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

1. Text-Based Image Segmentation

In this supplementary work, we illustrate some more qualitative text-based image segmentation results using the proposed LD-ZNet model on a diverse set of images. Specifically, we focus on segmenting and localizing 1) objects described by their attributes 2) objects in AI-generated images from the AIGI dataset and 3) multiple different things and stuff in a scene. We also perform a visual comparison with the RGBNet baseline on a diverse set of images, including examples from the PhraseCut test dataset.

1.1. Attributes

Figure 1 depicts attribute-based segmentation. Specifically, objects described by attributes based on color or relative properties (such as height) or actions are well segmented by LD-ZNet.

1.2. AI-Generated Images

We show more qualitative comparisons on our AIGI dataset in Figure 2. We observe similar trend. While MDETR fails to segment the text prompts *"Spiderman"*, *"tortoise"*, *"vespa"* and *"robot"* due to novel concepts and domain gap, CLIPSeg estimates a rough segmentation on the most discriminative regions with lower confidence. However, LD-ZNet accurately segments in all the cases.

1.3. Qualitative Comparisons on Diverse Domains

Figure 3 demonstrates some specific cases where RGBNet fails to segment or poorly segments the object being referred to, where as LD-ZNet segments the objects better.

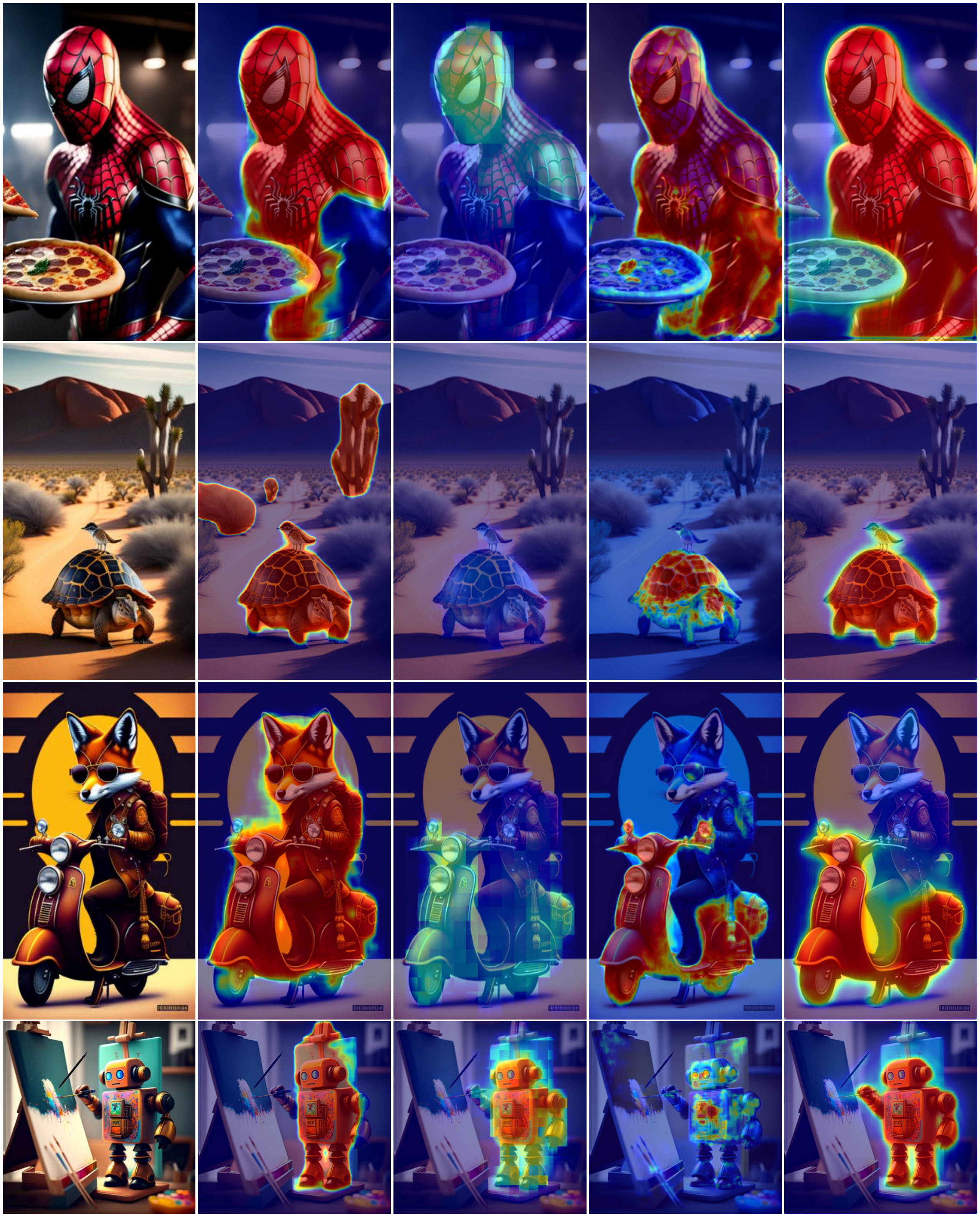
1.4. Scene understanding

Figure 4 shows the segmentation performance of LD-ZNet for several objects and regions in an image. Specifically, we show the segmentation for stuff classes such as "Clouds", "Mountains", "Chair", "Grass", "River" *etc*. and thing classes such as "Trees", "Bicycle", "Sofa", "Books" *etc*. The quality of the segmentation across multiple object classes suggests that LD-ZNet has a good understanding of the overall scene.

Figure 1: Qualitative results showing LD-ZNet correctly segmenting attribute based queries. Text prompts for objects based on color attribute (top row), relative property and action attributes (bottom row) are well segmented by LD-ZNet.

1.5. Qualitative Comparisons on Phrasecut

Figure 5 shows qualitative comparisons of ZNet and LD-ZNet with the RGBNet baseline on the test dataset of PhraseCut. Attributes such as "grey", "glass", "tallest", and "riding" are well understood and localized in LD-ZNet.



(a) Input **(b)** MDETR **(c)** CLIPSeg **(d)** SEEM **(e)** LD-ZNet

Figure 2: Qualitative comparison on the AI-generated images from AIGI dataset for text-based segmentation. The text prompts are *"Spiderman"*, *"tortoise"*, *"vespa"* and *"robot"* respectively.

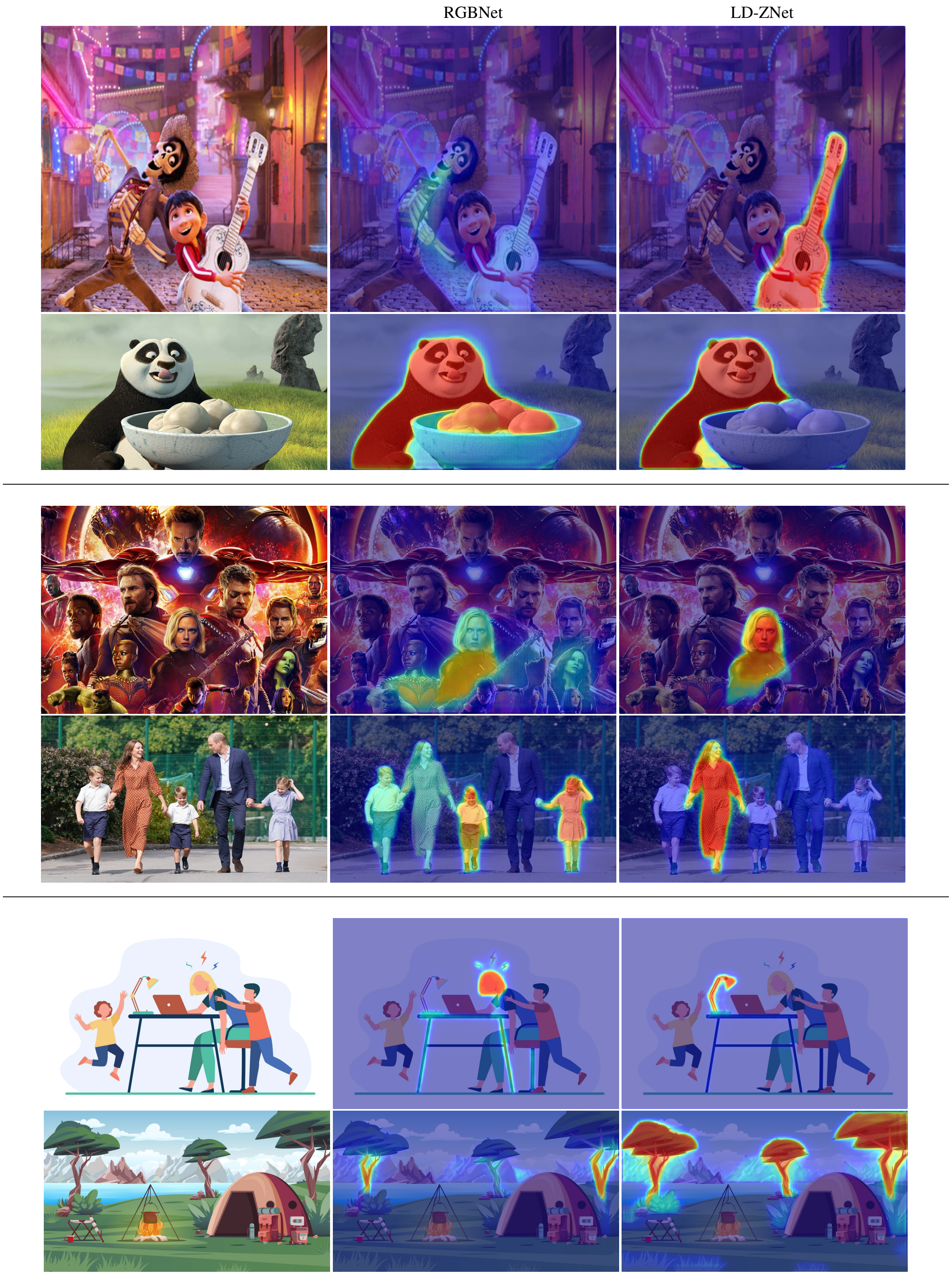


Figure 3: More qualitative examples where RGBNet fails to localize *"Guitar"*, *"Panda"* from animation images (top row), famous celebrities *"Scarlett Johansson"*, *"Kate Middleton"* (second row) and objects such as *"Lamp"*, *"Trees"* from illustrations (bottom row). LD-ZNet benefits from using z combined with the internal LDM features to correctly segment these text prompts.

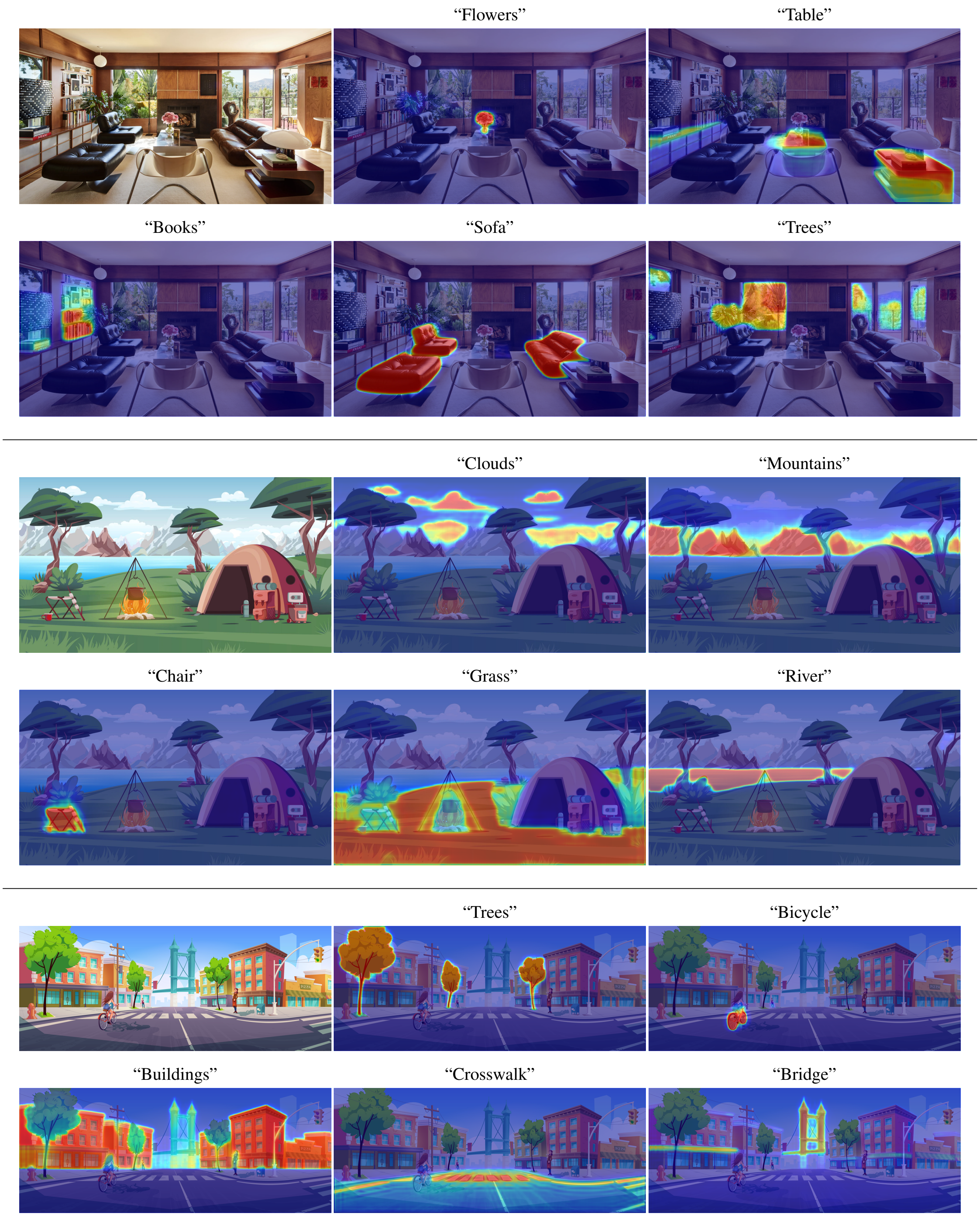
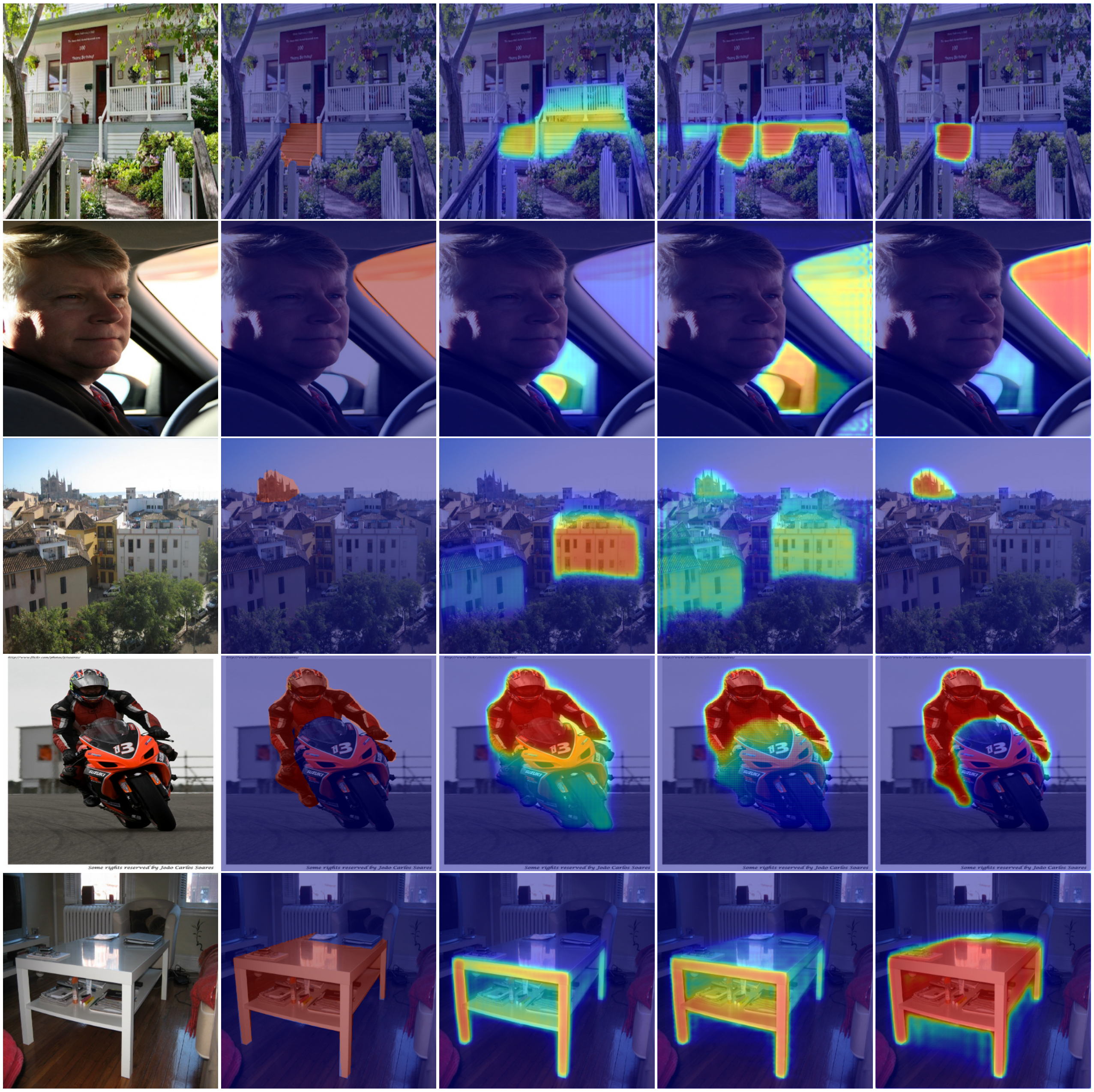


Figure 4: LD-ZNet text-based segmentation results for a diverse set of things and stuff classes across both real images (top row) and illustrations (middle and bottom rows). High-quality segmentation across multiple object classes suggests that LD-ZNet has a good understanding of the overall scene. Images used from google and freepik.



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

Figure 5: Qualitative comparisons on the PhraseCut test dataset. Each row contains an RGB image along with a reference text as an input, with the goal being to segment out the image regions corresponding to the reference text. The reference texts are *"grey steps"*, *"glass windshield"*, *"the tallest building"*, *"riding person"*, *"white stand"* for rows 1, 2, 3, 4 and 5 respectively. We show improvements using ZNet and LD-ZNet compared to the RGBNet.