# Zero-shot spatial layout conditioning for text-to-image diffusion models Supplementary Material

Guillaume Couairon*
Meta AI, Sorbonne Université

Marlène Careil*
Meta AI, LTCI, Télécom Paris, IP Paris

Matthieu Cord
Sorbonne Université, Valeo.ai

Stéphane Lathuilière
LTCI, Télécom Paris, IP Paris

Jakob Verbeek
Meta AI

## A. Societal impact

Our work advances the capabilities of generative image models, contributing to the democratization of creative design by offering tools for non-expert users. Generative image models, however, also pose risks, including using these tools to generate harmful content or deep-fakes, or models generating images similar to the training data which may contain personal data. These concerns have led us to steer away from using large-scale open-source generative image models trained on datasets scraped from the web, for which the licensing of the content is not always clear and which may contain harmful content. Instead, we trained models on a large in-house curated dataset which mitigates these concerns as far as possible.

## B. Implementation details

**Implementation details.** For all experiments that use our LDM diffusion model, we use 50 steps of DDIM sampling with classifier-free guidance strength set to 3. Regarding the construction of the text prompts, we follow [**?**] and concatenate the annotated prompt of COCO with the list of class names corresponding to the input segments. For instance, for the fourth example in Fig. **??**, the conditioning prompt would be "A person jumping a horse over a box. horse, fence, tree".

For our experiments with MultiDiffision we used the Huggingface demo code released by the authors, and only replaced the U-Net noise estimate with our text-to-image model instead of Stable Diffusion.

**Computation of metrics.** We compute FID with InceptionV3 and generate 5k images. The reference set is the original COCO validation set, and we use code from [**?**].

To compute the mIoU metric we use ViT-Adapter[**?**] as segmentation model rather than the commonly used DeepLabV2 [**?**], as the former improves over the latter by 18.6 points of mIoU (from 35.6 to 54.2) on COCO-Stuff. Each generated image is segmented by this model, and the mIoU metric is computed w.r.t. the ground-trutch segmentation mask.

All methods, including ours, generate images at resolution $512 \times 512$, except OASIS and SDM, for which we use available pretrained checkpoints synthesizing images at resolution $256 \times 256$, which we upsample to $512 \times 512$.

The evaluation protocol in our paper is similar but not the same as the one of MultiDiffusion. Our closest setting to MultiDiffusion is *Eval-few*, but it still differs in the validation set size (1k for MultiDiffusion, 5k for our method) and number of selected objects which is between 2 and 4 for MultiDiffusion, between 1 and 3 for our method (as used by SpaText).

For all evaluations of baseline methods that are not based on our text-to-image model, we report the results provided at https://cdancette.fr/diffusion-models/. Stable Diffusion-based baselines, like PwW, also use 50 steps of DDIM sampling, but with a classifier-free guidance of 7.5.

## C. Additional ablation experiments

For these additional ablation experiments, we use the *Eval-few* setting as presented in the paper, where $1 \leq K \leq 3$ spatial masks are used for conditioning.

**Attention layers used.** We first validate which layers are useful for computing our classifier guidance loss in Table 1. We find that whatever the set of cross-attention layers used for computing loss, the mIoU and FID scores are very competitive. In accordance with preliminary observations, it is slightly better to skip attention maps at resolution 8 when computing our loss.

*These authors contributed equally to this work.

| Layers used | ↓FID | ↑mIoU | ↑CLIP |
|---|---|---|---|
| All layers | 33.74 | 40.17 | 30.19 |
| Only decoder layers | 33.81 | 40.02 | 30.05 |
| Only encoder layers | 30.98 | 38.24 | 30.67 |
| Only res32 layers | **29.35** | 39.49 | **30.75** |
| Only res16 layers | 33.59 | 40.27 | 30.23 |
| res16 and res32 layers (ours) | 31.53 | **43.34** | 30.44 |

Table 1. Ablation on cross-attention layers used for estimating segmentation maps.

| Normalization | ↓FID | ↑mIoU | ↑CLIP |
|---|---|---|---|
| No normalization | 30.77 | 38.99 | 30.70 |
| L2 norm | **28.57** | 36.39 | **31.27** |
| L1 norm | 28.85 | 39.74 | 31.04 |
| $L_\infty$ norm (ours) | 31.53 | **43.34** | 30.44 |

Table 2. Impact of gradient normalization scheme on performance.

**Gradient normalization.** We validate the impact of normalizing gradient when applying classifier guidance with our $\mathcal{L}_{\text{Zest}}$ loss. Results are in Table 2.

**Impact of parameter $\tau$.** In our method, classifier guidance is only used in a fraction $\tau$ of denoising steps, after which it is disabled. Table 3 demonstrates that after our default value $\tau = 0.5$, mIoU gains are marginal, while the FID scores are worse. Conversely, using only $10\%$ or $25\%$ of denoising steps for classifier guidance already gives very good mIoU/FID scores, better than PwW for $\tau = 0.25$. As illustrated in Sec. D, this is because estimated segmentation maps converge very early in the generation process.

| Components | ↓FID | ↑mIoU | ↑CLIP |
|---|---|---|---|
| $\tau = 0.1$ | 30.54 | 34.25 | **31.18** |
| $\tau = 0.25$ | **30.36** | 40.75 | 30.77 |
| $\tau = 0.5$ | 31.53 | 43.34 | 30.44 |
| $\tau = 1$ | 34.75 | **44.58** | 29.99 |

Table 3. Ablation on parameter $\tau$, with fixed learning rate $\eta = 1$ in the *Eval-few* setting.

**Tokens used as attention keys.** Our estimated segmentation masks are computed with an attention mechanism over a set of keys computed from the text prompt embeddings. In this experiment, we analyze whether the attention over the full text-prompt is necessary, or whether we could simply use classification scores over the set of classes corresponding to the segments. We encode each class text separately with the text encoder, followed by average pooling to get a single embedding per class (in contrast to summing the attention values in Eq. (3) in the main paper). We still condition the diffusion model on the main caption but don't concatenate the class names to the prompt. But we use the all COCO class text embeddings to compute our loss. Computing our loss with these embeddings as attention keys results in a probability distribution over the segmentation classes. We find that the FID scores are worse (+ 3 pts FID), but the mIoU scores are very close (43.36 *vs*. 43.34). We conclude that our loss function primarily serves to align spatial image features with the relevant textual feature at each spatial location, and that the patterns that we observe in attention maps are a manifestation of this alignment.

## D. Additional visualizations

**Evolution of attention maps across timesteps.** We show in Fig. 1 and Fig. 2 average attention maps on the different objects present in the input segmentation during the first 12 denoising steps with and without our guidance scheme. We condition on the same Gaussian noise seed in both cases. We notice that attention maps quickly converges to the correct input conditioning mask when we apply ZestGuide and that the attention masks are already close to ground truth masks only after 12 denoising iteration steps out of 50.

**Additional visualizations on COCO.** In Fig. 3 and Fig. 4, we show additional qualitative samples generated with COCO masks comparing ZestGuide to the different zero-shot methods. We use up to three classes per image, corresponding to the *Eval-few* setting.

**Visualizations on hand-drawn masks.** In Fig. 5, we show generations conditioned on coarse hand-drawn masks, a setting which is closer to real-world applications, similar to Fig. 2 in the main paper. In this case the generated objects do not exactly match the shape of conditioning masks: the flexibility of ZestGuide helps to generate realistic images even in the case of unrealistic segmentation masks, see *e.g.* the cow and mouse examples.

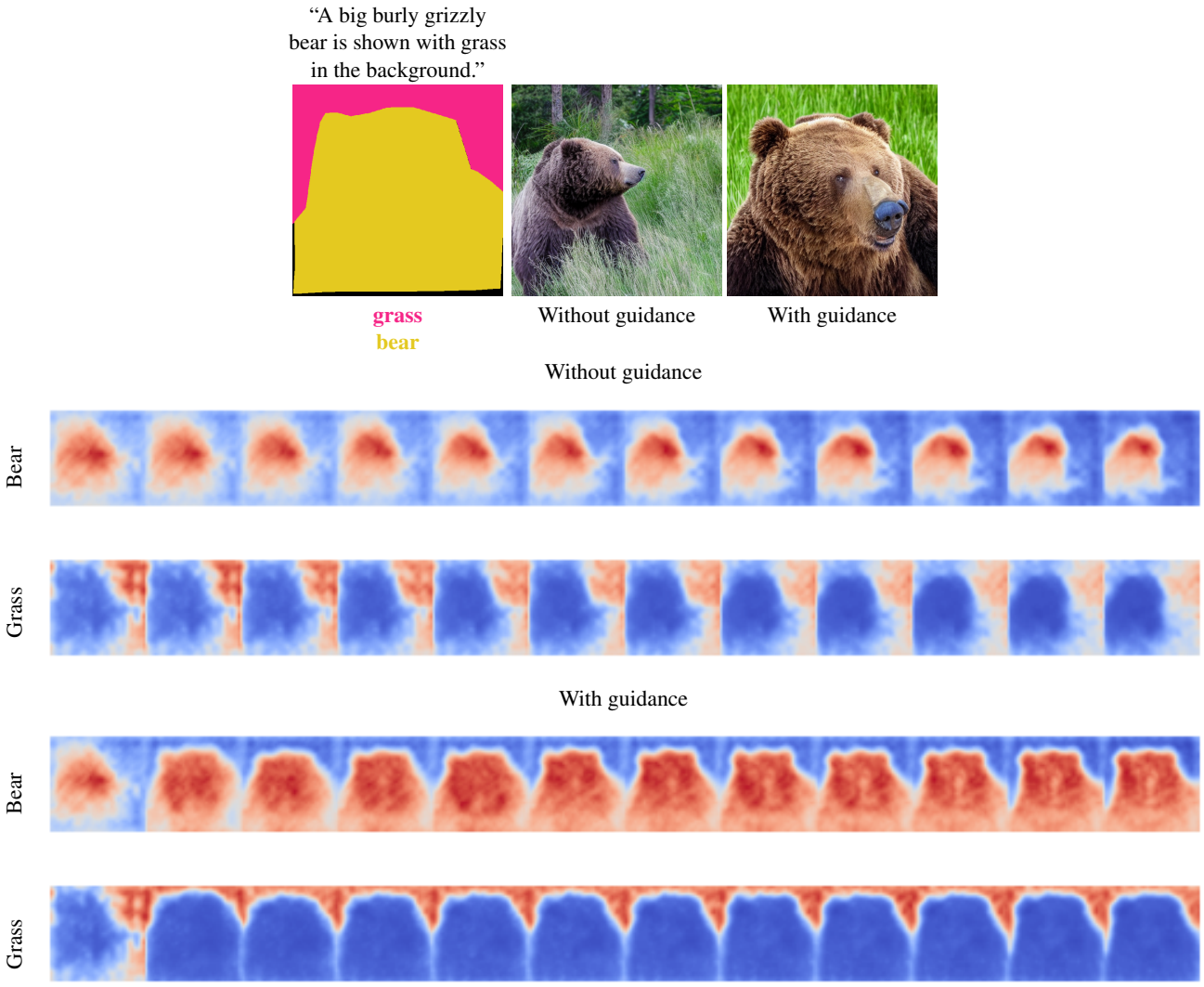


Figure 1. Visualization of first 12 denoising steps out of 50 steps. Same seed used with and without guidance.

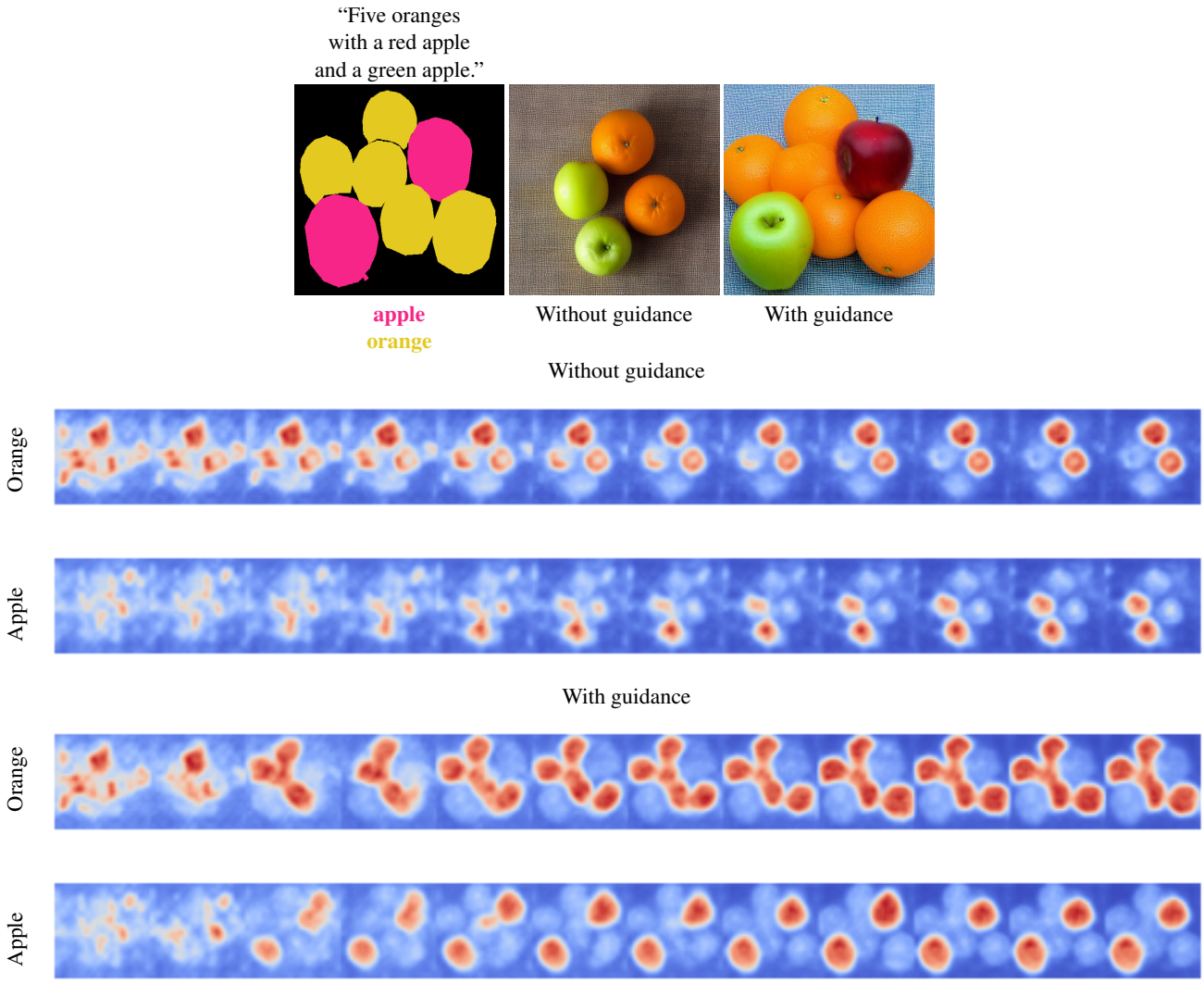


Figure 2. Visualization of first 12 denoising steps out of 50 steps. Same seed used with and without guidance.

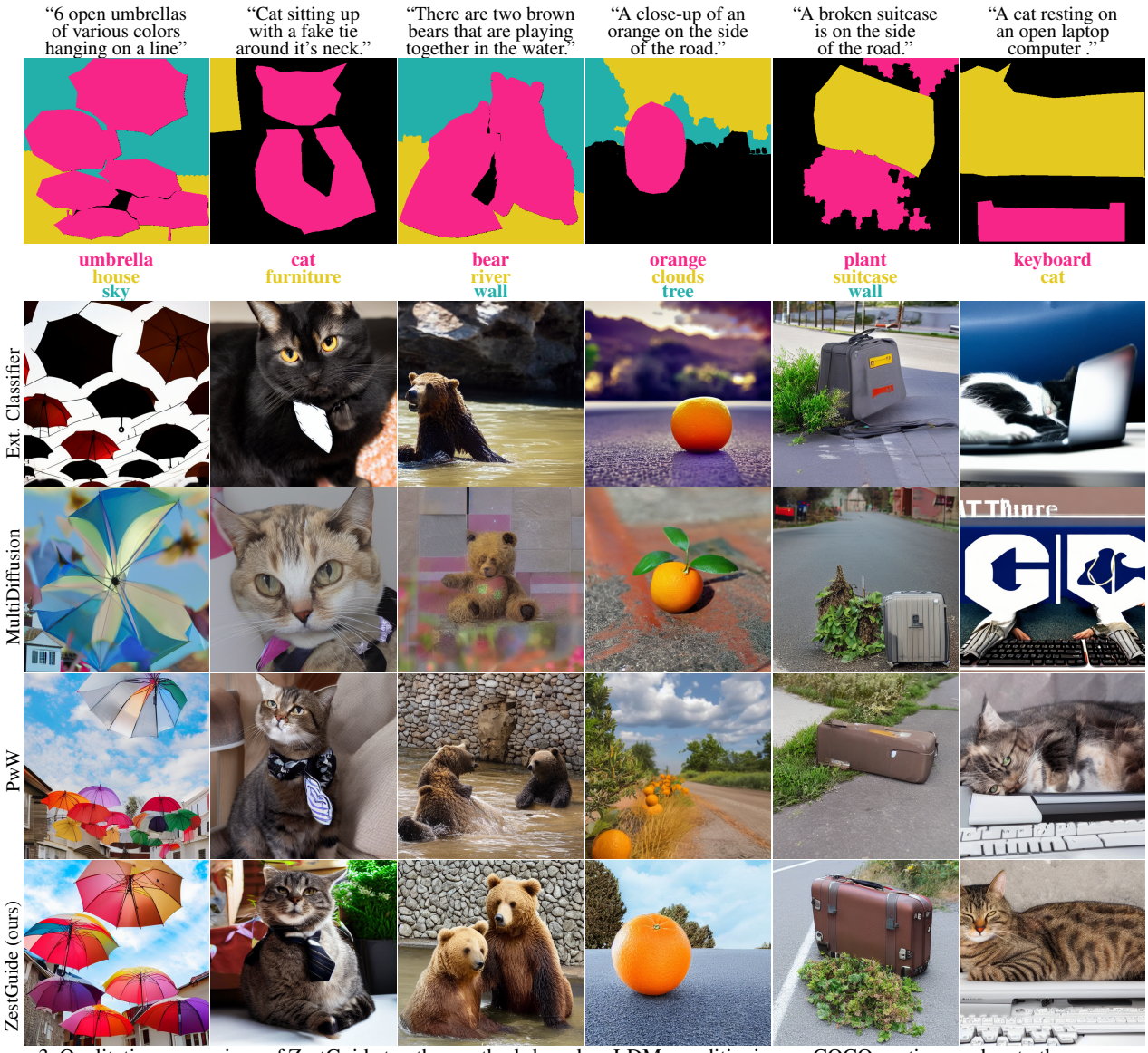


Figure 3. Qualitative comparison of ZestGuide to other methods based on LDM, conditioning on COCO captions and up to three segments.

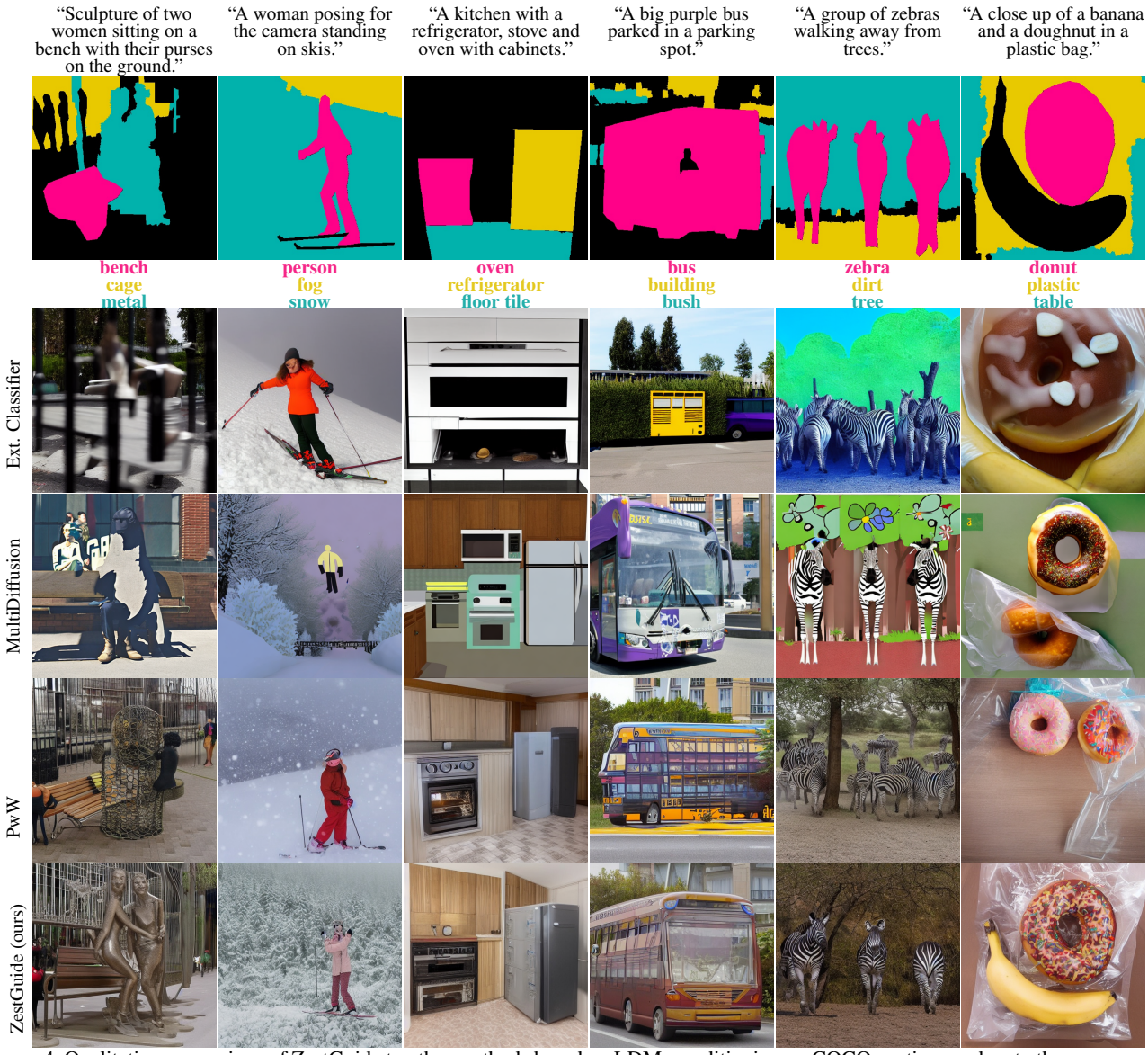


Figure 4. Qualitative comparison of ZestGuide to other methods based on LDM, conditioning on COCO captions and up to three segments.


Figure 5. ZestGuide generations on coarse hand-drawn masks.