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[†] Cruise LLC[‡] Amazon[§]

Generative pretraining learns object-level semantics

- Large-scale discriminative pretraining tasks such as image classification, captioning, or self-supervised techniques do not incentivize learning the semantic boundaries of objects
- Latest generative models pretrained using text-based latent diffusion techniques (LDMs) synthesize photorealistic objects → good object-level understanding!

Fine-grained semantic information in a pretrained LDM

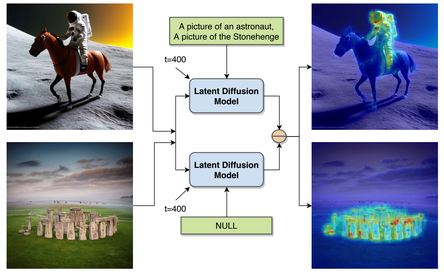

Figure 1. Coarse segmentation results from an LDM for two distinct images, demonstrating the encoding of fine-grained object-level semantic information within the model's internal features.

Observations and Contributions

- **z-space** in which the LDM operates is a compact and semantics-preserving → enables synthesis across several domains such as AI art, illustrations, cartoons etc.
- **Internal representations of a pretrained LDM** which are used to generate photorealistic images for various objects on the internet, also encode powerful visual linguistic semantic information

We want to exploit these two properties of an LDM and improve the text-based image segmentation on real and AI-generated imagery.

Visual-Linguistic information in LDM

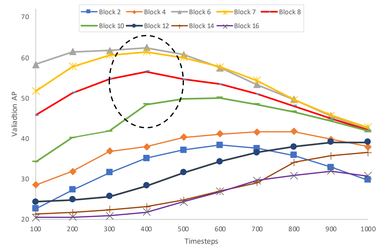

Figure 2. Semantic information present in the LDM features at various blocks and timesteps for the text-based image segmentation task. AP is measured on a small validation subset of the PhraseCut dataset.

LD-ZNet architecture for text-based segmentation

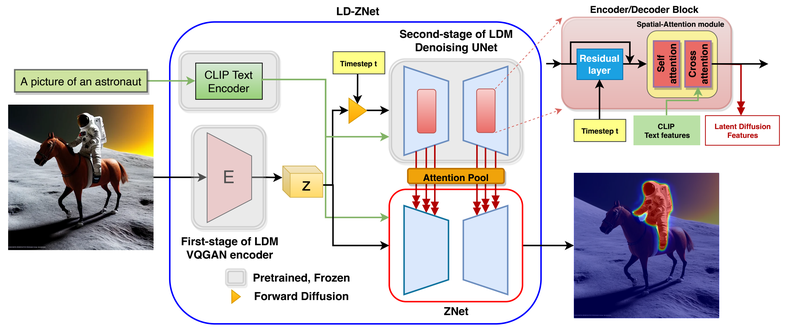

Figure 3. Overview of the proposed ZNet and LD-ZNet architectures. We propose to use the compressed latent representation z as input for our segmentation network ZNet. Next, we propose LD-ZNet, which incorporates the latent diffusion features at various intermediate blocks from the LDM's denoising UNet, into ZNet.

Text-based segmentation performance on phrasecut and AIGI datasets

Method	mIoU	IoU_{FG}	AP
MDETR	53.7	-	-
GLIPv2-T	59.4	-	-
RMI	21.1	42.5	-
Mask-RCNN Top	39.4	47.4	-
HulaNet	41.3	50.8	-
CLIPSeg (PC+)	43.4	54.7	76.7
CLIPSeg (PC, D=128)	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**

(a) PhraseCut

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

(b) Generalization to AIGI dataset

Multi-object segmentation

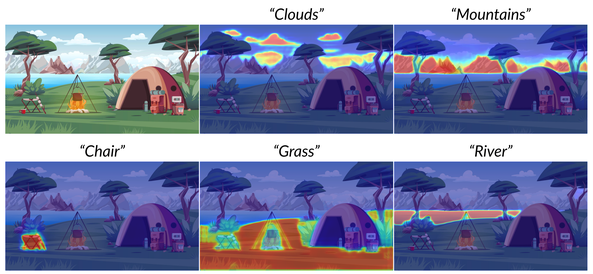

Figure 4. LD-ZNet text-based segmentation results for a diverse set of things and stuff classes. High-quality segmentation across multiple object classes suggests that LD-ZNet has a good understanding of the overall scene.

Segmentation on AI-Generated Images (AIGI)

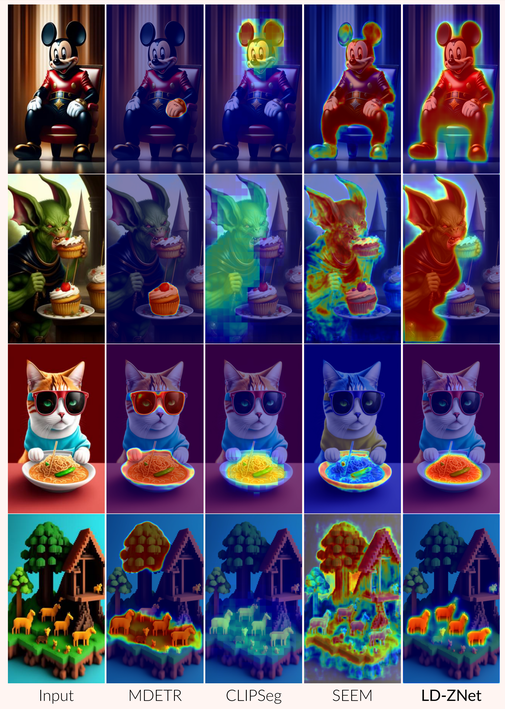

Figure 5. The text prompts are *"Mickey mouse"*, *"Goblin"*, *"Ramen"* and *"Animals"*.

Segmentation on other imagery

Figure 6. RGBNet fails to localize *"Panda"* in the animation image (top row), famous celebrity *"Scarlett Johansson"* (second row) and *"Lamp"* from illustrations (bottom row). LD-ZNet benefits from using z combined with the internal LDM features to correctly segment all of these images.

More predictions on AIGI

Project & AIGI dataset

SCAN ME