Learning 3D Texture-Aware Representations for Parsing Diverse Human Clothing and Body Parts

Kiran Chhatre[1,2], **Christopher E. Peters**[1], **Srikrishna Karanam**[2]
[1]KTH Royal Institute of Technology
[2]Adobe Research
{chhatre, chpeters}@kth.se, skaranam@adobe.com

Abstract

Existing methods for human parsing into body parts and clothing often use fixed mask categories with broad labels that obscure fine-grained clothing types. Recent open-vocabulary segmentation approaches leverage pretrained text-to-image (T2I) diffusion model features for strong zero-shot transfer, but typically group entire humans into a single *person* category, failing to distinguish diverse clothing or detailed body parts. To address this, we propose Spectrum, a unified network for part-level pixel parsing (body parts and clothing) and instance-level grouping. While diffusion-based open-vocabulary models generalize well across tasks, their internal representations are not specialized for detailed human parsing. We observe that, unlike diffusion models with broad representations, image-driven 3D texture generators maintain faithful correspondence to input images, enabling stronger representations for parsing diverse clothing and body parts. Spectrum introduces a novel repurposing of an Image-to-Texture (I2Tx) diffusion model—obtained by fine-tuning a T2I model on 3D human texture maps—for improved alignment with body parts and clothing. From an input image, we extract human-part internal features via the I2Tx diffusion model and generate semantically valid masks aligned to diverse clothing categories through prompt-guided grounding. Once trained, Spectrum produces semantic segmentation maps for every visible body part and clothing category, ignoring standalone garments or irrelevant objects, for any number of humans in the scene. We conduct extensive cross-dataset experiments—separately assessing body parts, clothing parts, unseen clothing categories, and full-body masks—and demonstrate that Spectrum consistently outperforms baseline methods in prompt-based segmentation.

Extended version — https://s-pectrum.github.io/

Introduction

Human parsing (Li et al. 2022c; Fang et al. 2018) decomposes a human image into distinct parts such as clothing items and body regions. This task provides rich descriptions for human-centric visual analysis (Lin et al. 2019) and has become increasingly important in applications such as person re-identification (Ye et al. 2021), human-object interaction (Peng et al. 2025), and portrait segmentation (Wang et al. 2024). A detailed understanding of human body structure and the vast diversity in apparel is critical for advancing fields that analyze human appearance in various contexts, especially in complex scenes like those in fig. 1 (top), where the human alphabet forming our model name, Spectrum, shows individuals wearing a range of formal, casual, and traditional attire.

Existing parsing methods typically rely on fixed, predefined categories. Although these are sufficient for human body parts, they struggle to handle the large variability and complexity of clothing in real world contexts, often producing broad or incorrect labels (e.g., labeling all clothing in the torso as '*upper clothes*' (Khirodkar et al. 2024)). Clothing fashion evolves rapidly, with new styles emerging while older ones become less common but still relevant, and accessories (e.g., *hats*, *bags*, *belts*, *jewelry*), as well as traditional attire (e.g., *police uniforms*, *sarees*), add further variation. Although recent high-quality annotated datasets (Li et al. 2024) capture this diversity, most current parsing architectures still work with a fixed set of labels, making it difficult to fully accommodate new or unseen clothing categories.

Emerging open-vocabulary approaches leverage diffusion models trained on large-scale text–image datasets, demonstrating strong compositional generalization and semantic control (Rombach et al. 2022). However, they fail to distinguish body parts or fine-grained clothing, instead labeling the entire human as a single '*person*' mask in datasets such as COCO (Mottaghi et al. 2014a), because their broad representations are not specialized for detailed human parsing. Effective human parsing requires both part-level segmentation—capturing diverse clothing and distinct body parts—and instance-level grouping (e.g., distinguishing each individual's hands in multi-person scenes). Recent multimodal text and image driven 3D texture generation models show faithful correspondence to input modalities (Tu et al. 2025). We observe that these strong 3D texture-based features can significantly enhance segmentation of diverse human clothing and body parts. To address existing limitations in fine-grained human parsing, we leverage diffusion models fine-tuned on high-quality 3D human texture maps from the ATLAS dataset (Liu et al. 2024), enabling accurate segmentation of diverse clothing categories and body parts.

In Spectrum, we utilize these specialized internal representations that capture body parts and clothing features more

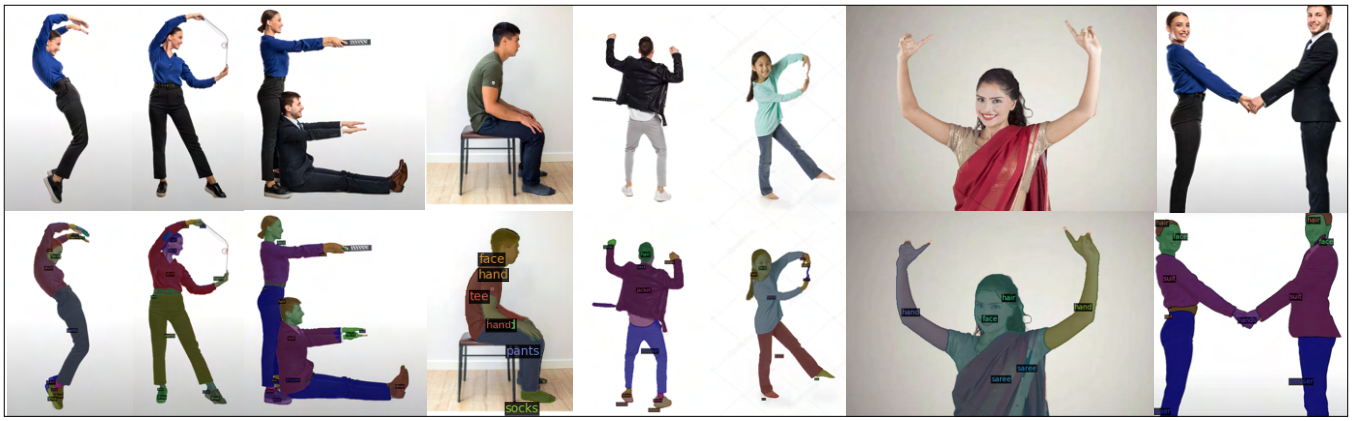

Figure 1: Human-alphabet parsing in the wild. Internet-video frames show people forming the letters of Spectrum, while the model segments diverse clothing and body parts, handling unseen items such as *sarees*, across complex poses and styles.

effectively than standard diffusion model representations. We leverage the large-scale CosmicManHQ dataset (Li et al. 2024), which contains detailed real-world images with text annotations describing human clothing and body parts, and focus exclusively on parsing on-body clothing, ignoring standalone garments or irrelevant items. Specifically, our model takes an image and a descriptive caption as input and outputs segmentation maps with semantic annotations for every visible body part and clothing category in scenes with any number of humans. To the best of our knowledge, our unified network is the first to offer this capability, which current human parsing and open-vocabulary segmentation methods lack. We use ground-truth masks for broad body-part categories and generic garment classes from CosmicManHQ to supervise the model, producing class-agnostic masks aligned with scene-specific prompts—enabling the parsing of both seen and unseen clothing categories. We validate our method through four evaluation setups: FPP (Full-Person Parsing, merging each person into one mask), BHP (Bare Human Parsing, separating each visible body part), CCP (COCO Category Parsing, focusing on human-relevant COCO accessory classes), and COP (Clothing-Only Parsing, isolating each clothing instance). Experiments on CosmicManHQ's test set and cross-dataset scenarios demonstrate our model's effectiveness in handling diverse clothing and body parts more effectively than existing human parsing and open-vocabulary segmentation approaches. As shown in fig. 1 (bottom), the model accurately segments formal (*suit*, *trouser*) and casual (*shorts*, *jacket*) attire in in-the-wild images, including unseen traditional clothing (*sarees*).

Our key contributions are: (1) We introduce Spectrum, the first unified human parsing method that performs pixel-level parsing of diverse clothing/body parts and instance-level part grouping in a single network. (2) We repurpose an Image-to-Texture (I2Tx) diffusion model, fine-tuned from a T2I backbone on 3D human texture maps, to align internal features with on-body clothing and parts for precise parsing. (3) Across four setups (FPP, BHP, CCP, COP) and cross-dataset tests, Spectrum consistently outperforms prior methods in multi-person scenes.

Related Work

Human Parsing

Early methods in human parsing focused on multiscale image features (Chen et al. 2018), human pose (Xia et al. 2016), or visual cues as weak supervision (Dai, He, and Sun 2015). Broadly, human parsing can be divided by the number of human instances in the data and input modality: single-human parsing (e.g., (Dickens and Hamad 2025)), multi-human instance-level parsing (e.g., (Liu et al. 2025)), and video-based parsing (e.g., (Gupta et al. 2023)). Our scope is image-based single- and multi-human parsing. In single-human parsing, CDGNet (Liu et al. 2022) accumulates horizontal/vertical human-part distributions with attention for accurate relationship modeling, SCHP (Li et al. 2022c) addresses annotation noise via self-correction, and SAPIENS (Khirodkar et al. 2024) exploits a ViT-based masked autoencoder pretrained on large-scale human images. In multi-human parsing, RP R-CNN (Yang et al. 2020) introduces a semantic-enhanced feature pyramid and parsing re-scoring network, while AIParsing (Zhang et al. 2022) takes a one-stage top-down approach for instance detection and part segmentation. Most previous methods rely on ground-truth image masks with fixed parsing labels, lacking dense annotations of human attributes and real-world complexity (Gong et al. 2017), thus producing coarse results that cannot handle new categories to align masks categorically. We address this gap by pursuing robust fine-grained parsing for body parts and diverse clothing, incorporating instance-level part grouping in a unified network.

Open-Vocabulary Segmentation

Background. Open-domain tasks have gained increasing attention for their ability to handle unseen categories during inference (Zhu and Chen 2024). These include open-set recognition (Geng, Huang, and Chen 2021), open-world recognition (Kim et al. 2022a), and open-vocabulary learning (Wu et al. 2024). Large vision–language models (VLMs) trained with contrastive image and text objectives provide strong zero shot transfer (Jia et al. 2021), and T2I models

further bind region–word correspondences (Ho, Jain, and Abbeel 2020). Below, we review open-vocabulary semantic and instance segmentation methods relevant to our work.

Semantic segmentation. Recent open-vocabulary semantic segmentation methods adapt pretrained VLMs or diffusion models to segment novel classes. ODISE (Xu et al. 2023) uses frozen T2I diffusion features and an implicit captioner, while MaskCLIP (Zheng Ding 2023) modifies CLIP's final attention layer for stronger local semantic consistency. IFSeg (Yun et al. 2023) processes target categories given only label sets, and GD (Li et al. 2023) aligns T2I model's visual and textual embeddings via a grounding module. OVSeg (Liang et al. 2023) fine-tunes CLIP with mask–category pairs to overcome domain gaps, and OVAM (Manchón et al. 2024) generates word-specific attention maps in T2I models without additional training. SED (Xie et al. 2024) applies a hierarchical encoder–decoder for cost-map generation, while SEEM (Zou et al. 2023) employs prompt-based interactive segmentation. OpenSeg (Ghiasi et al. 2022) randomly drops words during training to prevent overfitting. Broadly, some T2I-based methods synthesize labeled images (Li et al. 2022a) or perform image conditioning for mask generation (Ji et al. 2023).

Instance segmentation. GGN (Wang et al. 2022) uses pseudo supervision from learned pixel-level affinities, OLN (Kim et al. 2022b) applies classifier-free object localization, and OV-SAM (Yuan et al. 2024) merges CLIP and SAM for open-vocabulary recognition. Despite the successes of open-vocabulary methods, these approaches rarely address fine-grained human body parts or complex clothing categories, often collapsing people into a single *person* mask. We bridge this gap by adapting open-vocabulary segmentation to detailed body-part and clothing parsing for arbitrary humans. To this end, we adopt a fine-tuned T2I model on human textures, yielding specialized representations that enable accurate segmentation of detailed body parts and diverse clothing.

Method

Overview. We focus on prompt-based human parsing for diverse clothing categories and individual body parts. During training, ground truth masks for 17 generic base categories and dense captions are provided; body part masks are detailed, whereas clothing masks are weakly annotated because the base categories (e.g., *Special Clothings*, *One-piece Outfits*) are broad and require grounding to learn accurate labels. At test time, clothing may include both seen and unseen classes, and we ensure that the test subset explicitly includes unseen categories for separate evaluation.

Repurposing Image-to-Texture Models

Preliminaries. We leverage strong 3D texture-based features with faithful correspondence to input images to enhance segmentation of diverse clothing and body parts. Consequently, we repurpose diffusion features from TexDreamer (Liu et al. 2024), a dual-conditioned (text/image) 3D texture generator, for 2D human parsing. TexDreamer builds on Stable Diffusion (SD) (Rombach et al. 2022) with LoRA, trained on 3D human texture maps paired with images and captions from the ATLAS dataset. It conditions on image features from the IM2CONTEXT encoder, whose LoRA weights and pretrained encoder we reuse in our method.

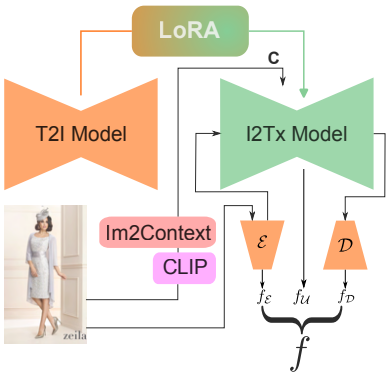

Figure 2: I2Tx feature extraction. Base Stable Diffusion (SD) weights are merged with 3D-texture LoRA matrices, and a context embedding $\mathbf{C}$ from IM2CONTEXT and CLIP-VISION drives a single forward pass in the I2Tx model. Concatenating encoder, denoiser, and decoder representations yields the texture-aligned feature f for clothing and body-part parsing.

I2Tx feature extraction. As shown in fig. 2, our feature-extraction pipeline uses the frozen I2Tx architecture in a single forward pass ($t = 0$), omitting iterative denoising. We first merge the base SD weights $W_{\phi_{\text{SD}}} \in \mathbb{R}^{d\times k}$ with TexDreamer's LoRA matrices $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ (with $r \ll \min(d,k)$), so the adapted projection becomes $W_{\phi_{\text{SD}}} + BA$. An input image x is then encoded by the CLIP-VISION encoder ϕ_{CV} and the IM2CONTEXT encoder ϕ_{I2C} to generate the context embedding $\mathbf{C}$ and the CLS token.

$$\mathbf{C} = \phi_{\text{I2C}}\left\{\phi_{\text{CV}}(x)\right\} \tag{1}$$

The encoder $\phi_{\mathcal{E}}$ produces features $f_{\mathcal{E}}$ and a latent code x_e. A noisy latent is sampled with noise schedule $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$.

$$\begin{aligned} f_{\mathcal{E}}, x_e &= \phi_{\mathcal{E}}(x), \\ x_t &\triangleq \sqrt{\bar{\alpha}_t}\, x_e + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{2}$$

The denoiser ϕ_{SD} is conditioned on $(x_t, \mathbf{C}, \text{CLS})$ to yield f_U, while the decoder $\phi_{\mathcal{D}}$ processes x_t to obtain $f_{\mathcal{D}}$. Concatenating the encoder, denoiser, and decoder streams yields a texture-aligned representation f that drives our segmentation head to parse on-body clothing and body parts.

$$\begin{aligned} f_U &= \phi_{\text{SD}}(x_t, \mathbf{C}, \text{CLS}) \\ f_{\mathcal{D}} &= \phi_{\mathcal{D}}(x_t) \\ f &= f_{\mathcal{E}} \parallel f_U \parallel f_{\mathcal{D}} \end{aligned} \tag{3}$$

Semantically Grounded Parsing

Parsing-head architecture. As shown in fig. 3, a pixel decoder $\mathcal{P}$ (implemented as a feature-pyramid network (Cheng et al. 2022)) converts the texture representation f into

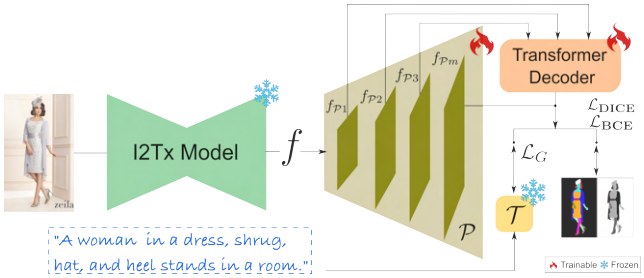

Figure 3: Training. Frozen I2Tx features from image x feed the pixel decoder $\mathcal{P}$ and transformer, producing N class-agnostic masks m optimized with $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{DICE}}$. Key prompt phrases, embedded by frozen $\mathcal{T}$, are contrastively aligned to m via $\mathcal{L}_{\text{G}}$.

multi-scale, high-resolution per-pixel embeddings. Its final mask feature $f_{\mathcal{P}m}$ together with the pyramid outputs $f_{\mathcal{P}1}, f_{\mathcal{P}2}, f_{\mathcal{P}3}$ feeds a transformer decoder that predicts N class-agnostic binary masks $\{m_i\}_{i=1}^{N}$. For each mask m_i, we obtain its embedding via masked average pooling, $z_i = \langle f, m_i \rangle$. We also extract auxiliary mask embeddings and pyramid features in the same manner. Following Cheng, Schwing, and Kirillov (2021), Hungarian matching assigns predictions to ground-truth masks. Class logits are optimized with binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$, and mask shape with Dice loss $\mathcal{L}_{\text{DICE}}$.

Prompt grounding. We train on CosmicMan-HQ, which provides BLIP-generated dense captions containing detailed hair and clothing attributes. For each image–caption pair $\{x^{(m)}, s^{(m)}\}_{m=1}^{B}$ in a batch of size B, we extract K_{phrase} key phrases (relevant nouns and adjectives) from $s^{(m)}$, yielding $\{p_k\}_{k=1}^{K_{\text{phrase}}}$. Each phrase p_k is formatted into a prompt (e.g., "*a photo of a tan purse*" or "*person hair length is above shoulders*") and embedded with OpenCLIP ($\mathcal{T}$). Following (Xu et al. 2023), we apply a contrastive loss that aligns phrase embeddings $\mathcal{T}(p_k)$ with their corresponding mask embeddings z_i. The grounding loss $\mathcal{L}_G$ over the batch is then computed as:

$$g\left(x^{(m)}, s^{(m)}\right) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbf{p}_{ik} \cdot \langle z_i, \mathcal{T}(p_k) \rangle,$$
$$\mathcal{L}_G = -\frac{1}{B} \sum_{m=1}^{B} \log \frac{e^{2 \cdot g_{mm}/\tau}}{\sum_{n=1}^{B} e^{g_{mn}/\tau} \cdot \sum_{n=1}^{B} e^{g_{nm}/\tau}}. \quad (4)$$

Here, $\mathbf{p}_{ik}$ is the softmax probability that mask embedding z_i matches base phrase p_k, with N denoting the number of class-agnostic binary masks. The similarity between mask and text embeddings is given by $\langle z_i, \mathcal{T}(p_k) \rangle$. We use g_{mm} as a shorthand for $g(x^{(m)}, s^{(m)})$, the similarity score for the matched image-caption pair, while g_{mn} and g_{nm} denote cross-similarity scores within the batch—g_{mn} being the similarity between the m-th image and other captions s^n, and g_{nm} the similarity between the m-th caption and other images x^n. τ is a learnable temperature scaling the similarity scores. The total loss $\mathcal{L}_{\text{total}}$ is computed using weights λ_{BCE}, λ_{DICE}, and λ_{G} for $\mathcal{L}_{\text{BCE}}$, $\mathcal{L}_{\text{DICE}}$, and $\mathcal{L}_{\text{G}}$, respectively.

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\mathbf{p}, \mathbf{y}), \quad \mathcal{L}_{\text{DICE}} = 1 - \frac{2\,\mathbf{y} \cdot \mathbf{p} + 1}{\|\mathbf{y}\| + \|\mathbf{p}\| + 1},$$
$$\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} + \lambda_{\text{DICE}} \mathcal{L}_{\text{DICE}} + \lambda_{\text{G}} \mathcal{L}_{\text{G}}. \quad (5)$$

Here, y_i is the ground-truth mask. In $\mathcal{L}_{\text{DICE}}$, the numerator is the pixel-wise dot product ($\sum_{i=1}^{N} y_i\, \mathbf{p}_{ik}$), and the denominator sums the pixel values of $\mathbf{y}$ and $\mathbf{p}_{ik}^{b}$.

Implementation Details

I2Tx model. We use CLIP-ViT-H/14 (trained on the 224×224 LAION-2B English subset of LAION-5B) as the base CLIP-Vision model. TexDreamer trains an IM2CONTEXT model (linear projection + 3-layer Transformer decoder; visual/text dims: 1280/1024) and fine-tunes CLIP-Vision (scaling factor $\alpha = 16$, rank $r = 16$) and SD ($\alpha = 128$, $r = 128$) with LoRA, all on the ATLAS dataset (Liu et al. 2024) (50K 1024×1024 3D human textures in SMPL UV space). We merge the LoRA adapters into CLIP-Vision and the texture-finetuned SD UNet via Diffusers, using a diffusion time step of 0, a 256-dimensional latent, and seed 777. Captions are processed by extracting $K_{\text{phrase}} = 9$ nouns/adjectives with NLTK, and text embeddings come from ViT-H/14 OpenCLIP.

Semantic parsing. I2Tx features are upsampled via MSDeformAttn (Zhu et al. 2021). The transformer decoder (8 heads, 9 layers) predicts $N = 100$ class-agnostic binary masks and dynamically assigns unique colours to each mask for body parts and clothing.

Datasets. We train on 100K images randomly drawn from 27 parquet files (514K total) in LAION2B-en and LAION1B-nolang (CosmicMan-HQ subset of LAION-5B), yielding 6M images overall, obtained via img2dataset tool. Seventeen base categories are listed in table 4, and GT masks are resized to the original image size during training.

Training details. Images are augmented with horizontal/vertical flips and resized to 1024; test images keep aspect ratio with a 1024 short side. Loss weights are λ_{BCE}=2.0, λ_{DICE}=5.0, λ_{G}=1.0. Following (Cheng et al. 2022), mask loss is computed on 112×112 (12,544) random points from matched masks. We use AdamW with a 1e-4 learning rate and 0.05 weight decay. The model has 2.005B parameters (29.55M trainable, 0.86%) and trains for 12.2 days (370K iters, batch size 8) on eight A100 GPUs.

Experiments

Evaluation settings. We assess Spectrum under four tasks (see fig. 4): FPP—Full-Person Parsing (one mask per person), BHP—Bare-Human Parsing (body part masks),

Figure 4: FPP, BHP, CCP, and COP masks.

Method	CosmicManHQ						Cross-dataset: GranD-f					
	COP			BHP			COP			BHP		
	mIoU	mAcc	mAP^{SS}	mIoU	mAcc	mAP^{SS}	mIoU	mAcc	mAP^{SS}	mIoU	mAcc	mAP^{SS}
PPP-SCHP (Mottaghi et al. 2014b)	—	—	—	45.2	49.7	34.0	—	—	—	20.4	22.4	15.3
ATR-SCHP (Liang et al. 2015)	44.3	48.7	34.0	50.2	55.4	37.6	19.5	21.4	14.7	22.8	25.1	17.2
LIP-SCHP (Gong et al. 2017)	42.0	45.4	31.1	46.6	50.8	34.7	18.5	20.3	13.9	21.6	23.7	16.2
CIHP-PGN (Gong et al. 2018)	43.1	47.1	30.6	48.4	53.1	37.1	19.0	20.8	13.3	22.2	23.4	16.0
Sapiens-1B (Khirodkar et al. 2024)	52.4	57.1	40.3	58.5	62.2	42.1	25.1	27.3	18.4	27.5	30.5	20.5
M2F-C_{close} (Cheng et al. 2022)	49.7	54.4	33.6	54.1	57.2	34.1	24.8	26.9	17.6	25.0	27.0	18.0
ODISE-C_{open} (Xu et al. 2023)	65.7	72.2	49.3	73.4	77.0	53.7	21.5	27.2	12.2	23.5	29.2	14.2
Ours	**77.5**	**79.4**	**58.2**	**86.3**	**88.2**	**65.1**	**27.1**	**29.8**	**21.0**	**30.2**	**33.1**	**22.9**

Table 1: Human parsing on CosmicManHQ (test) and GranD-f (cross) under COP and BHP. M2F-C_{close} and ODISE-C_{open} are retrained on our CosmicManHQ split. Best scores are highlighted in **teal**, and second-best scores are shown in gray.

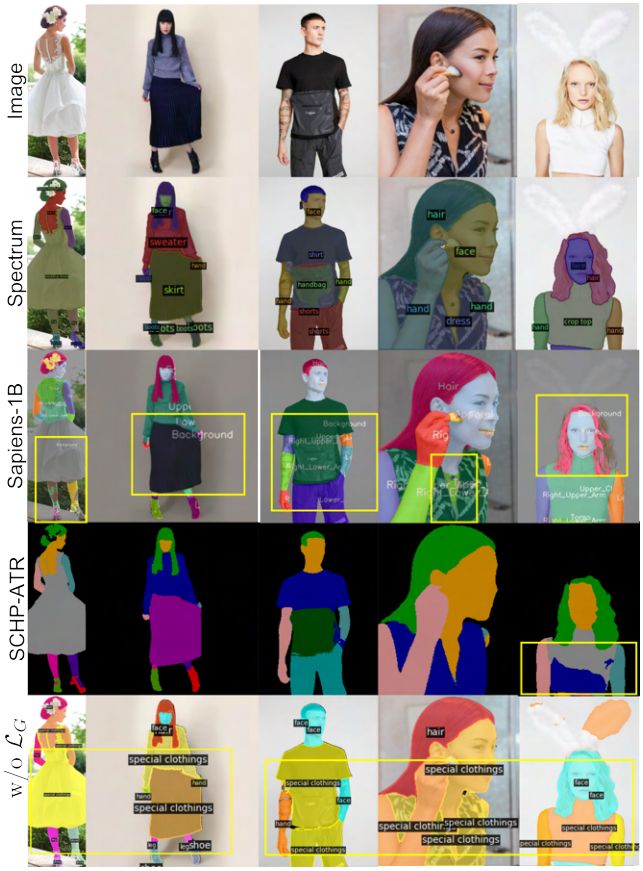

Figure 5: Standard parsers (R2–4); w/o $\mathcal{L}_G$ ablation (R5).

CCP—COCO Category Parsing (accessories: *backpack*, *umbrella*, *shoe*, *eye glasses*, *handbag*, *tie*, *suitcase*), and COP—Clothing-Only Parsing (one mask per clothing). To ensure fair comparison, we unify baseline labels by re-labelling equivalent categories (e.g., *pants*→*bottoms*, *upper/lower arms*→*hand*) and ignore classes a baseline was never trained to predict. Spectrum outputs BHP, CCP, and COP masks, which we merge to form an FPP mask for comparison. All scores use pixel-wise agreement with the GT.

Cross-dataset evaluation. The model is trained on single-human CosmicManHQ (Li et al. 2024) and tested on its hold-out split, on GranD-f (Rasheed et al. 2024) (multi-human, all four settings), and on standard human parsing datasets ATR (Liang et al. 2015), LIP (Gong et al. 2017), PPP (Mottaghi et al. 2014b), and CIHP (Gong et al. 2018). GranD-f poses challenging crowded scenes, heavy occlusion, and incomplete GT masks—resulting in lower absolute scores for all methods. The prompts follow the original annotations (CosmicManHQ: BLIP (Li et al. 2022b); GranD-f: Vicuna (Chiang et al. 2023)); the lengthy captions are compressed while retaining every clothing term. Qualitative figures show the prompt (teal) and the extended body-parts list (EBP); the full EBP is *face*, *faces*, *hand*, *hands*, *hair*, *hairs*, *wavy*, *ponytail*, *bob*, *bald*, *curly*, *afro-hair*, *leg*, *legs*, *back*, *chest*, *belly*, *stomach*, and *feet*.

Method	CosmicManHQ – COP			Cross data: GranD-f – COP		
	mIoU	mAcc	mAP^{SS}	mIoU	mAcc	mAP^{SS}
GD† (Li et al. 2023)	14.6	18.0	11.5	7.6	9.8	6.5
IFSeg (Yun et al. 2023)	32.0	35.2	24.1	13.7	15.0	10.4
MCLIP* (Zheng Ding 2023)	29.1	32.0	21.9	13.2	14.5	10.0
ODISE† (Xu et al. 2023)	33.3	36.6	25.1	14.0	15.4	10.6
OVSeg* (Liang et al. 2023)	25.6	27.5	19.3	12.8	14.1	9.7
OVAM† (Manchón et al. 2024)	25.6	29.5	19.3	8.5	11.8	7.9
SED (Xie et al. 2024)	44.0	48.4	33.1	14.6	16.8	11.0
SEEM (Zou et al. 2023)	29.8	36.0	21.4	11.4	12.8	9.9
LISA (Lai et al. 2024)	40.5	44.0	30.4	15.2	17.7	10.9
Ours‡	**77.5**	**79.4**	**58.2**	**27.1**	**29.8**	**21.0**

Table 2: Open-vocabulary semantic parsing. Legend: † diffusion-based; ∗ CLIP-based; ‡ uses both methods.

Method	CosmicManHQ				Cross-dataset: GranD-f			
	COP		FPP		COP		FPP	
	mAP^{IS}	AR@100	mAP^{IS}	AR@100	mAP^{IS}	AR@100	mAP^{IS}	AR@100
GGN	—	—	20.4	33.0	—	—	11.7	26.3
MCLIP*	3.9	6.0	27.8	39.4	2.2	5.4	19.5	29.2
OLN	—	—	23.6	30.7	—	—	11.3	24.7
ODISE†	4.2	7.7	36.5	59.5	3.6	6.9	22.9	30.1
Ours†	**24.7**	**30.3**	**42.4**	**66.1**	**17.1**	**26.2**	**31.2**	**39.8**

Table 3: Open-vocabulary instance parsing.

Closed-Set Segmentors

We benchmark Spectrum against state of the art human-parsing models that assume a fixed label set (BHP and COP only). Specifically, we compare against (i) the original Sapiens-1B (Khirodkar et al. 2024) and CIHP-PGN (Gong et al. 2018); (ii) three SCHP (Li et al. 2022c) variants retrained on the standard ATR, LIP, and PPP human parsing datasets; and (iii) two models trained on

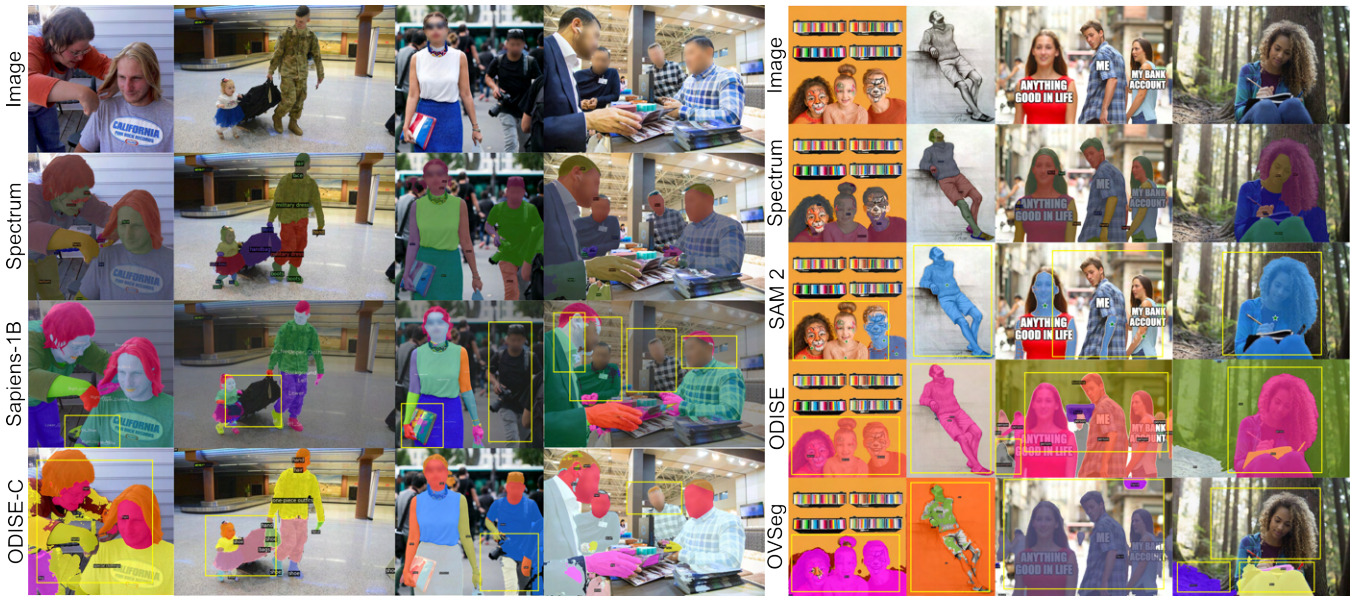

Figure 6: (Left) Multihuman GranD-f: ours surpasses Sapiens-1B and retrained ODISE-C. (Right) Open-vocabulary parsing in the wild: baselines merge all people into one mask, whereas ours preserves distinct clothing and body-part masks.

Model	Face	$_L$Hand$_R$		Hair	Bags	Special Clothings	Tops	Eyewear	$_L$Leg$_R$		Hats	Belts	$_L$Shoe$_R$		One-piece Outfits	Scarf	Bottoms	Avg
w/o **C**	16.71	19.63	16.46	15.11	18.5	11.59	17.48	19.58	15.19	16.12	18.81	13.91	18.11	16.19	17.07	19.43	18.93	16.99
w/o $\mathcal{L}_G$	52.99	62.24	52.2	47.91	58.67	36.76	55.43	62.09	48.18	51.13	59.67	44.1	57.44	51.35	54.14	61.61	60.05	53.88
w/o f	66.25	77.82	65.26	59.9	73.35	45.95	69.3	77.63	60.23	63.93	74.59	55.13	71.82	64.2	67.68	77.02	75.07	67.36
Ours	84.47	91.22	83.21	84.37	93.52	62.59	88.36	94.98	76.8	81.51	95.11	72.3	91.57	81.86	86.3	96.21	95.72	**85.89**

Table 4: Per-class IoU (%) for Spectrum and ablations (w/o **C**, $\mathcal{L}_G$, f). L/R subscripts indicate left/right parts.

Figure 7: Diffusion timesteps ablation. Parsing at $t = 0, 100, 200, 500$; $t = 0$ performs best.

(a) Unseen (U)/ Seen (S) COP

Method	Cosmic		GranD-f	
	U	S	U	S
ODISE	24.6	38.5	7.7	18.1
ODISE-C	51.2	69.0	15.2	23.3
Ours	**60.8**	**80.8**	**18.9**	**29.9**

(b) Cross-dataset eval.

Method	LIP	ATR	PPP	CIHP
SCHP	59.36	82.29	71.46	–
PGN	–	–	–	55.8
Ours	**85.6**	**85.9**	**82.5**	**80.3**

Table 5: Unseen/seen and cross-dataset mIoU evaluation.

our CosmicManHQ split—Mask2Former in a closed-set setting (M2F-C$_{\text{close}}$) (Cheng et al. 2022) and the open-vocabulary segmentor ODISE-C (Xu et al. 2023). (PPP-SCHP outputs BHP masks only due to label mismatches with COP.) Table 1 reports results on the CosmicManHQ test split and on GranD-f. Each COP/BHP entry lists (mIoU, mAcc, mAP$^{\text{SS}}$@0.5–0.95). Spectrum outperforms all baselines in every metric; Sapiens-1B ranks second, while ODISE-C, though open-vocabulary, still falls short.

Method	FPS	GFLOPS	#Params
OVSeg (Liang et al. 2023)	1.1	0.830	531M
ODISE (Xu et al. 2023)	0.6	0.953	1522M
MCLIP (Zheng Ding 2023)	**2.5**	**0.542**	**367M**
Ours	0.5	0.976	1571M

Table 6: Model size comparison.

Cross-dataset results in table 5b show that our method achieves higher mIoU than the three SCHP variants on their respective training datasets (LIP, ATR, and PPP) and also surpasses the PGN baseline on CIHP, highlighting its strong generalization ability. Qualitative comparisons in fig. 5 (rows 1–4) reflect this trend: Spectrum delivers complete, semantically correct masks for all visible body parts and novel garments (*wedding dress*, *skirt*, *shirt*), dynamically assigning new colours (e.g., different colours for multiple hands). In comparison, Sapiens-1B and other baselines show fragmented regions (yellow box highlight). See the Sup. Mat. for dynamic-color mask assignments and captions for all qualitative results.

Open-Vocabulary Segmentors

Most open-vocabulary models output a single *person* mask (FPP) and lack BHP support; we therefore evaluate them only under COP. **Semantic segmentation.** We benchmark IFSeg (Yun et al. 2023), MCLIP (Zheng Ding 2023), OVSeg (Liang et al. 2023), SED (Xie et al. 2024),

SEEM (Zou et al. 2023), LISA (Lai et al. 2024), and the T2I-based GD (Li et al. 2023), OVAM (Manchón et al. 2024), and ODISE (Xu et al. 2023). **Instance segmentation.** We include GGN (Wang et al. 2022), OLN (Kim et al. 2022b), MCLIP, and ODISE. Table 2 lists COP scores (mIoU, mAcc, mAP^{SS}@0.5–0.95) on the CosmicManHQ test split and on GranD-f; Spectrum outperforms all semantic baselines, with SED and LISA closest. Table 3 reports instance metrics (mAP^{IS}@0.5–0.95, AR@100); Spectrum again leads by a wide margin (GGN, OLN cannot output COP masks). Qualitative comparisons in fig. 6 show complex multi-human scenes from GranD-f (left) and in-the-wild images (right). Spectrum yields coherent masks for every garment and body part, whereas parsing baselines (Sapiens-1B, ODISE-C) fragment garments and open-vocabulary segmentors (ODISE, OVSeg) often collapse all people into one mask (highlighted in yellow boxes). We also tested SAM 2 (Ravi et al. 2025) with point prompts on a small subset; results are shown qualitatively only, as its task setting is not directly comparable.

Unseen Clothing

Because closed-set parsers cannot handle novel classes, we evaluate only open-vocabulary models under COP, separating seen vs. unseen clothing. A BERT-NER filter (Devlin et al. 2019) removes test-set labels that overlap the training set, leaving categories such as *maxi*, *floral shirt*, *jumpsuit*, *loafers*, *sneakers*, *polo shirt*, *midi*, *camisole*, *hoodie*, *leather jacket*, *slippers*, *military dress*, *saree*, *baby dress*, *jersey*, *fur coat*, *brief*, *bandana*, etc. Table 5a reports mIoU on CosmicManHQ and GranD-f for unseen/seen COP splits, comparing Spectrum with two top-performing baselines, ODISE and retrained ODISE-C. Spectrum achieves the highest scores, with ODISE-C second. Qualitative examples in figs. 1, 5 and 6 further show accurate masks on novel clothing, thanks to the repurposed I2Tx features and prompt grounding.

Model Size Comparison

Table 6 reports FPS, GFLOPs, and parameters. Although Spectrum is not efficiency-focused, it runs on a single 20GB RTX4000 with compute close to ODISE, while MCLIP has the smallest footprint.

Ablation Studies

We ablate Spectrum on the CosmicManHQ test split and report per-class IoU is in table 4. Additional FPP/CCP metrics, qualitative examples (single/multi-human), prompt variants, limitations, and ethical notes are in the Sup. Mat.

Standard SD features. Substituting our I2Tx texture features with vanilla SD latent features reduces mIoU to 67.36, showing that texture-aligned latents are substantially more informative for diverse clothing and body-part parsing.

Without context embedding (C and CLS). Removing the context embedding drops mIoU from 85.89 to 16.99, as it provides essential conditioning for the I2Tx model.

Without grounding loss ($\mathcal{L}_{\text{G}}$). Omitting the contrastive prompt-grounding loss drops mIoU to 53.88 and causes clothing to collapse to generic *Special Clothing* label (fig. 5 R5), highlighting the importance of prompt alignment.

Diffusion timesteps. We extract I2Tx features in a single forward pass at t=0, skipping denoising. Sampling $t \in \{100, 200, 500\}$ from the 1000-step schedule yields fragmented masks and poorer segmentation (fig. 7).

Low exposure ablation. Our training data (CosmicManHQ split) and ATLAS 3D textures are well lit, so darker scenes create a domain gap. Applying gamma correction ($0 < \gamma < 1$) to simulate under-exposure drops mIoU to 79.11 at $\gamma = 0.75$, 54.20 at $\gamma = 0.50$, and 30.90 at $\gamma = 0.25$, showing that low light markedly harms performance.

Prompt dependency. Spectrum is insensitive to prompt count, mIoU is stable (77.5@K_{phrase}=9 vs. 76.8@K_{phrase}=5), and dropping 30% of phrases lowers mIoU by only 0.8.

Conclusion

We present Spectrum, a unified network for part-level pixel parsing of diverse clothing and body parts, as well as instance-level part grouping. It leverages the internal representation of an I2Tx diffusion model—trained to generate 3D textures from input images—to generate accurate segmentation maps for visible body parts and clothing categories, while ignoring standalone garments and irrelevant objects. Extensive cross-dataset experiments, including single and multi-human scenarios, evaluate body parts, clothing parts, unseen categories, and full-body masks. Spectrum consistently outperforms state-of-the-art human parsers, open-vocabulary segmentors, and instance segmentation baselines in prompt-based mask prediction. In particular, it achieves superior generalization to unseen clothing categories in cross-dataset settings. These findings suggest that repurposing 3D texture features from fine-tuned diffusion models can support more detailed and generalizable human parsing.

Acknowledgments

We thank Michael J. Black for valuable feedback. We also thank Peter Kulits, Zitian Zhang, Zhening Huang, and Samuel Sartor for proofreading. This work was conducted during an internship at Adobe Research.

References

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4): 834–848.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation.

Cheng, B.; Schwing, A. G.; and Kirillov, A. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Adv. Neural Inform. Process. Syst.*, 17864–17875.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Dai, J.; He, K.; and Sun, J. 2015. BoxSup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation. In *Int. Conf. Comput. Vis.*, 1635–1643. IEEE Computer Society.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. Association for Computational Linguistics.

Dickens, J.; and Hamad, K. 2025. Part Segmentation of Human Meshes via Multi-View Human Parsing. arXiv:2507.18655.

Fang, H.; Lu, G.; Fang, X.; Xie, J.; Tai, Y.; and Lu, C. 2018. Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 70–78.

Geng, C.; Huang, S.-J.; and Chen, S. 2021. Recent Advances in Open Set Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10): 3614–3631.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. In *Eur. Conf. Comput. Vis.*, 540–557. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-20058-8.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-Level Human Parsing via Part Grouping Network. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Eur. Conf. Comput. Vis.*, volume 11208 of *Lecture Notes in Computer Science*, 805–822. Springer.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6757–6765. IEEE Computer Society.

Gupta, A.; Wu, J.; Deng, J.; and Fei-Fei, L. 2023. Siamese Masked Autoencoders. In *Adv. Neural Inform. Process. Syst.*

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. DDP: Diffusion Model for Dense Visual Prediction. In *Int. Conf. Comput. Vis.*, 21684–21695. IEEE.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In Meila, M.; and Zhang, T., eds., *Int. Conf. on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 4904–4916. PMLR.

Khirodkar, R.; Bagautdinov, T. M.; Martinez, J.; Zhaoen, S.; James, A.; Selednik, P.; Anderson, S.; and Saito, S. 2024. Sapiens: Foundation for Human Vision Models. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Eur. Conf. Comput. Vis.*, volume 15062 of *Lecture Notes in Computer Science*, 206–228. Springer.

Kim, D.; Lin, T.; Angelova, A.; Kweon, I. S.; and Kuo, W. 2022a. Learning Open-World Object Proposals Without Learning to Classify. *IEEE Robotics Autom. Lett.*, 7(2): 5453–5460.

Kim, D.; Lin, T.-Y.; Angelova, A.; Kweon, I. S.; and Kuo, W. 2022b. Learning Open-World Object Proposals without Learning to Classify. *IEEE Robotics and Autom. Lett.*

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9579–9589. IEEE.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022a. BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 21298–21308. IEEE.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022b. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *Int. Conf. on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2022c. Self-Correction for Human Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6): 3260–3271.

Li, S.; Fu, J.; Liu, K.; Wang, W.; Lin, K.; and Wu, W. 2024. CosmicMan: A Text-to-Image Foundation Model for Humans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 6955–6965. IEEE.

Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Open-vocabulary Object Segmentation with Diffusion Models.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7061–7070. IEEE.

Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; and Yan, S. 2015. Deep Human Parsing with Active Template Regression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(12): 2402–2414.

Lin, L.; Zhang, D.; Luo, P.; and Zuo, W. 2019. *Human Centric Visual Analysis with Deep Learning*. Springer Publishing Company, Incorporated, 1st edition. ISBN 9789811323867.

Liu, K.; Choi, O.; Wang, J.; and Hwang, W. 2022. CDGNet: Class Distribution Guided Network for Human Parsing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 4473–4482.

Liu, K.; Wang, J.; Jin, R.; Hwang, W.; and Chung, T.-S. 2025. SCHNet: SAM Marries CLIP for Human Parsing. arXiv:2503.22237.

Liu, Y.; Zhu, J.; Tang, J.; Zhang, S.; Zhang, J.; Cao, W.; Wang, C.; Wu, Y.; and Huang, D. 2024. TexDreamer: Towards Zero-Shot High-Fidelity 3D Human Texture Generation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Eur. Conf. Comput. Vis.*, volume 15105 of *Lecture Notes in Computer Science*, 184–202. Springer.

Manchón, P. M.; Alcover-Couso, R.; SanMiguel, J. C.; and Martínez, J. M. 2024. Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9242–9252.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014a. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 891–898. IEEE Computer Society.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014b. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Peng, X.; Xie, Y.; Wu, Z.; Jampani, V.; Sun, D.; and Jiang, H. 2025. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2878–2888.

Rasheed, H. A.; Maaz, M.; Mullappilly, S. S.; Shaker, A. M.; Khan, S. H.; Cholakkal, H.; Anwer, R. M.; Xing, E. P.; Yang, M.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 13009–13018. IEEE.

Ravi, N.; Gabeur, V.; Hu, Y.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.; Girshick, R. B.; Dollár, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In *Int. Conf. Learn. Represent.* OpenReview.net.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10684–10695.

Tu, M.; Ye, S.; Jung, H.; and Kim, J. 2025. SMPLGPTexture: Dual-View 3D Human Texture Estimation using Text-to-Image Generation Models. arXiv:2504.13378.

Wang, W.; Feiszli, M.; Wang, H.; Malik, J.; and Tran, D. 2022. Open-World Instance Segmentation: Exploiting Pseudo Ground Truth From Learned Pairwise Affinity. *IEEE Conf. Comput. Vis. Pattern Recog.*

Wang, Z.; Miao, Q.; Xi, Y.; and Zhao, P. 2024. Eformer: Enhanced transformer towards semantic-contour features of foreground for portraits matting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3880–3889.

Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; Ghanem, B.; and Tao, D. 2024. Towards Open Vocabulary Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(7): 5092–5113.

Xia, F.; Wang, P.; Chen, L.; and Yuille, A. L. 2016. Zoom Better to See Clearer: Human and Object Parsing with Hierarchical Auto-Zoom Net. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Eur. Conf. Comput. Vis.*, volume 9909 of *Lecture Notes in Computer Science*, 648–663. Springer.

Xie, B.; Cao, J.; Xie, J.; Khan, F. S.; and Pang, Y. 2024. SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 3426–3436. IEEE.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and Mello, S. D. 2023. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2955–2966. IEEE.

Yang, L.; Song, Q.; Wang, Z.; Hu, M.; Liu, C.; Xin, X.; Jia, W.; and Xu, S. 2020. Renovating Parsing R-CNN for Accurate Multiple Human Parsing. In *Eur. Conf. Comput. Vis.*

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6): 2872–2893.

Yuan, H.; Li, X.; Zhou, C.; Li, Y.; Chen, K.; and Loy, C. C. 2024. Open-Vocabulary SAM: Segment and Recognize Twenty-Thousand Classes Interactively. In *Eur. Conf. Comput. Vis.*, 419–437. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-72774-0.

Yun, S.; Park, S. H.; Seo, P. H.; and Shin, J. 2023. IFSeg: Image-free Semantic Segmentation via Vision-Language Model. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Zhang, S.; Cao, X.; Qi, G.-J.; Song, Z.; and Zhou, J. 2022. AIParsing: Anchor-free instance-level human parsing. *IEEE Trans. Image Process.*, 31: 5599–5612.

Zheng Ding, Z. T., Jieke Wang. 2023. Open-Vocabulary Universal Image Segmentation with MaskCLIP. In *International Conference on Machine Learning*.

Zhu, C.; and Chen, L. 2024. A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12): 8954–8975.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Int. Conf. Learn. Represent.* OpenReview.net.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment Everything Everywhere All at Once. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Adv. Neural Inform. Process. Syst.*