" rather than "the second man from the left" which makes it more challenging. Unlike RefCOCO and RefCOCO+, the average length of sentences in G-Ref is 8.4 words, which have more words about locations and appearances. While we adopt the UNC partition for RefCOCO and RefCOCO+ in this paper, we use the UMD partition for G-Ref.

Metrics: We follow the evaluation methodology of [26]

Method	mIoU	IoU_{FG}	AP
MDETR [18]	53.7	-	-
GLIPv2-T [62]	59.4	-	-
RMI [51]	21.1	42.5	-
Mask-RCNN Top [51]	39.4	47.4	-
HulaNet [51]	41.3	50.8	-
CLIPSeg (PC+) [26]	43.4	54.7	76.7
CLIPSeg (PC, D=128) [26]	48.2	56.5	78.2
RGBNet	46.7	56.2	77.2
ZNet (Ours)	51.3	59.0	78.7
LD-ZNet (Ours)	**52.7**	**60.0**	**78.9**

Table 1: Text-based image segmentation performance on the PhraseCut testset. The performance of ZNet and LD-ZNet is highlighted in gray. Both these models outperform the baseline RGBNet on all the metrics.



(a) Input **(b)** GT mask **(c)** RGBNet **(d)** ZNet **(e)** LD-ZNet

Figure 6: Qualitative comparison on the PhraseCut test set. Each row contains an input image with a text prompt as an input, with the goal being to segment the image regions corresponding to the reference text. The text prompts are *"hanging clock"* and *"castle"* for the top and bottom rows. We show improvements using ZNet and LD-ZNet compared to the RGBNet.

and report best foreground IoU (IoU_{FG}) for the foreground pixels, the best mean IoU of all pixels (mIoU), and the Average Precision (AP).

5. Results

5.1. Image Segmentation Using Text Prompts

On the PhraseCut dataset, we compare the performance of previous approaches with our ZNet and LD-ZNet for the text-based image segmentation task (Table 1). In order to showcase the performance improvement of our proposed networks, we create a baseline named RGBNet with the same architecture as ZNet except we use the original images as the input instead of its latent space z. For RGBNet, we use additional learnable convolutional layers to map the original image to match the input resolution of ZNet. From Table 1, we observe that our ZNet and LD-ZNet significantly outperform RGBNet. Specifically, the performance improvement from using the latent representation z over the original images is clear (i.e. ZNet vs RGBNet baseline). Performance further improves upon incorporating the LDM visual-linguistic representations (LD-ZNet) - by 6% overall on the $mIoU$ metric compared to RGBNet. We also highlight this qualitatively in Figure 6. In the figure, we show the original image and the GT mask along with outputs from the RGBNet baseline followed by ZNet and LD-ZNet, where both ZNet and LD-ZNet help improve results consistently. For example in the top row, RGBNet detects light fixtures for the "hanging clock" prompt, and although ZNet does not have as strong activations for these incorrect detections, it is LD-ZNet that correctly segments the "clock". Similarly in the bottom row, while RGBNet completely got the "castle" wrong, ZNet correctly has activations on the right buildings, but with lower confidence. However, LD-ZNet improves it further.

We outperform in all the metrics when compared to previous works, other than MDETR [18] and GLIPv2 [62]. Notably, these works are pre-trained on detection and phrase grounding for predicting bounding boxes on huge corpus of text-image pairs across various publicly available datasets with bounding box annotations and are later fine-tuned on the Phrasecut dataset for the segmentation task. However, our work is orthogonally focused towards exploring and utilizing LDMs and its internal features for improving the text-based segmentation performance. Note that object detection datasets have a good overlap with the visual content in PhraseCut, however, they are not representative of the diversity in images available on the internet. For example, while they could learn common concepts like sky, ocean, chair, table and their synonyms, methods like MDETR would not understand concepts like Mikey Mouse, Pikachu etc., which we will show in Section 6.

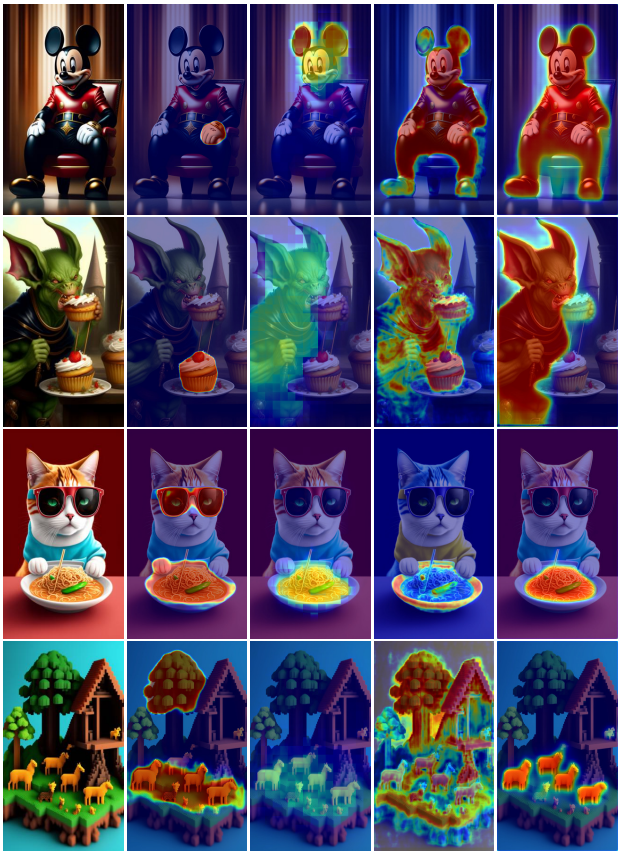
Method	mIoU	AP
MDETR [18]	53.4	63.8
CLIPSeg (PC+) [26]	56.4	79.0
SEEM [67]	57.4	70.0
RGBNet	63.4	84.1
ZNet (Ours)	68.4	85.0
LD-ZNet (Ours)	**74.1**	**89.6**

Table 2: Generalization of the proposed LD-ZNet on our AIGI dataset when compared with other state-of-the-art text-based segmentation methods.

5.2. Generalization to AI Generated Images

With the growing popularity of AI generated images, text-based image segmentation is extensively being used by content creators in their daily workflows. Many public libraries [3] widely employ methods such as CLIPSeg [26] for performing segmentation in AI-generated images. So we study the generalization ability of our proposed segmentation approach on AI-generated images. To this extent, we

[3] https://github.com/brycedrennan/imaginAIry, https://github.com/AUTOMATIC1111/stable-diffusion-webui



Input MDETR [18] CLIPSeg [26] SEEM [67] LD-ZNet

Figure 7: Qualitative comparison on the AI-generated images for text-based segmentation. The text prompts are *"Mickey mouse"*, *"Goblin"*, *"Ramen"* and *"animals"* respectively.

first prepare a dataset of 100 AI-generated images from lexica.art and manually annotate them using 214 text-prompts. We name this dataset AIGI and release it on our project website [4] for future research. Next, we evaluate our approaches ZNet and LD-ZNet along with our RGBNet baseline and other text-based segmentation methods - CLIPSeg (PC+) [26], MDETR [18] and SEEM [67]. Glipv2 and the SAM model [22] with textual input were not publicly available for us to evaluate at the time of this submission. All these methods are trained on the Phrasecut dataset except for SEEM and we measure the IoU metric as shown in Table 2. It can be seen that RGBNet outperforms CLIPSeg, MDETR and SEEM because its built on the UNet architecture initialized from the LDM weights that contains semantic information for good generalization. Our methods ZNet and LD-ZNet further improve the generalization to these AI-generated images by more than 20% compared to MDETR. This is largely due to the robust z-space of the LDM that resulted from a VQGAN pre-training on a variety

[4] https://koutilya-pnvr.github.io/LD-ZNet/

Method	RefCOCO		RefCOCO+		G-Ref	
	IoU	AP	IoU	AP	IoU	AP
CLIPSeg (PC+) [26]	30.1	14.1	30.3	15.5	33.8	23.7
RGBNet	36.3	15.7	37.1	16.7	41.9	27.8
ZNet (Ours)	40.1	16.8	40.9	17.8	47.1	29.2
LD-ZNet (Ours)	**41.0**	**17.2**	**42.5**	**18.6**	**47.8**	**30.8**

Table 3: Generalization of our proposed approaches to different types of expressions from other datasets. Z-Net and LD-ZNet outperform both the RGBNet baseline and CLIPSeg on the generalization across all datasets.

Diffusion features via	mIoU	IoU_{FG}	AP
LD-ZNet with concatenation	50.2	59.0	78.1
LD-ZNet with cross-attention	**52.7**	**60.0**	**78.9**

Table 4: Incorporating LDM features into ZNet via cross-attention (LD-ZNet) leverages the visual-linguistic information present in them, compared to concatenation, leading to better performance on the text-based image segmentation task.

of domains like art, cartoons, illustrations *etc*. Furthermore, the latent diffusion features that contain useful semantic information for the synthesis task, also help in segmenting the AI-generated images. We show the qualitative comparison of these methods in Figure 7 for four AI-generated images from our dataset. While CLIPSeg can estimate most distinctive regions such as face of the *Mickey mouse* or rough locations of *Goblin*, *Ramen* and *animals*, MDETR and SEEM incorrectly segment them because these concepts are unknown to them and because of the domain gap between their training data and AIGI images respectively. In both such cases, our proposed LD-ZNet estimates accurate segmentation. More qualitative results for LD-ZNet on images from the AIGI dataset are shown in Figure 9.

5.3. Generalization to Referring Expressions

Reference expression segmentation task is aimed for robot-localization kind of applications, where segmenting at instance-level is performed through distinctive referring expression. Many works such as [54, 50] also train the text encoder to learn the complex positional references in the text. However, we are focused on generic text-based segmentation that has support for stuff categories as well as for multiple instances. We study the generalization ability of the proposed approach - using LDM features, to this complex task. Specifically, we use the models trained on the PhraseCut dataset and evaluate them on the RefCOCO [20], RefCOCO+ [20] and G-Ref [30] datasets whose complex referring expressions are for single-instance localization and segmentation. We also evaluated the generalization of CLIPSeg (PC+) [26] model that was trained on an extended version of the PhraseCut dataset (PC+), to further demonstrate the generalization capability of our methods. Table 3 summarizes the performance for our models along

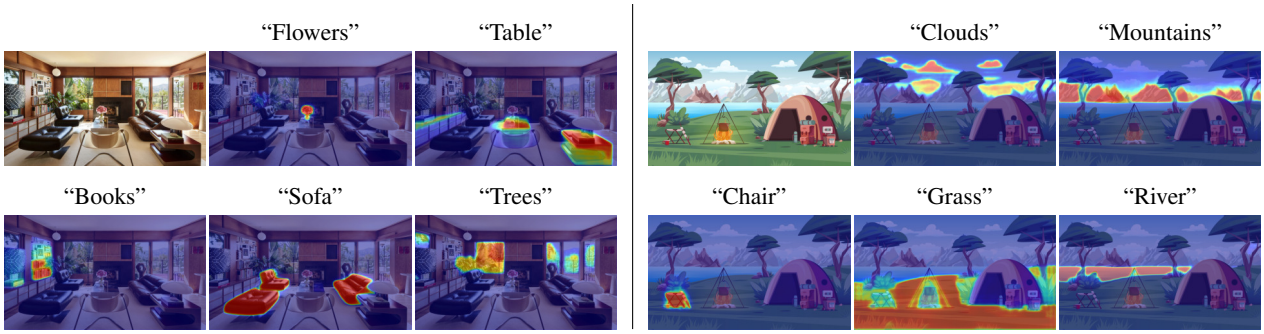


Figure 8: LD-ZNet text-based image segmentation results for a real image and a cartoon on diverse set of things and stuff classes. High quality segmentation across multiple classes suggests that LD-ZNet has a good understanding of the overall scene.

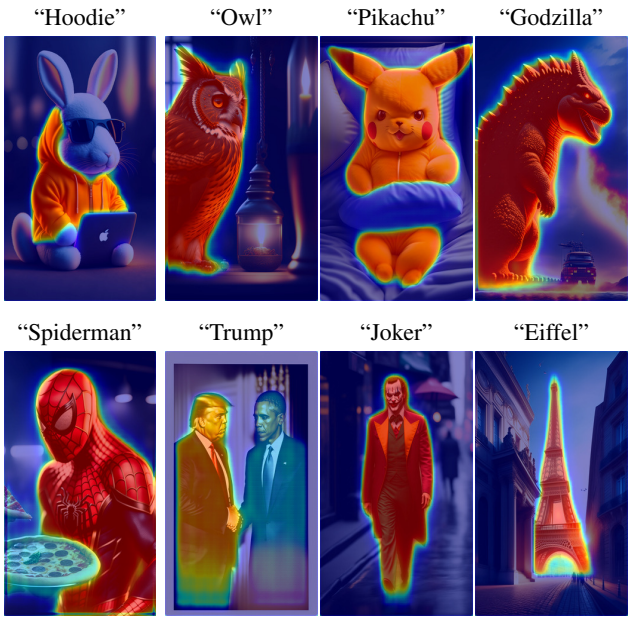


Figure 9: More qualitative results of LD-ZNet from AIGI dataset.

with the RGBNet baseline. We observe a similar trend in performance improvements across RGBNet $<$ ZNet $<$ LD-ZNet. These experiments demonstrate that the LDM features enhance the generalization power of the LD-ZNet model even on complex referring expressions.

5.4. Inference Time

During inference, our proposed LD-ZNet relies on the LDM to extract the internal features for just a single time step (as opposed to around 50 reverse diffusion time steps for the text-to-image synthesis task). We then use these LDM features for further cross-attention into LD-ZNet via the attention pool layer to extract the final mask. Therefore, using the diffusion model increases the overall run time by only a small amount. For the stable-diffusion model, inference takes 2.57s for 50 timesteps to synthesize an image (roughly 51ms per timestep), whereas the average inference times for RGBNet, ZNet and LD-ZNet are only 62ms, 55ms and 101ms, respectively, per image on the AIGI dataset with an RTX A6000 gpu. SEEM [67] takes 293ms for the same task. Since we use an architecture similar to UNet (from the second stage of the LDM), as our segmentation network, the proposed LD-ZNet has 925M trainable parameters.

5.5. Cross-attention vs Concat for LDM features

In LD-ZNet, we inject LDM features into the ZNet model using cross-attention (Figure 5). In order to understand the importance of the cross-attention layer, we also train and evaluate another model where the LDM features are concatenated with the features of the ZNet right before the spatial-attention layer. The results are summarized in Table 4 and it shows that concatenating the LDM features yields inferior results compared to the proposed method. This is because of the *attention pool* layer which serves as a learnable layer and also encodes positional information into the LDM features for setting up the cross-attention. Moreover, the cross-attention layer learns how feature pixels from the ZNet attend to feature pixels from the LDM, thereby leveraging context and correlations from the entire image. With concatenation however, we only fuse the corresponding features of LDM and ZNet which is sub-optimal.

6. Discussion

In this section we present more qualitative results to demonstrate several interesting aspects of our proposed technique when applied towards downstream segmentation tasks. In Figs. 7 to 10, we visualize results of text-based image segmentation on a diverse set of images, which include AI generated images, illustrations and generic photographs. In Figure 8, we show that when LD-ZNet is applied on the same image with various text prompts, it is able to correctly segment the object and stuff classes being referred to in both examples. This capability is crucial for open-world segmen-

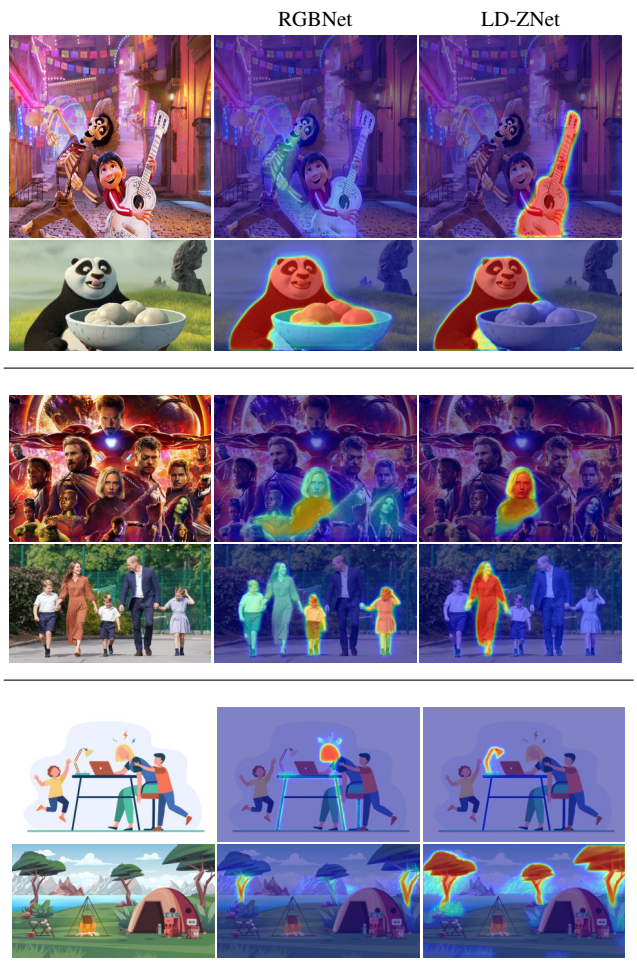


Figure 10: More qualitative examples where RGBNet fails to localize *"Guitar"*, *"Panda"* from animation images (top row), famous celebrities *"Scarlett Johansson"*, *"Kate Middleton"* (second row) and objects such as *"Lamp"*, *"Trees"* from illustrations (bottom row). LD-ZNet benefits from using z combined with the internal LDM features to correctly segment these text prompts.

tation and overall understanding of the scene. The results also highlights that the algorithm works remarkably well on other domains like cartoons/illustrations. It is noteworthy that LD-ZNet can perform accurate segmentation for text prompts which include cartoons (Pikachu, Godzilla), celebrities (Donald Trump, Spiderman), famous landmarks (Eiffel Tower), as seen in Figure 9. Finally, Figure 10 shows the advantages of leveraging semantic information present in the latent diffusion features. Compared to our baseline RGBNet, the proposed LD-ZNet generates better segmentation maps across animations, celebrity images and illustrations.

7. Conclusion

We presented a novel approach for text-based image segmentation using large scale latent diffusion models. By training the segmentation models on the latent z-space, we were able to improve the generalization of segmentation models to new domains, like AI generated images. We also showed that this z-space is a better representation for text-to-image tasks in natural images. By utilizing the internal features of the LDM at appropriate time-steps, we were able to tap into the semantic information hidden inside the image synthesis pipeline using a cross-attention mechanism, which further improved the segmentation performance both on natural and AI generated images. This was experimentally validated on several publicly available datasets and on a new dataset of AI generated images, which we will make publicly available.

8. Acknowledgments

Koutilya PNVR and David Jacobs were supported in part by the National Science Foundation under grant number IIS-1910132 and IIS-2213335.

References

[1] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3, 4

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2, 3

[3] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 3

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[6] Fernando De la Torre and Michael J Black. Robust principal component analysis for computer vision. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 362–369. IEEE, 2001. 4

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 3

[9] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835. Curran Associates, Inc., 2021. 3

[10] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016. 4

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3, 4

[12] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 3

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 3

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 3

[15] 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

[16] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pages 108–124. Springer, 2016. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 4

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2, 6, 7

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 5

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5, 7

[21] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004. 4

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3, 7

[23] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[24] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2

[25] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1271–1280, 2017. 2

[26] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 2, 5, 6, 7

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022. 3

[28] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018. 2

[29] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 3

[30] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 5, 7

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[33] Daniil Pakhomov, Sanchit Hira, Narayani Wagle, Kemar E. Green, and Nassir Navab. Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip, 2021. 3

[34] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 5

[35] 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. 3

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[37] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[42] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[43] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018. 2

[44] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 3

[45] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16515–16525, 2022. 3

[46] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4475–4485, 2021. 3

[47] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 4

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[49] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *International Conference on Machine Learning*, pages 10596–10606. PMLR, 2021. 3

[50] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022. 2, 7

[51] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 5, 6

[52] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. *arXiv preprint arXiv:2212.05034*, 2022. 3

[53] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[54] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2, 7

[55] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *The 32nd British Machine Vision Conference (BMVC)*, 2021. 3

[56] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 2

[57] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 3, 5

[58] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 4

[59] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2

[60] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 1

[61] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 3

[62] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2, 6

[63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[64] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 3

[65] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022. 3

[66] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[67] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once, 2023. 3, 6, 7, 8