SAPIENS2

Anonymous authors
Paper under double-blind review

ABSTRACT

We present Sapiens2, a model family of high-resolution transformers for human-centric vision focused on generalization, versatility, and high-fidelity outputs. Our model sizes range from 0.4 to 5 billion parameters, with native 1K resolution and hierarchical variants that support 4K. Sapiens2 substantially improves over its predecessor in both pretraining and post-training. First, to learn features that capture low-level details (for dense prediction) and high-level semantics (for zero-shot or few-label settings), we combine masked image reconstruction with self-distilled contrastive objectives. Our evaluations show that this unified pretraining objective is better suited for a wider range of downstream tasks. Second, along the data axis, we pretrain on a curated dataset of 750 million high-quality human images and improve the quality and quantity of task annotations. Third, architecturally, we incorporate advances from frontier models that enable longer training schedules with improved stability. Our 4K models adopt windowed attention to reason over longer spatial context and are pretrained with 2K output resolution. Sapiens2 sets a new state-of-the-art and improves over the first generation on pose ($+4$ mAP), body-part segmentation ($+22.3$ mIoU), normal estimation ($+29.2$ relangular error) and extends to new tasks such as pointmap and albedo estimation.

1 INTRODUCTION

Sapiens introduced a foundation model for human-centric vision (Khirodkar et al., 2024). The overarching goal is to build models that operate across *any* human task and *any* human imagery while maintaining *highest* output fidelity. In this work, we present SAPIENS2, which advances this objective along all three axes—task, image, and fidelity.

Any human task. Sapiens primarily relied on MAE (He et al., 2022) pretraining, a form of masked image modeling (MIM) (Hondru et al., 2025). MIM preserves signal and spatial details by optimizing reconstruction and thus primarily learns by compression (Zhang et al., 2022). Unlike language—where tokens are discrete and largely self-semantic and masked modeling has become a default—visual semantics are denser, context-dependent and under-constrained by pixel prediction alone; consequently, MIM features often require moderate-to-high supervision to express semantics reliably. In contrast, contrastive learning (CL) (Chen et al., 2020a) injects semantics by enforcing instance-level invariances using positives and negatives(Chen et al. (2020b), Chen et al. (2021)), yet its global invariance objectives tend to underperform on dense prediction, where fine spatial detail and photometric fidelity matter. This gap has motivated hybrids that combine global CL and MIM - such as iBOT's masked student–teacher matching (Zhou et al., 2021) and successors such as DINOv2 (Oquab et al., 2023) and v-JEPA (Bardes et al., 2024). While these approaches narrow the gap, performance at high resolution remains mixed and can exhibit *representation drift*: aggressive invariances (notably appearance augs.) decouple teacher and student from the true observations, eroding cues -such as color- that are critical for human-centric dense tasks (*e.g.* photorealistic avatar creation). SAPIENS2 addresses these limitations by coupling a reconstruction objective with contrastive objectives, anchoring features in pixel space (Huang et al., 2023) while organizing them semantically. The result is a general-purpose representation that transfers across zero-shot, few-shot (Song et al., 2023), and fully supervised regimes and a broad spectrum of human-centric tasks.

Any human image. Generalization scales with data and model capacity. During pretraining, we curate *750M* high-quality human images from a web-scale corpus via multi-stage filtering. The collection spans diverse ages, ethnicities, backgrounds, and real-world conditions, subject to a single constraint: each image contains at least one prominent person. Beyond this human-centric requirement, we use no task labels and inject no human-specific priors during pretraining. For post-training, we target fundamental human tasks—pose estimation (Zheng et al., 2023), body-part

Figure 1: **SAPIENS2 for dense-prediction tasks.** We compare 1B models from both generations on segmentation, depth, and normals. Sapiens2 improves over Sapiens with stronger generalization and sharper segmentation of rare classes (lips, tongue, earrings), achieving pixel-accurate hair segmentation. On geometric tasks (depth, normals), it captures subtler facial, clothing, and hair details—all without task-specific architectures.

segmentation Thisanke et al. (2023), surface-normal (Bae & Davison, 2024), pointmap (per-pixel XYZ) (Wang et al., 2024) and albedo estimation (Ran et al., 2024). Relative to Khirodkar et al. (2024), we scale task-specific supervision by $10\times$, typically on the order of 1M labels per task, and improve synthetic assets with more detailed geometry and photorealism. On the model axis, our largest variant has 5B parameters, accompanied by 0.4B, 0.8B, and 1B models for different compute settings and broader use. At a native resolution of 1K, our largest model achieves among the highest FLOPs reported for vision transformers. Fig. 1 showcases improvements over Sapiens for segmentation, depth and normals. Our models segment tiny accessories such as chains and earrings, and separate teeth and gums with pixel accuracy. Additionally, the predicted normals better capture facial wrinkles and hair details. Our evaluations show that, learning at scale yields strong generalization across unconstrained human images and challenging in-the-wild conditions.

Highest fidelity. Prediction fidelity scales with the number of visual tokens a model processes, which in turn grows with input resolution (Zhao et al., 2018). Beyond standard 1K backbones (Khirodkar et al., 2024), we introduce a 4K backbone pretrained and post-trained for dense prediction, with task heads that decode to 2K resolution across tasks. To make 4K tractable, we adopt a hierarchical design (Li et al., 2022): an initial stack of windowed self-attention layers operates locally to capture texture and fine boundaries, from each window we pool a summary token and then apply global self-attention—mirroring our 1K models—to fuse long-range context. This layout is naturally compatible with MAE-style pretraining: after the local stage, masked tokens can be dropped so that information does not flow across masked regions, avoiding the leakage that convolutional backbones typically require masked convolutions to prevent (Gao et al., 2022). We additionally incorporate targeted efficiency and stability upgrades—RMSNorm in place of LayerNorm (Meta, 2025), grouped-query attention for higher throughput (Ainslie et al., 2023), QK-Norm for robust high-resolution training (Henry et al., 2020)—and employ a pixel-shuffle (Shi et al., 2016) decoder for sub-pixel reasoning. Together, these choices fully exploit our high-resolution setting while keeping memory in check.

Figure 2: **k-NN comparison using [CLS] token**. SAPIENS2 learns a more discriminative, human-semantic feature space—grouping visually similar concepts and improving retrieval performance at high resolution

We extensively evaluate SAPIENS2 across various tasks and benchmarks. Figure 2 qualitatively visualizes nearest neighbors retrieved using [CLS] tokens from 1K-resolution Sapiens and SAPIENS2. Our contrastive pretraining yields a feature space that captures human semantics and returns plausible neighbors. Figure 3 further shows that, without any supervision, our model produces human-centric attention maps. Overall, our contributions are summarized as follows.

- SAPIENS2 is a family of transformers (0.4B–5B parameters) pretrained on 750 million high-quality human images. Our models support 1K native resolution and 4K hierarchical resolution and are designed for high-resolution dense predictions.
- We use masked reconstruction with contrastive objectives to learn features that generalize in zero-shot settings on human tasks while preserving fine details in dense predictions.
- We fine-tune with high-quality annotations for pose, part segmentation, pointmaps, normals, and albedo, achieving state-of-the-art performance across benchmarks.

2 RELATED WORK

Self-Supervised Learning. Recent breakthroughs in self-supervised learning at scale fall into two families: (1) Masked Image Modeling (MIM) and (2) Contrastive Learning (CL). MIM follows masked language modeling in NLP, but unlike language—where tokens are self-semantic—image patches are context-dependent. Visual representations are thus denser and more ambiguous. MIM objectives are commonly viewed as a form of compression (Zhang et al., 2022) of the input tokens. Among popular approaches, BEiT (Bao et al., 2021) uses a dVAE tokenizer to discretize image patches and trains the model to predict the codebook indices of masked patches, while MAE (He et al., 2022) masks a large fraction of patches (75%) and reconstructs the missing pixels directly. Numerous studies adopt this paradigm for pretraining—e.g., U-MAE, CAE, SiamMAE, MR-MAE, and Sapiens (Khirodkar et al., 2024). Representative methods in CL include BYOL (Grill et al., 2020), SimCLRv2 (Chen et al., 2020b), MoCov3 (Chen et al., 2021), and DINO (Caron et al., 2021). Given their complementarity, combining the objectives is natural; for instance, iBOT (Zhou et al., 2021) combines MIM with CL-style self-distillation, aligning student and teacher features via the masked objective rather than reconstructing pixels or codewords, consistent with JEPA (Assran et al., 2023) and v-JEPA2 (Assran et al., 2025). DINOv2 (Oquab et al., 2023) adopt iBOT objective as their primary pretraining strategy. However, latent-space objectives risk abstract drift: the representations are not anchored to observations (images or sentences), inducing lossy compression and discarding cues (*e.g.* color) critical for dense prediction. In Sapiens2, we combine the image-anchored MAE objective with the semantic CL objective. Prior work such as CMAE (Huang et al., 2023) explores this combination but evaluates primarily on classification. In contrast, we study a unified objective at billion-scale across multiple human-centric tasks.

Human-Centric Vision Models. Many recent works focus on building models for human-centric vision. These models often outperform general models of similar scale on human-related tasks. For instance, HAP (Yuan et al., 2023) uses 2D keypoints to guide the mask sampling process during masked image modeling, encouraging the model to focus on body structure information. Geoman (Kim et al., 2025a) uses a image-to-video diffusion model for geometry estimation. HCMoCo (Hong et al., 2022) and PBoP (Meng et al., 2024) employ multiple encoders to exploit multimodal human body consistency through a hierarchical contrastive learning framework. SOLIDER (Chen et al., 2023) introduces a human semantic classification loss to inject semantic information into the learned features. LiftedCL (Chen et al., 2022) incorporates an adversarial

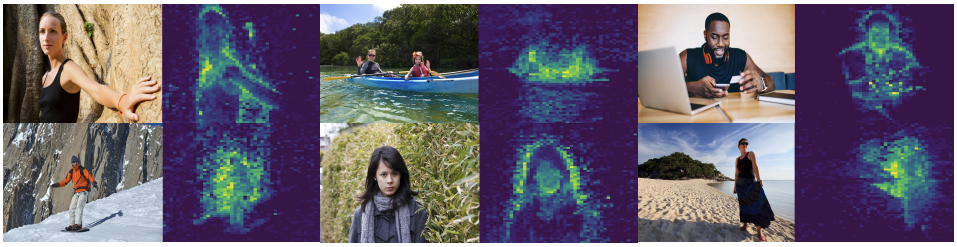


Figure 3: **Human-centric attention.** Visualization of [CLS]-token self-attention across heads in the final layer.

loss to supervise the lifted 3D skeletons, explicitly embedding 3D human structure information for human-centric pretraining. SapiensID (Kim et al., 2025b) trains a model specifically for person re-identification. In contrast to these approaches, Sapiens2 does not inject any explicit human priors beyond the data itself during pretraining. This truly inductive prior-free approach enables scaling to millions of images and model sizes without introducing handcrafted human-centric biases.

Vision Transformers at Scale. Although the largest vision backbones remain an order of magnitude smaller than language models (Lu et al., 2024), the field is scaling rapidly as both data and model sizes grow. To clarify the landscape, we position prior works along three axes: parameters, resolution, and data. Amongst notable recent works, the largest vision backbone in the Perception Encoder family (Bolya et al., 2025) has 2B parameters, is trained at 448 px resolution, and uses 5.4B samples. DINOv2 (Oquab et al., 2023) scales to 1B parameters at 512 px and is pretrained on 152M images. ViT-22B (Dehghani et al., 2023) remains the largest model by parameter count; it is trained at 224 px and is pretrained on 1M images from ImageNet (Russakovsky et al., 2015). Sapiens-2B (Khirodkar et al., 2024), at 1024 px, was the largest human-centric vision backbone, pretrained on 300M human images. In Sapiens2, we scale to 5B parameters and extend the input resolution to 4K, yielding a vision backbone with the largest FLOPs, trained on 750M human images.

3 Pretraining

This section details our pretraining data and methodology, with emphasis on human-centric curation and design choices that preserve output fidelity and strengthen semantic understanding.

3.1 Humans-750M Dataset

Scale helps only when the data distribution is diverse, balanced, and high quality (Touvron et al., 2023; Radford et al., 2021; Chuang et al., 2025). From a web-scale pool of ∼4B images, we isolate human-centric content via a multi-stage filter: bounding box detection, head-pose estimation, aesthetic and realism scoring, CLIP (Radford et al., 2021) features and text-overlay detection. We remove images that fail realism, quality or other checks. From the remainder, we retain instances where at least one person is ≥ 384 pixels on the short side; images may contain multiple people. We deduplicate via perceptual hashing and deep-feature nearest-neighbor pruning, and we cluster visual embeddings followed by selective sampling (Oquab et al., 2023) to balance content across poses, viewpoints, occlusion, clothing, scene types, and illumination. Thresholds and balance caps are calibrated with small human audits. The result is a curated, balanced corpus of ∼750M high-quality human images for pretraining.

3.2 Self-Supervised Learning

Let $\mathcal{I}$ denote the training set. We sample an image $\mathbf{x} \sim \mathcal{I}$ and draw V random augmentations to obtain views $\{\mathbf{x}_i\}_{i=1}^{V}$. Each view is patchified into N tokens indexed by $\mathcal{P} = \{1, \ldots, N\}$, i.e., $\mathbf{x}_i = \{\mathbf{x}_i^p\}_{p\in\mathcal{P}}$. Let $\{\mathbf{e}_{\text{pos}}^p\}_{p\in\mathcal{P}}$ be positional embeddings (Dosovitskiy et al., 2020) and Φ_{enc}, Φ_{dec}, Φ_{cls} be our transformer encoder, patch decoder and contrastive decoder respectively. Specifically, Φ_{cls} maps the encoder [CLS] token to K logits.

Masked Image Modeling. For each view $i \in \{1, \ldots, V\}$, we sample a binary mask $\mathbf{m}_i \in \{0, 1\}^N$ with masking ratio r. The masked and visible index sets are defined as $\mathcal{M}_i = \{p \in \mathcal{P} : m_i^p = 1\}$ and $\mathcal{V}_i = \mathcal{P} \setminus \mathcal{M}_i$. The encoder Φ_{enc} processes only visible tokens: $\mathbf{z}_i^{\text{vis}} = \Phi_{\text{enc}}(\{\mathbf{x}_i^p + \mathbf{e}_{\text{pos}}^p\}_{p\in\mathcal{V}_i})$. We then form a full sequence by scattering $\mathbf{z}_i^{\text{vis}}$ back to $\mathcal{V}_i$ and inserting a learned mask token at $\mathcal{M}_i$: $\mathbf{z}_i = \text{scatter}\big(\mathbf{z}_i^{\text{vis}}; \mathcal{V}_i\big) \cup \big\{\mathbf{e}_{[\text{MASK}]} + \mathbf{e}_{\text{pos}}^p\big\}_{p\in\mathcal{M}_i}$. The decoder Φ_{dec} reconstructs all patches, $\hat{\mathbf{x}}_i = \Phi_{\text{dec}}(\mathbf{z}_i)$ with outputs $\{\hat{\mathbf{x}}_i^p\}_{p\in\mathcal{P}}$. Following He et al. (2022), targets are normalized $\tilde{\mathbf{x}}_i^p$, and the loss averages MSE over *masked* tokens and views:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{|\mathcal{M}_i|}\sum_{p\in\mathcal{M}_i}\big\|\tilde{\mathbf{x}}_i^p - \hat{\mathbf{x}}_i^p\big\|_2.$$

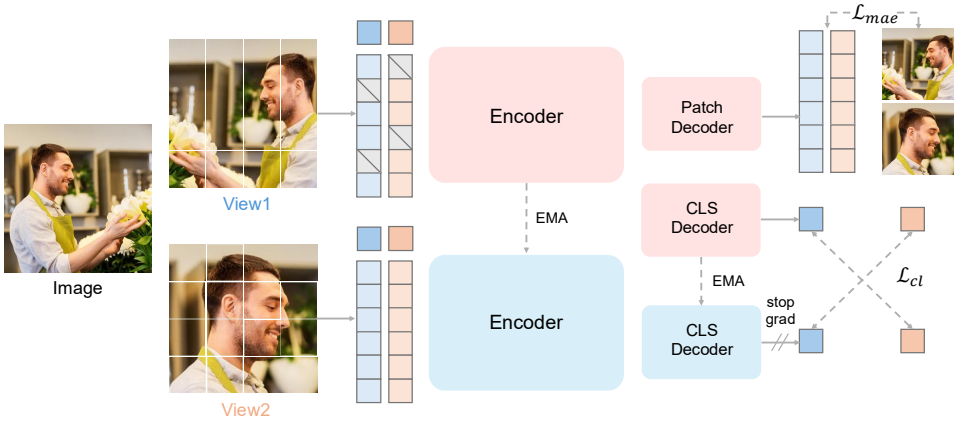


Figure 4: **SAPIENS2 Pretraining**. We combine the masked reconstruction loss ($\mathcal{L}_{\text{mae}}$) with a global contrastive loss on [CLS] ($\mathcal{L}_{\text{cl}}$). Multiple image views are generated, and a student–teacher framework matches predicted distributions across views. $\mathcal{L}_{\text{mae}}$ helps the model learns low-level details (*e.g.*texture) for high-fidelity dense tasks, while $\mathcal{L}_{\text{cl}}$ improves semantic understanding across human images.

Contrastive Learning. We adopt a student–teacher scheme (Caron et al.), the teacher has the same architecture $(\Phi_{\text{enc}}, \Phi_{\text{cls}})$, is *non-learnable*, and its parameters are an EMA of the student. For each view i, the student and teacher $[\text{CLS}]$ embeddings and logits are

$$\mathbf{c}_i^s = [\text{CLS}](\Phi_{\text{enc}}(\mathbf{x}_i)), \quad \mathbf{c}_i^t = [\text{CLS}](\Phi_{\text{enc}}^{\text{ema}}(\mathbf{x}_i)), \qquad \mathbf{s}_i = \Phi_{\text{cls}}(\mathbf{c}_i^s), \quad \mathbf{t}_i = \Phi_{\text{cls}}^{\text{ema}}(\mathbf{c}_i^t),$$

with $\mathbf{p}_i = \text{softmax}(\mathbf{s}_i)$ and $\mathbf{q}_i = \text{softmax}(\mathbf{t}_i)$. For the V-view (global + local) setting, we form the positive pair set $\mathcal{S}$ consisting of all cross-view global↔global and global↔local pairs (excluding same-view matches for global crops; local↔local pairs are skipped). The contrastive objective averages a teacher-to-student cross-entropy over these pairs:

$$\mathcal{L}_{\text{CL}} = \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} H(\mathbf{q}_j, \mathbf{p}_i), \qquad H(\mathbf{q}, \mathbf{p}) = -\sum_{k=1}^{K} q_k \log p_k.$$

Finally, Fig. 4 shows our pretraining setup for $V = 2$. We use a joint objective $\mathcal{L} = \mathcal{L}_{\text{MAE}} + \lambda \mathcal{L}_{\text{CL}}$, combining human-centric low-level fidelity with view-invariant semantics.

4 MODEL ARCHITECTURE

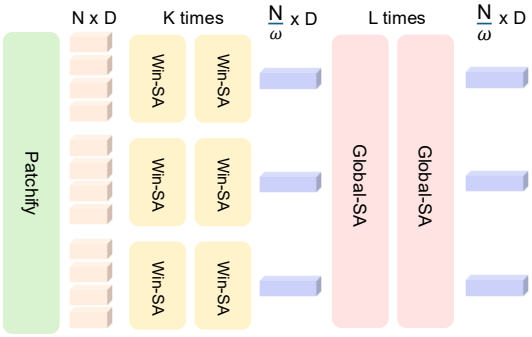
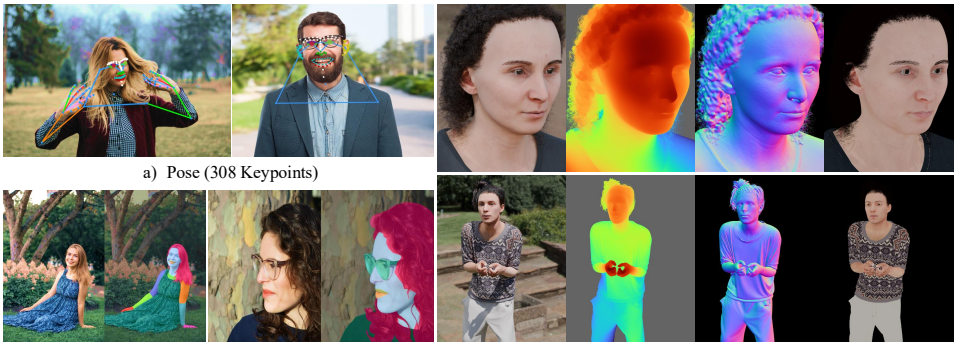


Figure 5: **Windowed self-attention** for 4K resolution.

We revise the backbone to stably scale to 5B parameters, increase the input resolution from 1K to 4K, and maintain compatibility with sparse masked pretraining. The mid-depth blocks use grouped-query attention (GQA) (Ainslie et al., 2023), while the early and late blocks use standard multi-head self-attention. We replace the feed-forward layers with gated SwiGLU-FFN variants (Shazeer, 2020). For long-schedule stability, we apply QK-Norm (Henry et al., 2020)—normalizing queries and keys before attention—and substitute LayerNorm with the parameter-efficient RMSNorm (Zhang & Sennrich, 2019). To scale to 4K inputs, we adopt a hierarchical attention design (Ryali et al., 2023): given an $H \times W$ image with patch size p, yielding $N = (H/p)(W/p)$ tokens, the first K layers apply windowed self-attention to capture local structure. We then downsample the 2D token grid by a spatial stride $\sqrt{\omega}$ via $[\text{CLS}]$-guided pooling to obtain N/ω tokens. Next L layers use global attention over this reduced sequence, refer Fig. 5. During pretraining, we apply token masking after the local stage, and include a brief masked-reconstruction phase at 2K to sharpen sub-pixel fidelity on dense tasks without degrading semantics. Finally, we increase decoder outputs to 1K for base backbones (from 0.5K) and to 2K for 4K backbones.

5 POST-TRAINING

We fine-tune the pretrained backbone on five human-centric tasks—pose estimation, body-part segmentation, depth, surface normals, and albedo—using lightweight task-specific heads while leaving the backbone unchanged. Relative to Khirodkar et al. (2024), we broaden supervision and refine task objectives.



a) Pose (308 Keypoints)

b) Segmentation (29 Classes)

c) Synthetic (Pointmap, Normal, Albedo)

Figure 6: **Post-Training Annotations.** We annotated 100K in-the-wild images with *pose (a)* and *segmentation (b)*, class vocabulary is also extended to include eyeglasses (in cyan). For *pointmap, normal, albedo (c)*, we improve our synthetic assets to capture finer geometric details and color variations.

Pose Estimation. We follow a top-down paradigm to estimate keypoint heatmaps from an input image. Our keypoint topology is a 308-keypoint full-body skeleton with dense coverage of the face (243) and hands (40 total), with the remainder spanning torso and lower-body. Unlike Khirodkar et al. (2024), which relied solely on capture-studio annotations, we add in-the-wild supervision (Fig. 6a) by newly annotating 100K high-resolution images from our pretraining corpus with the same vocabulary. This hybrid supervision improves generalization to unconstrained images. Our objective uses MSE over ground-truth heatmaps with OHEM (Chen et al., 2018) to focus supervision within a large keypoint set as $\mathcal{L}_{\text{pose}} = \sum_{u\in\Omega} \|\hat{\mathbf{H}}(u) - \mathbf{H}(u)\|_2$.

Body-Part Segmentation. Our segmentation vocabulary has 29 classes (previously 28 plus eyeglasses category; see Fig. 6b). The vocabulary targets part-specific supervision and precise localization of semantic human body parts. Similar to pose, we increase segmentation supervision to 20K in-the-wild images with segmentation labels. Our objective uses per-pixel weighted cross-entropy combined with Dice loss (Azad et al., 2023) for sharper boundaries.

Pointmap (Depth) Estimation. Rather than relative depth, we regress a per-pixel 3D pointmap $\hat{\mathbf{P}}(u) \in \mathbb{R}^3$ in the camera frame. Since metric scale is ambiguous with unknown intrinsics (Yin et al., 2023), we predict a focal-normalized pointmap $\tilde{\mathbf{P}}(u)$ and a scalar head s, forming $\hat{\mathbf{P}}(u) = s\,\tilde{\mathbf{P}}(u)$ (Bochkovskii et al., 2024). Supervision is entirely synthetic and use higher-fidelity assets (hair, eyes, fine facial wrinkles, Fig. 6c). The loss is $\mathcal{L}_{\text{pointmap}} = \sum_{u\in\Omega} \|\hat{\mathbf{P}}(u) - \mathbf{P}(u)\|_2 + \|\nabla\hat{\mathbf{P}}(u) - \nabla\mathbf{P}(u)\|_2$ where ∇ is finite differences along XY.

Normal Estimation. We predict per-pixel unit normals $\hat{\mathbf{N}}(u) \in \mathbb{R}^3$ for human pixels using the same high-fidelity synthetic assets; the decoder uses multiple PixelShuffle (Aitken et al., 2017) layers for artifact-free upsampling. The loss is defined as: $\mathcal{L}_{\text{normal}} = \sum_{u\in\Omega} 1 - \hat{\mathbf{N}}(u)\cdot\mathbf{N}(u) + \|\hat{\mathbf{N}}(u) - \mathbf{N}(u)\|_2 + \|\nabla\hat{\mathbf{N}}(u) - \nabla\mathbf{N}(u)\|_2$.

Albedo Estimation. We predict per-pixel diffuse albedo $\hat{\mathbf{A}}(u) \in [0,1]^3$, crucial for relighting (Kim et al., 2024). Training uses high-fidelity synthetic pairs $\mathbf{A}(u)$ (Fig. 6c) and encourages illumination-invariant recovery of skin tone and clothing. The loss is $\mathcal{L}_{\text{albedo}} = \sum_{u\in\Omega} \|\hat{\mathbf{A}}(u) - \mathbf{A}(u)\|_2 + \|\nabla\hat{\mathbf{A}}(u) - \nabla\mathbf{A}(u)\|_2 + \|\mu(\hat{\mathbf{A}}) - \mu(\mathbf{A})\|_2$, where $\mu(\cdot)$ is the spatial RGB mean for alignment.

6 Experiments

In this section, we initially outline implementation details, then evaluate pretrained feature generalization using dense probing and post-train performance across a variety of downstream tasks.

6.1 Implementation Details

Sapiens2 is implemented in PyTorch with HF-Accelerate (Gugger et al., 2022). All our models are trained on A100 GPUs using bfloat16 and FSDP for efficiency. We use fused AdamW (Loshchilov & Hutter, 2017) as the optimizer for all experiments, with a brief learning-rate warmup followed by cosine decay. We pretrain from scratch at 1024×768 (1K) and 4096×3072 (4K) resolutions. Starting from Sapiens–0.3B, 0.6B and 1B, we apply the architectural revisions in Sec. 2 to produce Sapiens2–0.4B, 0.8B and 1B. To push the frontier for human-centric vision models, we also introduce a 5B model that scales both network depth and token embedding dimensions. Sapiens2-5B is the highest-FLOPs vision transformer at 15 TFlops. Table 1 summarizes our model configurations at 1K resolution. Finally, we fine-tune the 1B–4K model for segmentation and normal estimation.

Model	Parent-Model	#Params	FLOPs	Hidden size	Layers	Heads
Sapiens2-0.4B	Sapiens-0.3B	0.398 B	1.260 T	1024	24	16
Sapiens2-0.8B	Sapiens-0.6B	0.818 B	2.592 T	1280	32	16
Sapiens2-1B	Sapiens-1B	1.462 B	4.715 T	1536	40	24
Sapiens2-5B	-	5.071 B	15.722 T	2432	56	32

Table 1: **SAPIENS2 architectural details**. Broadly, we base the smaller models on the first generation and introduce a 5B variant that scales both depth (layers) and width (token embeddings).

Evaluation. We construct task-specific test sets to measure fidelity and generalization, and importantly go beyond existing benchmarks in annotation quality. Each set contains challenging in-the-wild samples. For pose, we evaluate on 11K images annotated with 308 keypoints, in contrast to the 5K capture-studio images used by SAPIENS. For segmentation, we use a similar in-the-wild test of 10K images with 29 classes. For pointmap, normals, and albedo, following Saleh et al. (2025), we evaluate on a 10K-image test set built from our photorealistic assets with higher geometric detail. Please refer to the appendix for additional details.

6.2 PRETRAINING GENERALIZATION: DENSE PROBING

To evaluate zero-shot generalization of the pretrained backbone, we perform dense probing and compare against state-of-the-art vision backbones—Sapiens (Khirodkar et al., 2024), PE (Bolya et al., 2025), and DINOv2 (Oquab et al., 2023)—across a variety of human tasks. For dense probing, we freeze the backbone and lightly train a task-specific decoder with identical hyperparameters across all methods. The tasks vary in their demands: for pose estimation, high-level human semantics aid keypoint localization, whereas for albedo recovery, the backbone must closely capture input appearance. Table 2 reports task-specific metrics across multiple model sizes. Among baselines, DINOv2 is strongest for pose and geometric understanding (e.g., pointmaps), despite its non-human-centric pretraining, likely due to its contrastive objective. Sapiens (Khirodkar et al., 2024), due to its masked-autoencoder pretraining, lacks semantic understanding but performs well on albedo estimation. With our combined pretraining objective, Sapiens2 consistently outperforms the baselines, and the 5B model exhibits strong zero-shot generalization and precision.

6.3 COMPARISON WITH STATE-OF-THE-ART METHODS

To understand performance and generalization across human-centric tasks, we compare our models against task-specific state-of-the-art methods in this section. We provide a brief summary here and refer to the appendix for detailed analysis.

Pose. We compare Sapiens2 with state-of-the-art whole-body top-down pose estimators in Table 3. We retrain baselines on our new keypoint set using recommended settings. Our models substantially improve over the first generation; specifically, Sapiens2-0.8B, despite its smaller parameter count, outperforms larger models due to architectural improvements and broader supervision. Consistent with scaling laws Kaplan et al. (2020), our results show predictable gains with increased scale. Our largest model, Sapiens2-5B, sets a new state of the art for dense 308-keypoint predictions in-the-wild, achieving 82.3 mAP on challenging poses.

Model	Pose	Seg	Pointmap	Normal	Albedo
	mAP ↑	mIoU (%) ↑	L2 ↓	MAE° ↓	MAE ($\times 10^{-2}$) ↓
PE-L (Bolya et al., 2025)	22.7	16.8	0.537	28.7	4.22
PE-H (Bolya et al., 2025)	36.2	19.7	0.529	27.4	4.14
Sapiens-1B (Khirodkar et al., 2024)	25.9	31.8	0.532	26.9	3.85
Sapiens-2B (Khirodkar et al., 2024)	33.6	34.7	0.515	26.0	3.72
DINOv2-H (Oquab et al., 2023)	30.4	37.6	0.468	26.3	4.01
DINOv2-G (Oquab et al., 2023)	42.9	43.7	0.432	24.1	3.92
Sapiens2-0.4B (Ours)	28.3	36.5	0.471	28.7	3.96
Sapiens2-0.8B (Ours)	39.6	39.1	0.435	24.2	3.89
Sapiens2-1B (Ours)	47.4	41.0	0.428	22.6	3.64
Sapiens2-5B (Ours)	**56.7**	**46.2**	**0.358**	**19.4**	**3.12**

Table 2: **Dense probing on human tasks.** We freeze the backbone and fine-tune a lightweight, task-specific decoder with identical hyperparameters across all methods.

Model	Input Size	mAP (%)	mAR (%)
ViTPose+-L, TPAMI23	256 × 192	47.8	53.6
ViTPose+-H, TPAMI23	256 × 192	48.3	54.1
DWPose-M, ICCV23	256 × 192	60.6	67.4
DWPose-L, ICCV23	384 × 288	66.5	72.8
RTMW-L, arxiv23	384 × 288	70.1	75.9
RTMW-X, arxiv23	384 × 288	70.2	76.1
Sapiens-1B*, ECCV24	1024 × 768	76.8	79.3
Sapiens-2B*, ECCV24	1024 × 768	78.3	82.1
Sapiens2-0.4B (Ours)	1024 × 768	76.9	81.3
Sapiens2-0.8B (Ours)	1024 × 768	79.4 (+1.1)	83.1 (+1.0)
Sapiens2-1B (Ours)	1024 × 768	80.4 (+2.1)	84.0 (+1.9)
Sapiens2-5B (Ours)	1024 × 768	**82.3** (+4.0)	**85.3** (+3.2)

Table 3: **Pose estimation** on 11K `test`. Flip test is used, same detections. *Denotes *v1* open-sourced models.

Model	mIoU (%)	mAcc (%)
SegFormer. Neurips21	45.2	68.3
Mask2Former, CVPR22	48.7	71.5
DeepLabV3+, ECCV18	42.8	66.9
HRNetV2+OCR	47.3	70.2
Sapiens-1B*, ECCV24	53.8	74.7
Sapiens-2B*, ECCV24	58.2	77.2
Sapiens2-0.4B (Ours)	74.9 (+16.7)	85.8 (+8.6)
Sapiens2-0.8B (Ours)	78.4 (+20.2)	89.1 (+11.9)
Sapiens2-1B (Ours)	79.3 (+21.1)	88.2 (+11.0)
Sapiens2-1B-4K (Ours)	79.8 (+21.6)	88.9 (+11.7)
Sapiens2-5B (Ours)	**80.5** (+22.3)	**89.3** (+12.1)

Table 4: **Segmentation** on 5K `test`. All methods have the same `train` set. *Denotes *v1* open-sourced models.

Segmentation. Table 4 compares our models to state-of-the-art methods on our segmentation vocabulary. For fairness, we train all baselines on our training set. Sapiens2 generalizes strongly to in-the-wild images with high-resolution outputs. Although the input resolution is the same (1K) for Sapiens and Sapiens2, Sapiens2–1B outperforms Sapiens-1B by 21.1% mIoU and 11% mAcc, owing to in-the-wild supervision and an increased output resolution of 1K (from 0.5K).

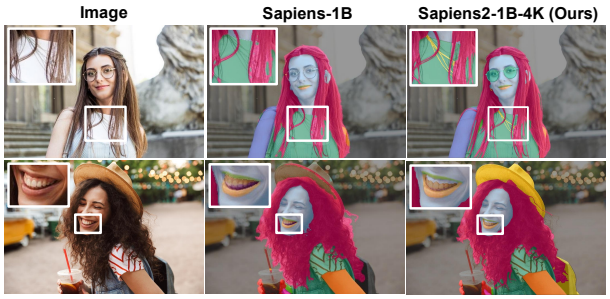


Figure 7: **Body-part segmentation** using our 1B-4K model.

Pointmap. Table 5 compares Sapiens2 with existing pointmap (XYZ) estimation methods such as UniDepth (Piccinelli et al., 2024), DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025a), and MoGe Wang et al. (2025b). This task is more challenging than relative depth estimation, as it requires reasoning about camera intrinsics. For fairness, we optimize for scale and evaluate all predictions in a focal-length-normalized canonical space. Interestingly, MoGe outperforms Sapiens2-0.4B despite being trained without human-centric data due to its affine-invariant design. However, our larger models substantially outperform all baselines. Fig. 8 qualitatively compares Sapiens2-1B with MoGe, showing that our predicted pointmaps better preserve human-specific geometric details.

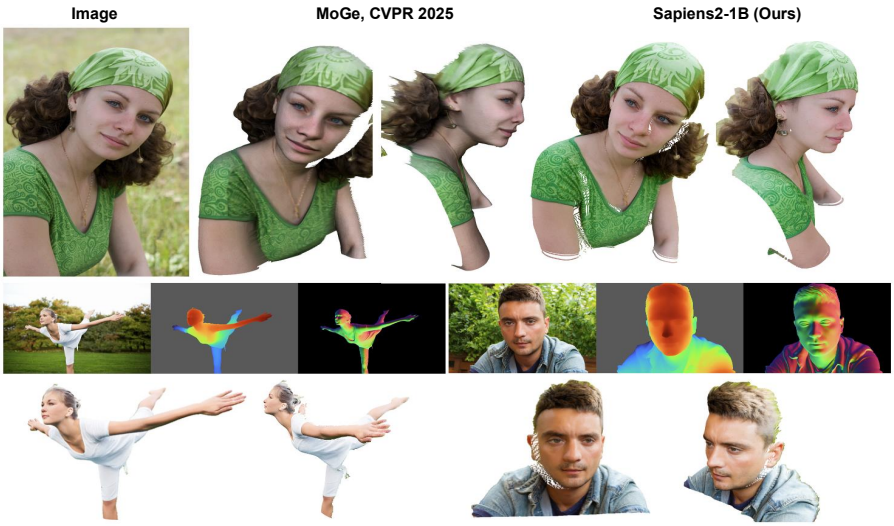


Figure 8: (*Top*) **Pointmap** qualitative comparison of Sapiens2-1B with MoGe (Wang et al., 2025b). (*Bottom*) Depth visualized from the predicted pointmap, along with surface normals and novel 3D viewpoints.

Method	Distance		Abs. Error (e^{-3})		
	L2	RMSE	X	Y	Z (cam)
UniDepth, CVPR24	0.368	0.689	8.34	10.92	5.23
DUSt3R, CVPR24	0.349	0.663	7.66	10.11	4.86
VGGT, CVPR25	0.217	0.515	3.79	4.96	2.19
MoGe, CVPR25	0.202	0.486	3.21	4.41	1.89
Sapiens2-0.4B (Ours)	0.203	0.491	3.24	4.44	1.90
Sapiens2-0.8B (Ours)	0.199	0.469	3.10	4.29	1.86
Sapiens2-1B (Ours)	0.194	0.492	3.12	4.24	1.81
Sapiens2-5B (Ours)	**0.181**	**0.470**	**2.93**	**3.99**	**1.69**

Table 5: **Pointmap evaluation** in focal-length normalized canonical coordinates on 10K `test`.

Method	Angular Error°		% Within t°		
	Mean	Median	5°	11.25°	30°
Marigold, CVPR24	18.83	15.27	9.41	39.87	45.21
DSINE, CVPR24	17.24	13.51	11.67	45.62	48.79
Sapiens-1B* ECCV24	13.62	10.11	32.18	69.34	82.14
Sapiens-2B* ECCV24	12.38	9.46	37.05	70.54	85.62
DAViD-L, ICCV25	10.73	7.49	42.91	72.16	89.27
Sapiens2-0.4B (Ours)	9.26	5.77	44.35	74.51	94.43
Sapiens2-0.8B (Ours)	8.78	5.25	47.63	76.24	94.77
Sapiens2-1B (Ours)	7.98	4.72	53.38	78.69	95.35
Sapiens2-1B-4K (Ours)	7.84	4.05	54.14	79.10	**95.46**
Sapiens2-5B (Ours)	**7.59**	**3.71**	**57.87**	**80.06**	95.23

Table 6: **Normal evaluations** on 10K whole-body `test` set at 4K ground-truth resolution.

Figure 9: **Normal prediction**. Qualitative comparison of Sapiens2-1B with DAViD (Saleh et al., 2025).

Normal. We compare our finetuned normal estimators with current state-of-the-art monocular methods in Table 6. Our evaluation set consists of whole-body scan images captured from random virtual camera viewpoints, with ground-truth normals available at $4K$ resolution. Our smallest model, Sapiens2-0.4B, outperforms existing methods by achieving a mean angular error of $9.26°$, with 94.43% of human pixels below the $22.5°$ threshold. Fig. 9 compares Sapiens2 with the baseline DAViD Saleh et al. (2025) and shows that it captures geometric details accurately and remains robust under varying lighting conditions.

Albedo. We present qualitative results of our 1B albedo estimation model in Fig. 10. Our model recovers true skin tone under varying lighting conditions, despite being trained solely on synthetic data. Additionally, compared to diffusion-based methods Liang et al. (2025), our feedforward model is significantly faster. Please refer to the appendix for the quantitative comparisons.

Figure 10: **Albedo estimation** using Sapiens2-1B. Our model effectively encodes low-level details crucial for albedo estimation and generalizes well to in-the-wild images, despite being trained on limited synthetic data.

Limitations and Declaration. Sapiens2 data sources typically contain 1–4 humans, with a prominent subject present in the image during both pre- and post-training. As a result, the models do not perform well in crowded, multi-person scenarios. All data used in our experiments was collected under appropriate licensing. Additionally, we used an LLM for light grammatical edits to the paper.

7 Conclusion

Sapiens2 introduces high-resolution, human-centric models pretrained on a 750-million-image dataset. Our models simultaneously learn appearance cues and semantics by combining masked reconstruction and contrastive objectives. They consistently outperform general-purpose models on human images and extend to tasks ranging from pose estimation to albedo recovery. Sapiens2 sets a new benchmark for high-fidelity dense predictions and provides a robust foundation for applications requiring a nuanced, detailed understanding of humans in unconstrained visual contexts.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Reza Azad, Moein Heidary, Kadir Yilmaz, Michael Hüttemann, Sanaz Karimijafarbigloo, Yuli Wu, Anke Schmeink, and Dorit Merhof. Loss functions in the era of semantic segmentation: A survey and outlook. *arXiv preprint arXiv:2312.05391*, 2023.

Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9535–9545, 2024.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Aleksei Bochkovskii, AmaÃĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

M Caron, H Touvron, I Misra, H Jégou, J Mairal, P Bojanowski, and A Joulin. Self-supervised vision transformers with dino. *GitHub Repository. Available online: https://github.com/facebookresearch/dino (accessed on 11 August 2022)*.

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*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15050–15061, 2023.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7103–7112, 2018.

Ziwei Chen, Qiang Li, Xiaofeng Wang, and Wankou Yang. Liftedcl: Lifting contrastive learning for human-centric perception. In *The Eleventh International Conference on Learning Representations*, 2022.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pp. 7480–7512. PMLR, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. *arXiv preprint arXiv:2205.03892*, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. Query-key normalization for transformers. *arXiv preprint arXiv:2010.04245*, 2020.

Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *International Journal of Computer Vision*, pp. 1–47, 2025.

Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Versatile multi-modal pre-training for human-centric perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16156–16166, 2022.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2506–2517, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024.

Gwanghyun Kim, Xueting Li, Ye Yuan, Koki Nagano, Tianye Li, Jan Kautz, Se Young Chun, and Umar Iqbal. Geoman: Temporally consistent human geometry estimation using image-to-video diffusion. *arXiv preprint arXiv:2505.23085*, 2025a.

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25096–25106, 2024.

Minchul Kim, Dingqiang Ye, Yiyang Su, Feng Liu, and Xiaoming Liu. Sapiensid: Foundation for human recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13937–13947, 2025b.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusion renderer: Neural inverse and forward rendering with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26069–26080, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xiaoding Lu, Zongyi Liu, Adian Liusie, Vyas Raina, Vineet Mudupalli, Yuwen Zhang, and William Beauchamp. Blending is all you need: Cheaper, better alternative to trillion-parameters llm. *arXiv preprint arXiv:2401.02994*, 2024.

Yihang Meng, Hao Cheng, Zihua Wang, Hongyuan Zhu, Xiuxian Lao, and Yu Zhang. Efficient multi-modal human-centric contrastive pre-training with a pseudo body-structured prior. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 114–128. Springer, 2024.

AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Zimin Ran, Xingyu Ren, Xiang An, Kaicheng Yang, Xiangzi Dai, Ziyong Feng, Jia Guo, Linchao Zhu, and Jiankang Deng. High-fidelity facial albedo estimation via texture quantization. *arXiv preprint arXiv:2406.13149*, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, pp. 29441–29454. PMLR, 2023.

Fatemeh Saleh, Sadegh Aliakbarian, Charlie Hewitt, Lohit Petikam, Antonio Criminisi, Thomas J Cashman, Tadas Baltrušaitis, et al. David: Data-efficient and accurate vision models from synthetic data. *arXiv preprint arXiv:2507.15365*, 2025.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023.

Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5261–5271, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9043–9053, 2023.

Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. Hap: Structure-aware masked image modeling for human-centric perception. *Advances in Neural Information Processing Systems*, 36:50597–50616, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 405–420, 2018.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM computing surveys*, 56(1):1–37, 2023.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A APPENDIX

A.1 PRETRAINING

A.1.1 IMPLEMENTATION DETAILS

We use the dense-probing evaluations as the final metrics to guide any design decisions during the pretraining stage. For instance, we pretrain the SAPIENS2–1B (embed dim 1536, 40 layers, 24 heads, patch size 16, final norm with [CLS]) at 1024×768. Training uses a joint MAE and contrastive objective: an 8-layer MAE decoder (dim 512) with ℓ_2 reconstruction, and a [CLS] projection head for contrastive learning. Loss weights are MAE: 1.0, CLS: 0.4, KoLeo: 0.04. We adopt multi-view training with 2 global and 4 local crops; global crops use random resize–crop in ratio $[0.5, 1.0]$, local crops in $[0.2, 0.7]$, with standard color/blur/solarize and horizontal flips. Inputs are normalized to ImageNet means/stds. Importantly, we do not use color augmentations on the global views - used for masked reconstruction objective.

Optimization uses fused AdamW (lr 1×10^{-4}, $(\beta_1, \beta_2)=(0.9, 0.95)$, wd 0.05) with zero-decay for norms, biases, positional and special tokens. We train for 5×10^5 iters with 10^3 warmup, cosine decay to 10^{-7}, and global grad-norm clip 5.0. The contrastive teacher EMA is 0.992 (center momentum 0.9); student temperature is 0.1, teacher temperature warms from 0.065 to 0.07 over the first 10^3 iters. We evaluate every checkpoint for downstream tasks with a frozen encoder and report results using the best checkpoint.

A.1.2 MASKING STRATEGY

Given the high resolution of our backbones, we use mixed blockwise/patchwise masking (blockwise prob 0.4) with a 75% mask ratio at patch size 16, refer Fig. 11. At 1024×768 ($64\times48=3072$ patches), this masks $\sim$ 2304 patches per image, yielding coarse occlusions that regularize MAE while leaving sufficient context for contrastive learning.

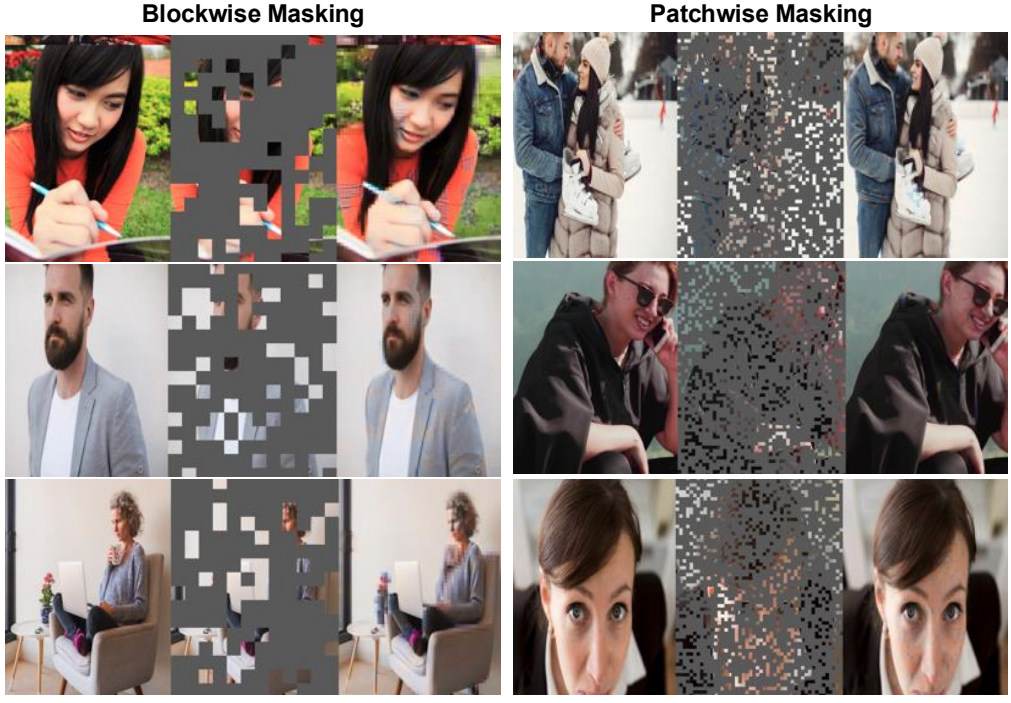


Figure 11: We randomly mix blockwise and patchwise masking to provide coarse occlusions. For MAE pretraining at high resolution (1024), we use a 75% mask ratio. Each sample represents (ground-truth image, masked input, reconstruction).

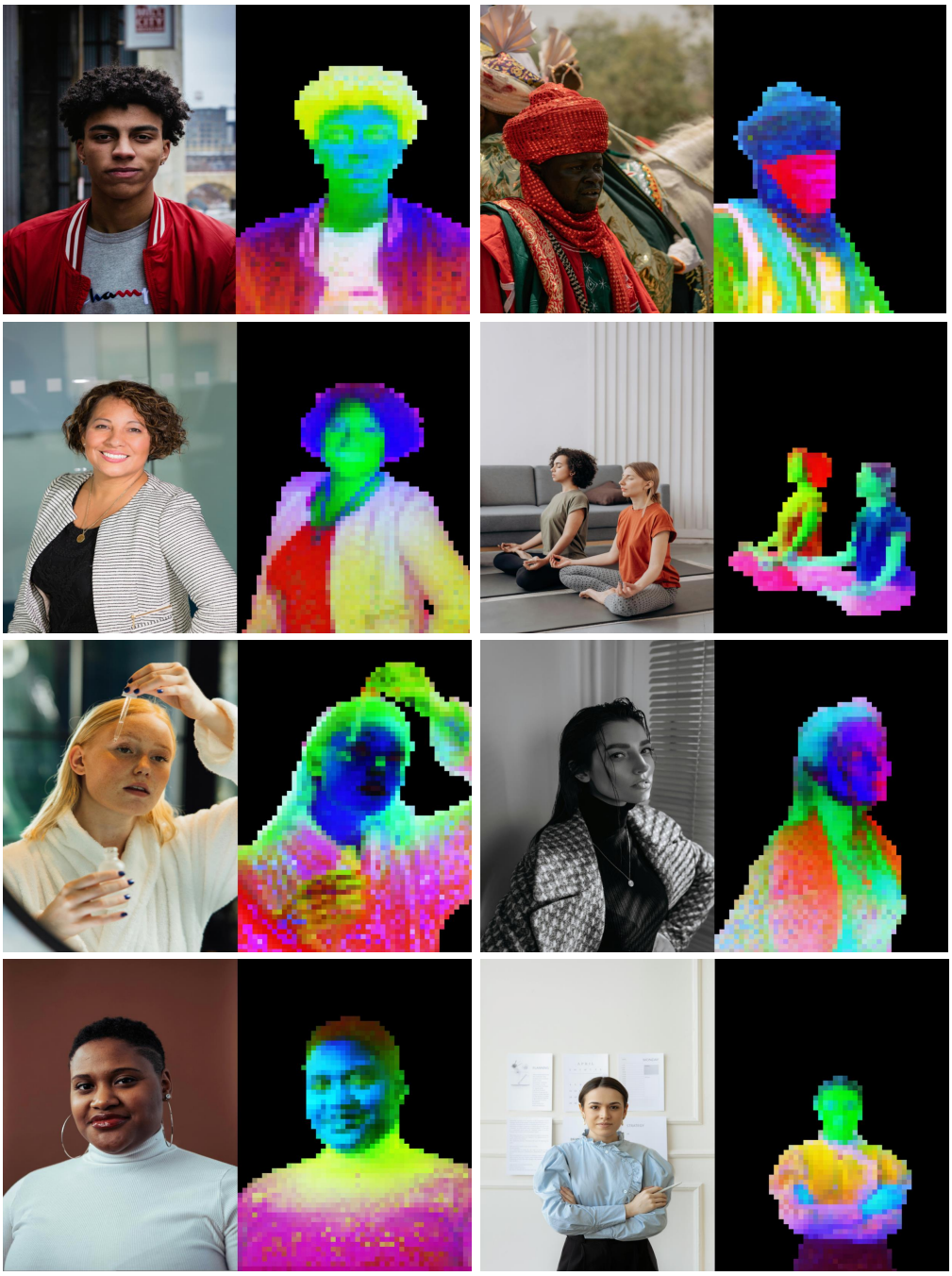


Figure 12: We visualize the encoder features using PCA (3 major components) with different colors. We use foreground masking to extract patch features for human pixels. Sapiens2 features capture texture and color information as well as showcase human semantics.

A.2 POSE ESTIMATION

We evaluate Sapiens2 using ground-truth bounding boxes on our in-the-wild `test` set for 308 keypoints. We fine-tune a top-down pose estimator initialized from a pretrained checkpoint with the `[CLS]` token disabled so the encoder outputs a feature map. The head is a heatmap decoder with in-channels 1536 and out-channels 308 (keypoints). It uses two deconvolution stages (kernel 4, stride 2) for $4\times$ upsampling, followed by 1×1 convolutions with channels $(768, 768, 512)$ and a final 1×1 projection to 308 heatmaps. We adopt UDP heatmaps (stride 4, σ=6) and optimize a weighted MSE loss. At test time, we enable flip testing with heatmap fusion.

Optimization uses AdamW (lr 5×10^{-4}, (β_1, β_2)=$(0.9, 0.999)$, weight decay 0.1) with layer-wise learning-rate decay and zero weight decay for biases, positional embeddings, relative position biases, and norms. We clip gradients to a global ℓ_2 norm of 1.0. The schedule warms up linearly for 500 iterations (start factor 10^{-3}), then follows polynomial decay (`power` 1.0) for the remainder. In addition to the main table, we provide fine-grained evaluations in Table 7, which compares Sapiens2 with Sapiens.

Model	Foot		Face		Left Hand		Right Hand		Whole Body	
	AP	AR	AP	AR	AP	AR	AP	AR	AP	AR
Sapiens-0.3B	72.1	77.6	82.4	86.7	66.8	72.9	67.3	73.2	70.5	77.0
Sapiens-0.6B	73.8	78.9	83.9	87.8	68.4	74.1	69.0	74.5	72.8	78.6
Sapiens-1B	75.0	80.1	85.1	88.6	69.7	75.3	70.2	75.7	74.1	79.4
Sapiens-2B	76.0	81.0	86.0	89.2	70.9	76.4	71.3	76.8	75.3	80.4
Sapiens2-0.4B	78.4	82.0	86.2	89.5	75.1	79.0	75.6	79.4	76.9	81.3
Sapiens2-0.8B	80.1	83.4	87.6	90.4	76.8	80.3	77.2	80.7	79.4	83.1
Sapiens2-1B	81.0	84.1	88.3	90.9	77.6	81.0	78.0	81.3	80.4	84.0
Sapiens2-5B	82.6	85.3	89.7	91.8	79.2	82.4	79.6	82.7	82.3	85.3

Table 7: Pose estimation results on 10K `test` set (K=308). Flip test is used.

Figure 13: In addition to in-the-wild annotations we also use capture-studio 3D triangulated ground-truth 308 keypoints for finetuning Sapiens2.

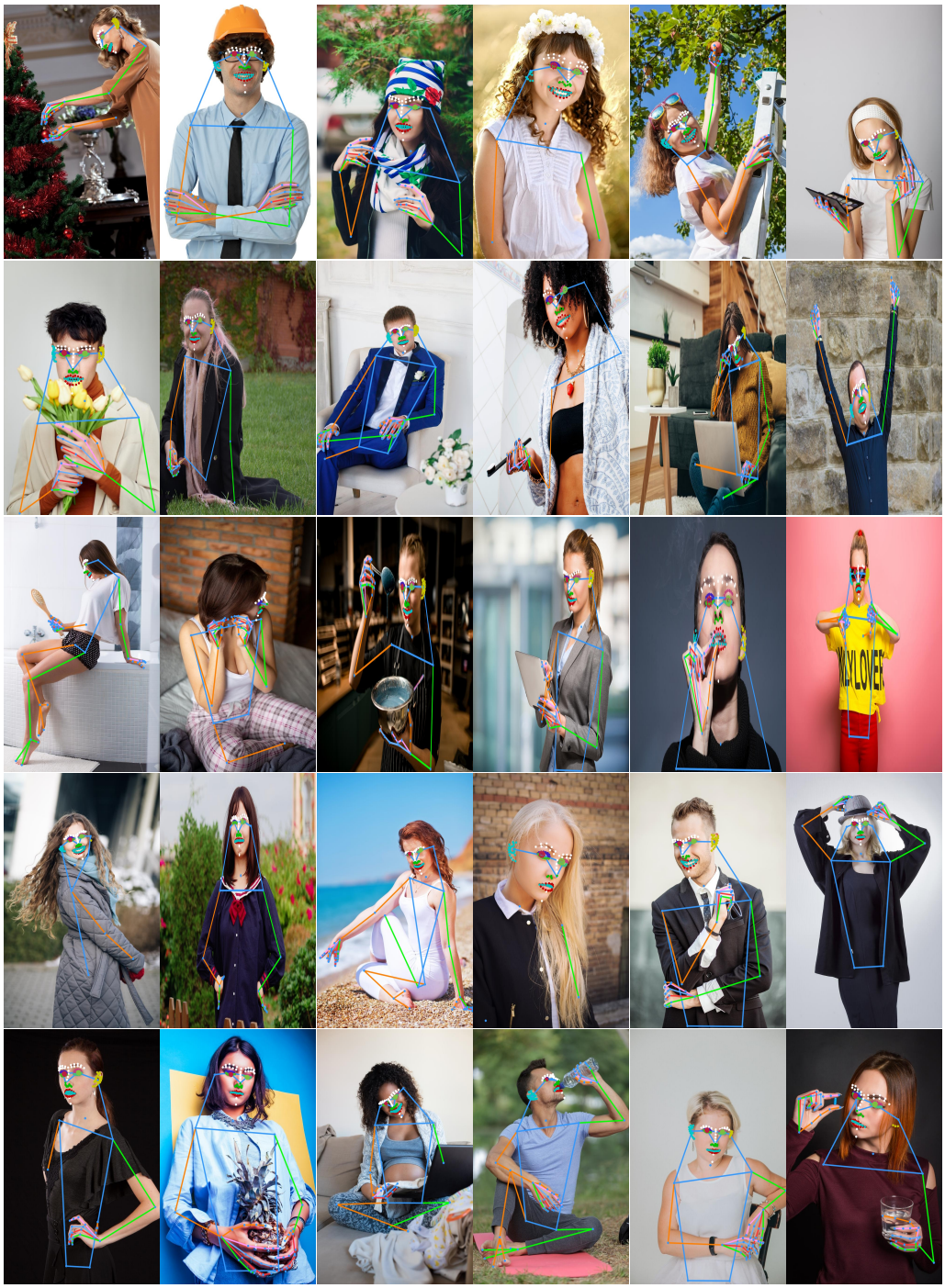


Figure 14: Top-down 308 keypoint predictions using Sapiens2-1B model on in-the-wild images.

A.3 BODY-PART SEGMENTATION

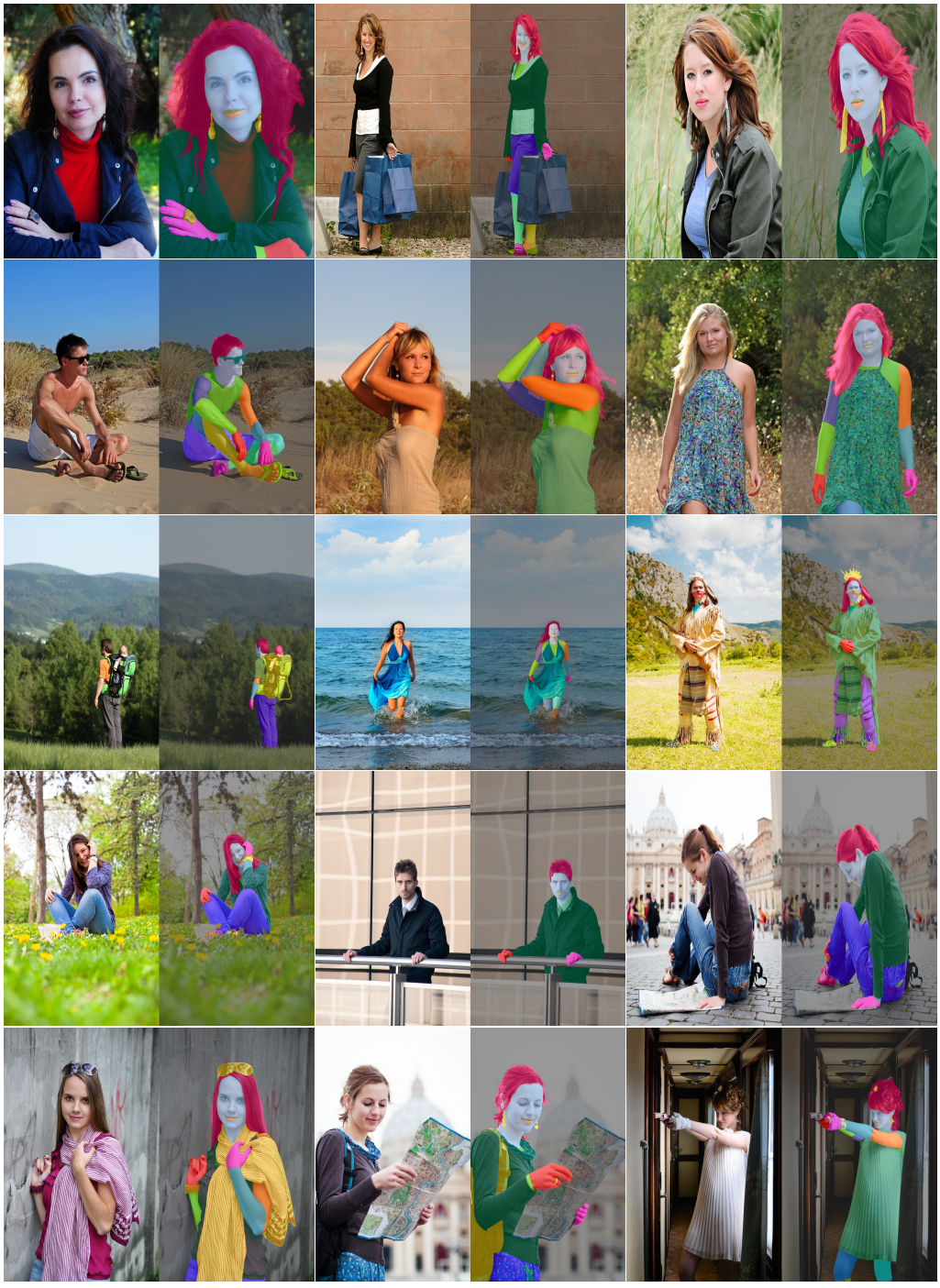


Figure 15: Body-part segmentation (29 classes) using Sapiens2-1B on real-world images.

A.4 POINTMAP ESTIMATION

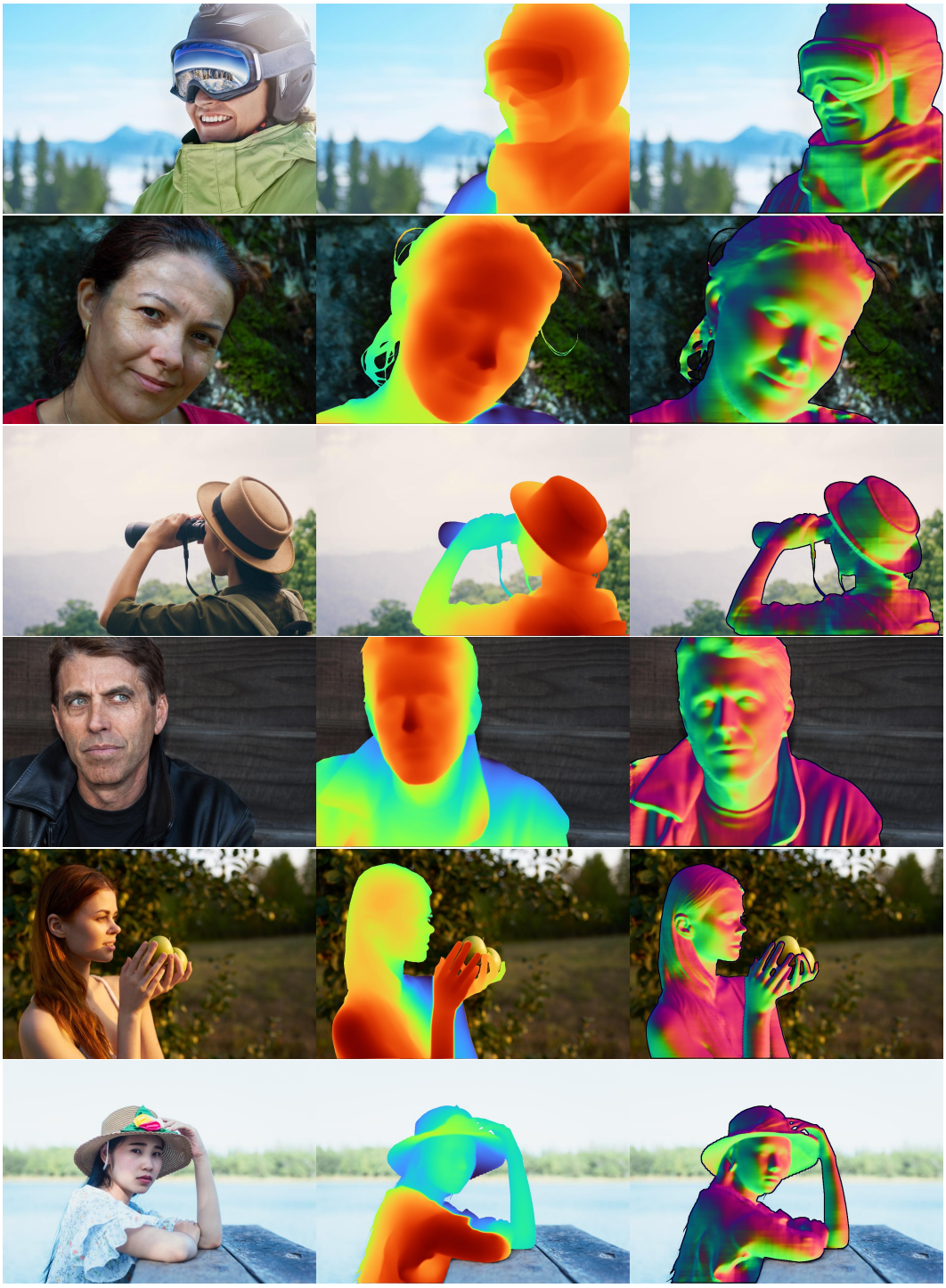


Figure 16: Pointmap using Sapiens2-1B. For each image, we visualize the absolute depth derived from the predicted XYZ pointmap as a heatmap and surface normals computed from depth.

Figure 17: Pointmap using Sapiens2-1B. For each image, we visualize the absolute depth derived from the predicted XYZ pointmap as a heatmap and surface normals computed from depth.

A.5 NORMAL ESTIMATION

Figure 18: Surface normal prediction using Sapiens2-1B.

A.6 ALBEDO ESTIMATION

We report quantitative results on our 10K `test` set for albedo estimation in Table 8 and qualitative results in Fig. 19.

Model	MAE	RMSE	PSNR	SSIM	Grad-L1
Sapiens2-0.4B	0.01825	0.03257	29.74412	0.88937	0.00642
Sapiens2-0.8B	0.01602	0.02876	30.82573	0.90337	0.00624
Sapiens2-1B	0.01224	0.02392	32.42509	0.91410	0.00612
Sapiens2-5B	0.01191	0.02341	32.61346	0.91524	0.00610

Table 8: Quantative evaluations for albedo estimation using Sapiens2 models.

Figure 19: Albedo (base color) prediction using Sapiens2-1B at 1024 × 768 resolution.