PARTCONCEPTS: A Unified Mechanism for Fine-Grained Part Localization and Generation

Vaibhav Agrawal[1] Varghese P Kuruvilla[1] Harsh Rangwani[2] Ravi Kiran S[1]
[1]IIIT Hyderabad [2]Adobe Research, Bengaluru

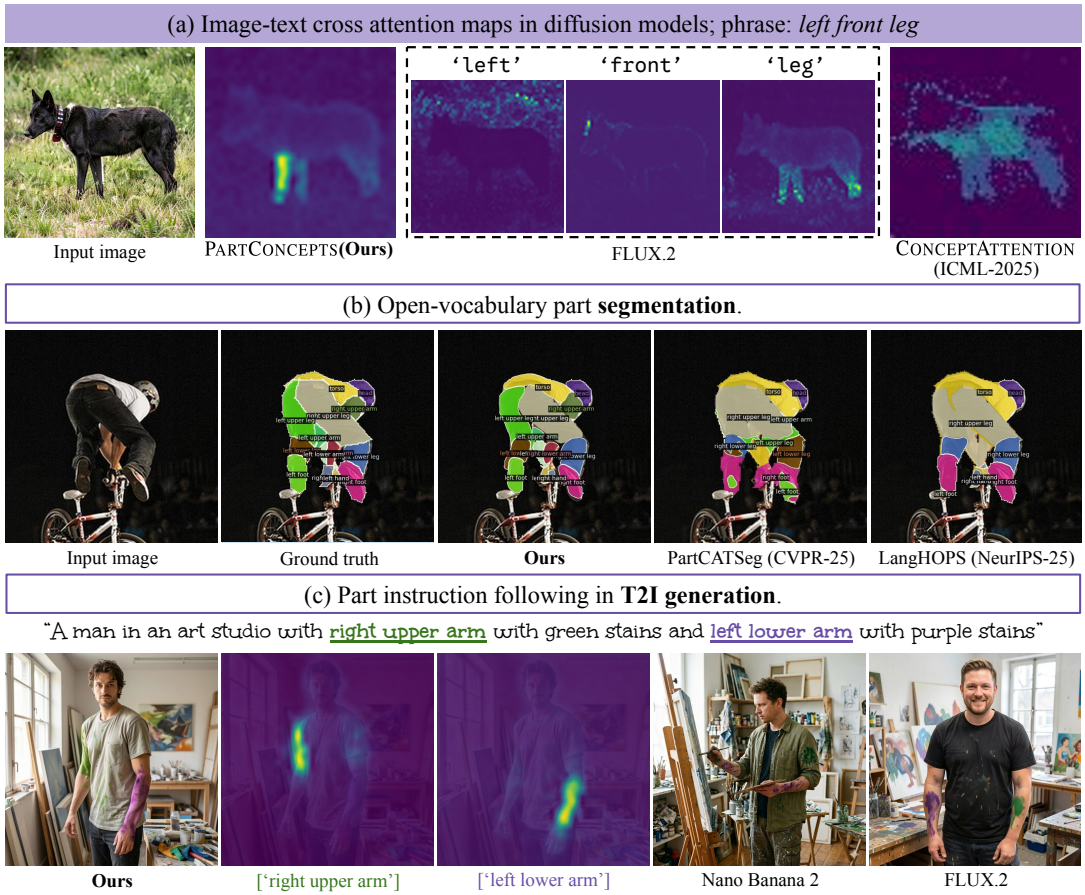

Figure 1. **PARTCONCEPTS.** Cross attention maps of state-of-the-art diffusion models [24] reveal that they lack fine-grained part understanding, shown in (a). To address this, we propose PARTCONCEPTS, a unified mechanism for fine-grained part localization (see b.) and generation (see c.).

Abstract

While text-to-image (T2I) diffusion models exhibit strong semantic disentanglement at the object level, they struggle to localize fine-grained object parts. Although the latent visual features in these models are sufficiently fine-grained, the text–image interaction (cross-attention) fails to effectively exploit this information. This results in coarse and ambiguous localization, particularly for spatially distinct components such as 'left' and 'right' limbs. To address this limitation, we introduce, **PARTCONCEPTS***, a mechanism that encodes textual part descriptions into compact, learnable representations. These* PARTCONCEPT *tokens are trained to selectively attend to their corresponding part regions in the image. To evaluate the part-level understanding of* PARTCONCEPT *tokens, we first probe its effectiveness for part segmentation. We then evaluate whether the improved localisation enables fine-grained instruction following in T2I generation. On standard Open-Vocabulary Part Segmentation (OVPS) benchmark, our method outperforms dedicated segmentation baselines, establishing a new state-of-the-art. We further evaluate our method on the more complex Pascal-Part benchmark for part instance segmentation which contains fine-grained spatial annotations (e.g. 'left lower arm'). Here,*

we outperform all baselines by a very significant margin. Finally, our qualitative and quantitative results show that this strong localization of PARTCONCEPT *tokens directly enables fine-grained instruction following in T2I generation: the textual attributes (e.g., color) inherently bind well to corresponding* PARTCONCEPT *tokens, without explicit attribute control mechanisms. This indicates that* PARTCONCEPT *tokens compose seamlessly with other text tokens, highlighting their applicability for fine-grained text-based generative control.*

1. Introduction

Text-to-image (T2I) diffusion models possess strong semantic priors that enable precise generation and editing [2, 22]. Beyond generation, recent work [21] shows that multi-modal Diffusion Transformer (mmDiT) [15, 41] cross-attention maps can serve as competitive object-instance segmentors. These results indicate that the diffusion latent space achieves effective semantic disentanglement at the object level. However, since objects are fundamentally defined by their constituent parts, any system aiming for part-level fine-grained reasoning and precise generative control must rely on semantic part-level representations. This raises a critical question: Does the image-text alignment of diffusion models actually extend to the part-level? Parts are uniquely challenging because quite often a single object contains multiple instances of the same component (e.g., 'left ear' vs. 'right ear'). Distinguishing these requires an understanding of object-centric spatial organization: the relative positioning of parts within the object's own frame of reference [35, 54].

An investigation of the image-text attention maps for state-of-the-art models like FLUX.2 [24] and CONCEPTATTENTION [21] reveals that they are coarse and poorly localized for fine-grained part descriptions (see fig. 1(a)). This poor localization manifests as a failure in downstream instruction following. As shown in fig. 1(c), even state-of-the-art models fail to follow prompts targeting fine-grained parts. **Hence, we conclude that the image-text alignment in diffusion models, while strong at the object level, does not persist at the part level**. To identify failure cause, we analyze FLUX.2 and make two key observations:

Observation O1. The latent visual features of diffusion models are semantically very fine-grained, as evidenced by their ability to support highly accurate fine-grained image-to-image pointwise semantic correspondence tasks [16, 19, 33, 47, 54, 55]), with minimal training supervision.

Observation O2. The text-image interaction (cross-attention) acts as a bottleneck; it fails to leverage the high-resolution latent visual features, resulting in coarse attention maps that cannot distinguish between spatially distinct parts like the 'left' and 'right' limbs, see fig. 1(a).

These observations imply that while the latent features possess sufficient resolution (O1), the text-image interaction acts as a bottleneck (O2). Therefore, our core research question is: **can we improve text aligned part understanding of a T2I model, while preserving its strong priors?** To address this, we introduce PARTCONCEPTS. Given a text prompt, we encode each textual part description (e.g., 'right upper arm') into a single PARTCONCEPT token embedding: this results in a structured token sequence, e.g., 'A photo of an artist with PARTCONCEPT['right upper arm'] ...', which is passed to the mmDiT. During training, this PARTCONCEPT token is optimized to attend to the corresponding part region in the image, providing the fine-grained grounding that base models lack. As we show in our results, our proposed method effectively generalizes to novel part phrases, object classes and prompts beyond what is seen in the training data.

To evaluate the part-level understanding of PARTCONCEPT tokens, we first demonstrate our method's effectiveness for *part segmentation*. We then show that this strong localization directly enables precise part-level attribute binding, thus providing *fine-grained instruction following* in T2I generation. Thus, our PARTCONCEPTs provide a unified mechanism for both visual grounding and generative control.

In the setting of *part segmentation*, we extract the part segmentation maps directly from PARTCONCEPT attention maps, using only a single forward pass through the mmDiT network; this enables efficient inference without the need for iterative sampling. PARTCONCEPTS outperforms dedicated segmentation baselines on the standard Open-Vocabulary Part Segmentation (OVPS) [52] benchmark, establishing state-of-the-art. However, we find that OVPS is too coarse for precise grounding: it typically removes spatial descriptors and collapses distinct instances such as 'left hand' and 'right hand' into a single 'hand' instance. Therefore, we evaluate for *part instance segmentation* using the original Pascal-Part [5] benchmark, which retains these fine-grained part instances and corresponding spatial descriptions. On this more rigorous benchmark, our method significantly outperforms all baselines and establishes a new state-of-the-art.

Finally, we utilize the effective localization of PARTCONCEPTs to enable *fine-grained instruction following in text-to-image generation*. We find that our localization-focused training inherently enables the model to bind attributes to specific parts, without a specialized architectural mechanism for attribute control. For example, when provided with a prompt such as 'an artist with PARTCONCEPT['right upper arm'] colored green ...' (see fig. 1(c)), the model successfully binds the color attribute to the correct part region during the generative process. These results demonstrate that PARTCONCEPTs preserve the compositional nature of text space, while maintaining the high-fidelity priors of the base T2I generative model. Our model enables fine-grained instruc-

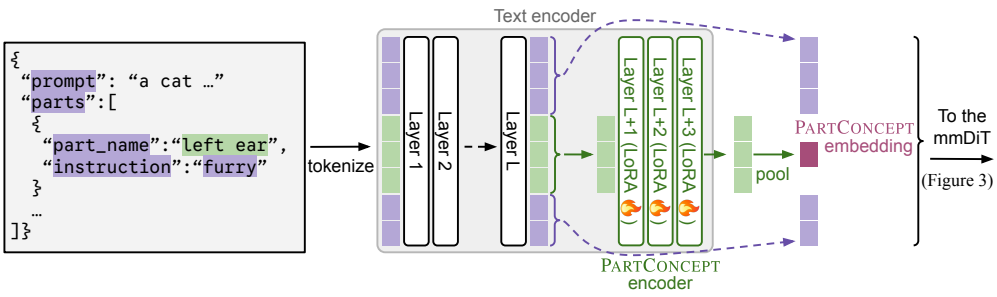

Figure 2. **Encoding textual part descriptions into PartConcepts.** Given an input text prompt (left) with part phrases, our method encodes these textual part phrases (e.g. 'left ear') into PartConcept embeddings, which are trained to attend to corresponding image regions. We extend the text encoding in the base T2I model to encode the PartConcepts: the base T2I model uses intermediate text encoder outputs (at layer L) for conditioning the mmDiT; we use three subsequent text-encoder layers (L+1, L+2, L+3) for forming our PartConcept encoder. We only train a LoRA on these subsequent layers; this design ensures that the output PartConcept does not drift too much from the text embedding space, thus preserving the T2I generative capabilities (see Sec. 4.3).

tion following to part descriptions not seen during training, indicating its strong generalization and prior preservation.

2. Related work

Semantic concepts in text-to-image (T2I) diffusion models. Numerous studies have demonstrated that text-to-image (T2I) diffusion models encode strong semantic priors. These priors have also been shown to be effective for downstream tasks such as semantic correspondence [14, 47, 55]. A line of research works introduce learnable prompt tokens, and optimize them to attend to specific visual concepts in the image [1, 19, 20, 30]. However, these works require training a separate embedding for new every visual concept, hence not scalable. Further, the learnt prompt tokens are typically not very composable with the text embeddings, hence reducing their applicability for localized attribute binding in T2I generation. A more scalable approach is to use the open vocabulary priors of the T2I models for learning visual concepts. Recent work, ConceptAttention [21] has shown the applicability of mmDiT cross attention maps for object-instance segmentation; however, it fails for fine-grained part segmentation (see fig. 2(a)).

Part instance segmentation in images is a long standing problem in computer vision [13, 26, 29, 34, 45, 49, 56], with various datasets proposed for this task [5, 18, 28, 43]. Among these, Pascal-Part [5] stands out for its fine-grained part descriptions (e.g. 'left ear', 'right front upper leg'). Open-Vocabulary Part Segmentation (OVPS) [52] is the standard benchmark used by recent methods [8, 10, 27, 37, 40, 51]; however, OVPS is too coarse, as it removes spatial descriptors and collapses distinct instances such as 'left hand' and 'right hand' into a single 'hand' instance. Among recent works, LangHOPS [37] parses object hierarchies using an MLLM to generate queries which are decoded into part-level segmentation masks. PartCATSeg [10] introduces a cost aggregation framework with a compositional loss and structural guidance using DINO [4]. In contrast, our work leverages strong priors of generative models using a simple architecture to achieve fine-grained open-set part segmentation.

Part level instruction following in T2I generation. Part level instruction following has been explored in prior work [44], which uses self attention clustering followed by attribute injection through cross attention. However, it assumes image-text cross attention to be highly localized, which is violated in many cases, especially for fine-grained part descriptions. While there are works which focus on learning part-level concepts from images, and compose them for creative generations [30, 39], they do not support text-based instruction following, hence not directly relevant. Recent work [12] uses learnable part concepts for localized image editing. However, it requires an embedding optimization procedure for every new part concept, hence not scalable. Further, it can only enable editing in a provided reference image. In contrast, our method directly enables instruction following in text-to-image generation, which is challenging due to absence of a reference structure.

3. Method

Given an input text prompt with part phrases (e.g., 'left ear') (see fig. 2 left), our goal is to ensure that the part phrases attend strongly to the respective part regions in the image, enabling part-level localization and consequently, fine-grained instruction following in T2I generation.

As observed in FLUX.2 cross attention maps fig. 1(a), the part phrase is tokenized into discrete tokens, which attend coarsely to disparate regions in the image. Spatial descriptors in part phrase ('left', 'right', etc.) particularly have very diffused attention maps. As a result, the part description as a whole does not localize strongly to any specific region in the image.

To address this, we encode the text embeddings corresponding to each part phrase into a PartConcept embed-

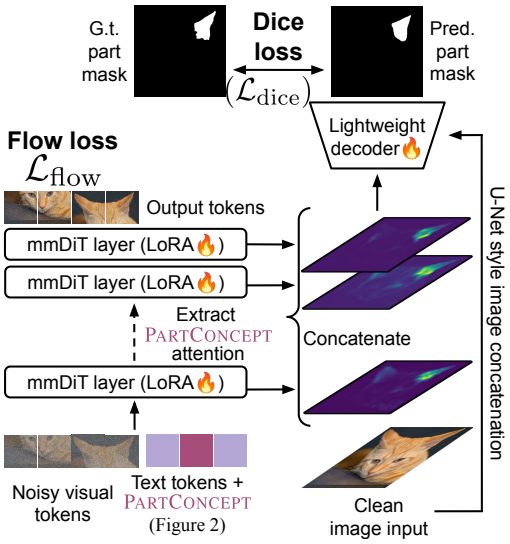

Figure 3. **The localization objective.** We pass the interleaved text and PartConcept tokens see fig. 2) through the text branch of the mmDiT model. During forward pass through the mmDiT, we extract layer-wise cross attention maps of the PartConcept tokens. These attention maps, along with clean input image are decoded into precise part segmentation maps using a multi-scale CNN decoder, trained using dice loss. **The generative objective.** To preserve the generative capability of the T2I model, we also include the flow loss (denoising MSE), enabling fine-grained *part-level instruction following* in T2I generation.

ding of the same dimension (see fig. 2). This results in a structured token sequence, where the discrete part phrase tokens are replaced by corresponding PartConcept token (see fig. 2, right). This token sequence is then passed into the multi-modal diffusion transformer (mmDiT), along with the noisy visual tokens (see fig. 3). To ensure that the PartConcept embedding does not drift too much from the text embedding space, we initialize the PartConcept encoder with successive layers of the base T2I text encoder, and only train a LoRA [23] over it (see fig. 2). Our parameter-efficient training preserves generative prior of the base T2I model, as discussed in Sec. 4.4.

The localization objective. Our aim is to ensure that the PartConcept token attends to the correct part region in the image. For this, we use Dice segmentation loss [38] to localize the PartConcept attention maps in the correct part region. A straightforward way is to apply the Dice loss directly on the layer-wise PartConcept attention maps; however, such an approach presents several limitations. First, the attention maps in mmDiT are computed using the compressed latents, hence they lack the resolution required for dense part mask prediction at the original image resolution. Second, directly using attention maps for downstream segmentation is quite brittle, as it necessitates extra hyperparameters for attention map thresholding and layer selection. To address this, we introduce a lightweight decoder, which decodes concatenated layer-wise PartConcept attention maps into a segmentation mask (see fig. 3 top-right). The decoder consists of stacked convolutions and up-sampling layers, followed by a sigmoid activation at the end. The clean input image is also introduced at every resolution in the decoder; this provides necessary cues for making precise segmentation boundaries.

To infer part-level segmentation masks given an input image and query part phrase(s), we first encode the query part phrase(s) into PartConcept embedding(s), as shown in fig. 2. Next, we encode the input image into the diffusion latent space and add a fixed amount of noise to it. We pass these noisy visual latents and the text tokens (the PartConcept tokens) into a single forward pass through the mmDiT, extract PartConcept attention maps and decode them into part segmentation maps (see fig. 3).

Preserving the generative capability. The setup described above ensures strong localization of the PartConcept tokens to the correct part regions. However, to enable high fidelity T2I generation, it is crucial that generative capability of the base model is preserved. Therefore, in addition to the localization loss, we also include the flow loss (denoising MSE), which preserves the generative model's prior. The final training objective is a weighted sum of the localization and the flow objectives $\mathcal{L} = \mathcal{L}_{\text{dice}} + \lambda \cdot \mathcal{L}_{\text{flow}}$. We find that the flow loss weight λ establishes a tradeoff between generation quality and localization performance, further discussed in Sec. 4.4.

Training dataset. To train our model, we leverage an existing part-level segmentation dataset, Pascal-Part [5], considering its fine-grained annotations with spatial identification of part instances. However, training on Pascal-Part images presents a significant hurdle: the low resolution and skewed aspect ratios of these images can degrade the high-fidelity priors of modern generative models. To address this, we apply an automated restoration technique using FLUX.2 for upsampling and outpainting the images, while preserving the original annotations. Training on this enhanced dataset, dubbed 'Pascal Part SR' (SR = Super-Resolution), ensures compatibility with the high visual quality of state-of-the-art generative models [24].

Attribute binding. While the mechanisms described above ensure that generative capability of the base T2I model is preserved, there is still one final requirement for accurate *part-level instruction following*: the PartConcept tokens must compose well with other text tokens. For instance, in a prompt specifying different colors for the 'right upper arm' and 'left lower arm' (Fig. 1c), the model must automatically bind each attribute to its respective PartConcept. In other words, the PartConcept embeddings should preserve the compositional nature of the text embeddings. To achieve this, we augment training prompts with part-level captions gener-

Method	ADE20K [52]			PPS-116 [52]			PartImageNet [18]		
	seen	unseen	h-IoU	seen	unseen	h-IoU	seen	unseen	h-IoU
VLPart [46]	*	*	*	42.6	18.7	26.0	*	*	*
ZSSeg+ [36]	43.2	27.8	33.9	54.4	19.0	28.2	*	*	*
CLIPSeg [32]	38.2	30.9	34.2	48.9	27.5	35.2	53.9	37.2	44.0
CAT-Seg [7]	33.8	25.9	29.3	43.8	27.7	33.9	47.3	35.1	40.3
PartCLIPSeg [9]	38.4	38.8	38.6	50.0	31.7	38.8	56.3	51.7	53.9
PartGLEE [27]	51.3	35.3	41.8	57.4	27.4	37.1	*	*	*
PartCATSeg [11]	53.1	47.2	50.0	57.5	44.9	50.4	73.8	71.5	72.7
LangHOPS [37]	49.3	49.7	49.5	59.2	46.5	52.1	71.9	**73.7**	72.8
Ours	**53.4**	**61.9**	**57.3**	**65.0**	**49.7**	**56.3**	**75.4**	70.6	**72.9**

Table 1. **Zero-shot evaluation results.** Our method outperforms dedicated segmentation baselines on Open-Vocabulary Part Segmentation (OVPS) [52] benchmark, establishing state-of-the-art.

Method	PartImageNet (OOD) [18]	PartImageNet → PPS-116 [52]
	Unseen	Unseen
CLIPSeg [32]	55.86	14.87
CAT-Seg [7]	39.12	12.50
PartCLIPSeg [9]	59.16	19.86
PartCATSeg [11]	66.15	22.88
Ours	**67.30**	**30.00**

Table 2. **Cross-dataset evaluation results.** We consider two settings: PartImageNet (OOD) [18] and PartImageNet (train) → PPS-116 [52] (eval).

ated by passing cropped image regions through a VLM [53]. Notably, this augmentation implicitly enables attribute binding in T2I generation, without explicit architectural mechanisms for attribute control.

4. Experiments

We first evaluate our method's effectiveness on part-level localization using the standard *part segmentation* [52] (Sec. 4.1) and *part instance segmentation* [5] benchmarks (Sec. 4.2). We then show that the proposed PARTCONCEPTs' strong localization directly enables fine-grained *part-level instruction following* in T2I generation (Sec. 4.3). Note that we use the Pascal-Part SR data (Sec. 3) only for experiments on T2I generation (Sec. 4.3); for part segmentation we train on the respective benchmarks only.

4.1. Open vocabulary part segmentation (OVPS) [52]

Implementation and training details. We use FLUX.2 Klein 9B [24] as the base model for all experiments in this paper. For the PARTCONCEPT encoder, we train a LoRA (rank 128) over the text-encoder layers (see fig. 2). Additionally, we train a smaller LoRA (rank 4) for the mmDiT, following standard practice. We train the models until convergence, at batch size 16, using default optimizer settings from HuggingFace Diffusers [50].

Method	Pascal-Part (org.) [5]		
	seen	unseen	h-IoU
PartCATSeg [11]	33.2	35.5	34.3
LangHOPS [37]	38.4	38.7	38.5
Ours	**55.2**	**54.2**	**54.7**

Table 3. **Fine-grained part *instance* segmentation**, with spatial identification of part instances (e.g. 'left lower arm').

Datasets and evaluation. Following OVPS [52], we evaluate two setups: *zero-shot*, where train and test categories do not overlap, and *cross-dataset*, where evaluation is performed on a dataset different than the one used for training, hence testing OOD generalization. In the context of *zero-shot* part segmentation, we use Pascal-Part-116 (PPS-116) [52], ADE20K-Part-234 [52] and PartImageNet [18] datasets. Following [10], we evaluate *cross-dataset* transfer for 1) PartImageNet (training) to PPS-116 (evaluation), and 2) PartImageNet (OOD), a split of the PartImageNet dataset which was proposed originally for few-shot part segmentation [18]. Similar to [37], we use the oracle-object setting [52], where ground truth object segmentation and class are provided. Following prior work [8, 10, 37, 52], we use the IoU metric, reported for both seen and unseen categories separately, as well as their harmonic mean (h-IoU).

Baseline comparison. We compare against standard part segmentation baselines; recent works include PartCATSeg [10] and LangHOPS [37]. Evaluation results are presented in Tab. 1 (*zero-shot*) and Tab. 2 (*cross-dataset*). Our method outperforms dedicated segmentation baselines in almost all categories, establishing state-of-the-art. We visualize segmentation results in fig. 8.

4.2. Fine-grained part *instance* segmentation [5]

While our method is competitive on (OVPS) [52] benchmark, we find that the OVPS suite is too coarse for precise grounding: it typically removes spatial descriptors and collapses

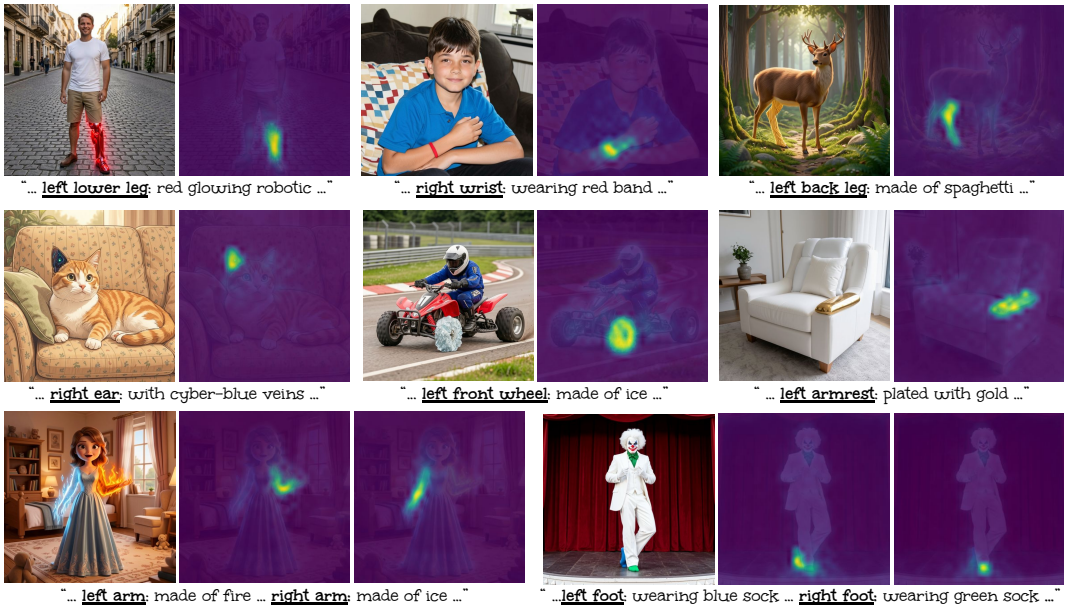

Figure 4. **Qualitative results for *part-level instruction following*, with PARTCONCEPT attention maps.** Our method is able to disambiguate spatial identifiers in part descriptions, (e.g., 'left front wheel'). Notably, PARTCONCEPTs generalize to novel part phrases; specifically, ***armrest, left front wheel* and *wrist* are not seen in our training data**, indicating strong generalization. Even in case of multiple parts, the attributes are correctly bound to respective PARTCONCEPTs; this demonstrates that PARTCONCEPT embeddings preserve the compositional nature of text embeddings.

distinct instances such as 'left hand' and 'right hand' into a single 'hand' instance. Therefore, we evaluate for part *instance* segmentation using the original Pascal-Part [5] benchmark, which retains these fine-grained part instances and corresponding spatial descriptions. On this more rigorous benchmark, our method significantly outperforms dedicated perception baselines and establishes a new state-of-the-art. We visualize segmentation results in fig. 8.

4.3. Fine-grained instruction following

Implementation and training details. We use the same settings as described in Sec. 4.1. Additionally, to preserve the generative capability of the base T2I model, we set the loss balancing coefficient $\lambda = 0.25$ (Sec. 3).

Evaluation dataset and metrics. For evaluation, we construct an evaluation set of 500 text prompts with fine-grained part-level instructions. For this, we utilize the text labels in Pascal-Part dataset: for each example, we randomly sample a part label (from the exhaustive list of labels provided by Pascal-Part), introduce a randomized instruction, such as 'stain with `{color}`' to make a text prompt. The resulting text prompts follow the template 'a photo of `{class_name}`, with `{instruction}`'. We use this dataset to quantitatively evaluate our method for *part-level instruction following* along two axes: 1) spatial localization (is the attribute applied to the correct part?) and 2) attribute following (is the attribute correctly followed?). Spatial localization is difficult to evaluate using current text-based

Figure 5. **Qualitative comparisons.** Our method exhibits strong spatial localization, while baselines frequently struggle, failing to disambiguate spatial identifiers in part descriptions (e.g., 'left back wheel', 'right ear', etc.).

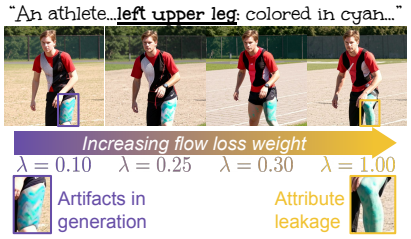

Figure 6. **Effect of increasing flow loss weight** (λ). Low values lead to visual artifacts, while high values cause attribute leakage.

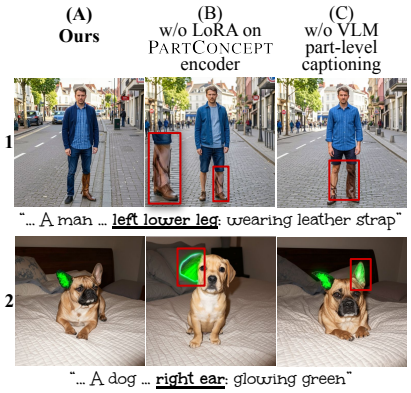

Figure 7. **Ablations. B.** LoRA on the PartConcept is crucial for preserving generation fidelity. **C.** Training prompt augmentation using VLM captioning improves attribute binding.

grounding models [3, 6, 31], due to their lack of fine-grained part understanding [37]; hence we perform a user study for measuring spatial localization. The study consists of A/B type binary choice questions. Overall, we collect 2040 responses from 40 users. For measuring attribute following, we use CLIP [42] similarity between text prompts (containing the attribute) and generated images.

Baselines. We compare our method against our base T2I model FLUX.2 Klein [24], Nano Banana 2 [17, 48], and FLUX.1 [dev] [25].

Qualitative results. As shown in fig. 4, the PartConcept strongly localizes to correct part regions, shown by the sharp attention maps, while preserving T2I generation fidelity. Our PartConcept encoder exhibits strong generalization to novel part phrases (e.g., ***armrest***, ***left front wheel***, ***wrist*** **(fig. 4) and** ***handle***, ***tyre*** **(fig. 5) are not present in our training data**). Further, our method effectively leverages the T2I generative prior to generalize to arbitrary attributes: none of the shown part attributes in figs. 4 and 5 were seen in our training data. Notably, correct attribute binding even in multi-part scenarios (third row, fig. 4) demonstrates that PartConcept embeddings preserve compositional nature of text embeddings.

Baseline comparison. We present qualitative comparisons (see fig. 5), user study (see fig. 9) and CLIP evaluation (see Tab. 4). The qualitative results and the user study highlight that our method exhibits strong spatial localization, while baselines consistently fail to disambiguate spatial identifiers in part descriptions (e.g. 'left front leg', 'right ear'). Further, FLUX.1 [25] fails to localize the attribute; instead, it applies the attribute to the entire object or image (fig. 5(D)). The CLIP scores demonstrate that our method is competitive in attribute following compared to the base models. Interestingly, FLUX.2 has higher CLIP score than all other models; this is because the generated images zoom in to the specific part mentioned in the prompt (see fig. 5 2-4B); as a result, the generated attribute is much more prominent, but with low spatial localization accuracy, highlighted in fig. 9. Ultimately, our method enhances part-instance understanding of the base T2I model (FLUX.2);

Model	CLIP sim.
FLUX.1	21.25
N.B. 2	23.53
FLUX.2	**26.07**
Ours	25.00

Table 4. **CLIP similarity score.** Evaluation of attribute following.

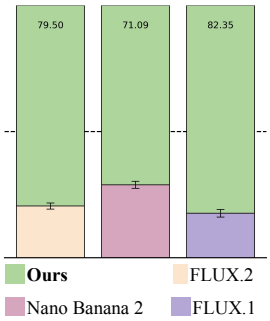

Figure 9. **User study.** Evaluation of part-level spatial localization.

4.4. Ablations

We validate our approach through several ablative experiments. First, the flow loss weight (λ) balances localization and denoising objectives during training (see fig. 6). We find that λ dictates a tradeoff between generation fidelity and attribute binding (see fig. 6); low values degrade the generative prior, causing visual artifacts, while high values reduce localization pressure, leading to attribute leakage. Second, LoRA finetuning for PartConcept encoder (see fig. 2) is critical; full fine-tuning causes significant visual artifacts (highlighted with red boxes in fig. 7(B)),

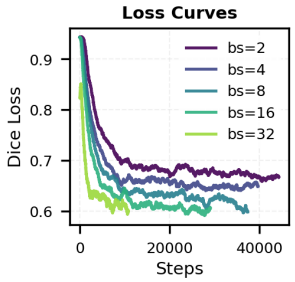

Figure 10. **Effect of increasing flow loss weight** (λ). Flow loss preserves the generative prior; low flow loss weight fails to preserve the generative prior, causing visual artifacts (left), while higher values fail to ensure strong attribute binding, due to reduced localization pressure.

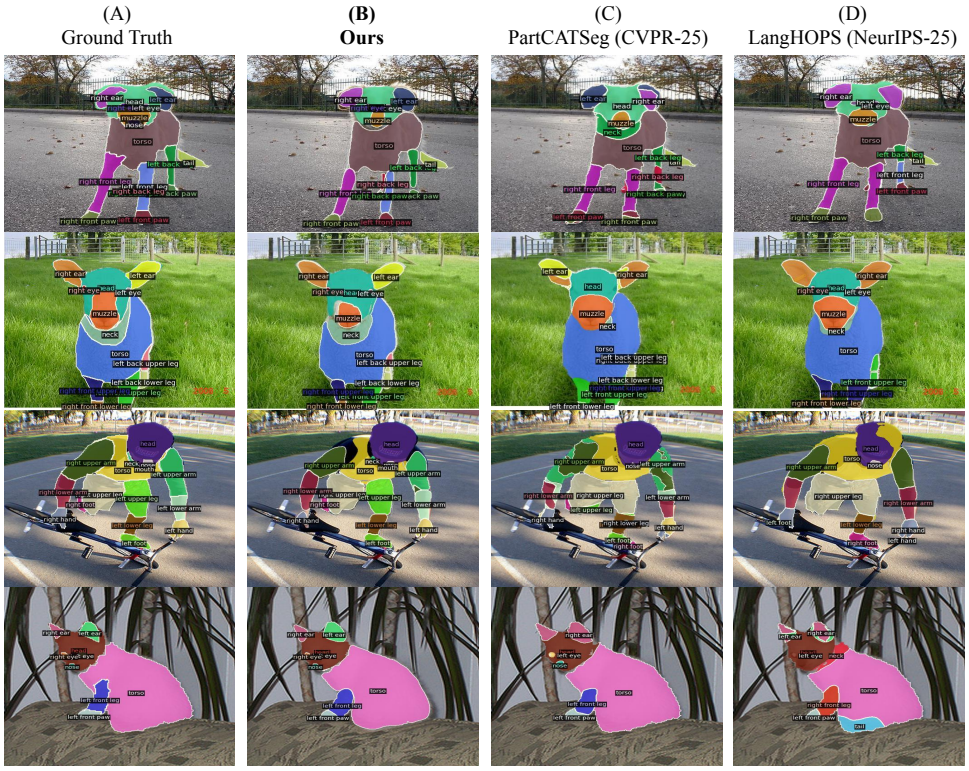

Figure 8. **Qualitative comparison.** Our method outperforms all baselines on various benchmarks, establishing state-of-the-art. Notably, we are significantly better for *part instance segmentation* [5], effectively disambiguating spatially distinct part instances like 'left lower arm' and 'right upper arm'.

as PARTCONCEPT embedding drifts too much from the text space, disrupting the T2I prior. Third, our VLM-based prompt augmentation is essential for attribute binding; without it, the training lacks incentive to associate specific attributes with PARTCONCEPTs, as shown in fig. 7(C). Finally, we find batch size significantly impacts localization performance. Unlike standard FLUX.2 [24] fine-tuning scripts [50] that use small batches, our localization loss trajectories significantly improve with larger batch sizes (fig. 10, we also expect improvements with scaling).

5. Limitations

Despite its effectiveness, PARTCONCEPTS faces two primary limitations. First, scaling beyond four part-attribute pairs occasionally causes attribute leakage, even when spatial localization is accurate. This likely stems from a distributional shift, as prompts with more than four attributes are out-of-distribution (OOD) relative to our training data. Second, while we provide a unified mechanism for segmentation and T2I part-level control, the best-performing loss weights (λ) differ for each task. This reflects a slight trade-off between discriminative grounding and generative prior preservation. Future research should try to mitigate this gap by designing the model in a way that it leads to optimal discriminative localization capabilities without having any tradeoff on generative prior preservation.

6. Conclusion

This work stems from observing that while T2I cross attention maps are strongly disentangled for objects, however this semantic clarity collapses at the part-level, particularly for spatially distinct part instances (e.g., 'left front leg' vs. 'right front leg'). We identify the bottleneck as text-image interaction (cross attention) rather than the visual feature resolution. Hence, we propose a lightweight mechanism to correct this: encoding part phrases (e.g. 'left front leg') into a compact representation: the PARTCONCEPT tokens, trained for spatial grounding. Our approach effectively leverages the base T2I's generative priors to achieve state-of-the-art performance on various *part (instance) segmentation* benchmarks, outperforming dedicated perception baselines. We further show that this effective localization directly enables robust attribute binding and *fine-grained instruction following* in T2I generation. These results demonstrate the effectiveness of the proposed approach for improving part-level understanding of generative models.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple

concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3

[2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7877–7888, 2025. 2

[3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 7

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 3

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3, 4, 5, 6, 8

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024. 7

[7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 5

[8] Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 37:137161–137189, 2024. 3, 5

[9] Jiho Choi, Seonho Lee, Seungho Lee, Minhyun Lee, and Hyunjung Shim. Understanding multi-granularity for open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 37:137161–137189, 2024. 5

[10] Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9782–9793, 2025. 3, 5

[11] Jiho Choi, Seonho Lee, Minhyun Lee, Seungho Lee, and Hyunjung Shim. Fine-grained image-text correspondence with cost aggregation for open-vocabulary part segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9782–9793, 2025. 5

[12] Aleksandar Cvejic, Abdelrahman Eldesokey, and Peter Wonka. Partedit: fine-grained image editing using pre-trained diffusion models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 3

[13] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5485–5494, 2021. 3

[14] Olaf Dünkel, Thomas Wimmer, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. Do it yourself: Learning semantic correspondence from pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5834–5844, 2025. 3

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2

[16] Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[17] Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026. Accessed: April 29, 2026. 7

[18] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 3, 5

[19] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 36:8266–8279, 2023. 2, 3

[20] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22820–22830, 2024. 3

[21] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. *arXiv preprint arXiv:2502.04320*, 2025. 2, 3

[22] 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

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1 (2):3, 2022. 4

[24] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025. 1, 2, 4, 5, 7, 8

[25] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 7

[26] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Semi-local affine parts for object recognition. In *British Machine Vision Conference (BMVC'04)*, pages 779–788. The British Machine Vision Association (BMVA), 2004. 3

[27] Junyi Li, Junfeng Wu, Weizhi Zhao, Song Bai, and Xiang Bai. Partglee: A foundation model for recognizing and parsing any objects. In *European Conference on Computer Vision*, pages 475–494. Springer, 2024. 3, 5

[28] Xiao Li, Yining Liu, Na Dong, Sitian Qin, and Xiaolin Hu. Partimagenet++ dataset: Scaling up part-based models for robust recognition. In *European Conference on Computer Vision*, pages 396–414. Springer, 2024. 3

[29] Xiangtai Li, Shilin Xu, Yibo Yang, Haobo Yuan, Guangliang Cheng, Yunhai Tong, Zhouchen Lin, Ming-Hsuan Yang, and Dacheng Tao. Panoptic-partformer++: A unified and decoupled view for panoptic part segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):11087–11103, 2024. 3

[30] Junyu Liu, R Kenny Jones, and Daniel Ritchie. Partcomposer: Learning and composing part-level concepts from single-image examples. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025. 3

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 7

[32] 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 (CVPR)*, pages 7086–7096, 2022. 5

[33] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2

[34] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3

[35] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19521–19530, 2024. 2

[36] Fangyun Wei Yutong Lin Yue Cao Han Hu Mengde Xu, Zheng Zhang and Xiang Bai. A simple baseline for open vocabulary semantic segmentation with pre-trained vision-language model. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2022. 5

[37] Yang Miao, Jan-Nico Zaech, Xi Wang, Fabien Despinoy, Danda Pani Paudel, and Luc Van Gool. Langhops: Language grounded hierarchical open-vocabulary part segmentation. *arXiv preprint arXiv:2510.25263*, 2025. 3, 5, 7

[38] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[39] Kam Woh Ng, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Partcraft: Crafting creative objects by parts. In *European Conference on Computer Vision*, pages 420–437. Springer, 2024. 3

[40] Tai-Yu Pan, Qing Liu, Wei-Lun Chao, and Brian Price. Towards open-world segmentation of parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15392–15401, 2023. 3

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[42] 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. 7

[43] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 3

[44] Harsh Rangwani, Aishwarya Agarwal, Kuldeep Kulkarni, R Venkatesh Babu, and Srikrishna Karanam. Composing parts for expressive object generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13209–13219, 2025. 3

[45] Rishubh Singh, Pranav Gupta, Pradeep Shenoy, and Ravikiran Sarvadevabhatla. Float: Factorized learning of object attributes for improved multi-object multi-part scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1445–1455, 2022. 3

[46] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023. 5

[47] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[49] Andrea Vedaldi, Siddharth Mahendran, Stavros Tsogkas, Subhransu Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, David Weiss, et al. Understanding objects in detail with fine-grained attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3622–3629, 2014. 3

[50] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and

Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5, 8

[51] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36:21429–21453, 2023. 3

[52] Meng Wei, Xiaoyu Yue, Wenwei Zhang, Shu Kong, Xihui Liu, and Jiangmiao Pang. Ov-parts: Towards open-vocabulary part segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 3, 5

[53] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5

[54] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. 2023. 2

[55] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023. 2, 3

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3