FROM RGB GENERATION TO DENSE FIELD READOUT: PIXEL-SPACE DENSE PREDICTION WITH TEXT-TO-IMAGE MODELS

Anonymous authors
Paper under double-blind review

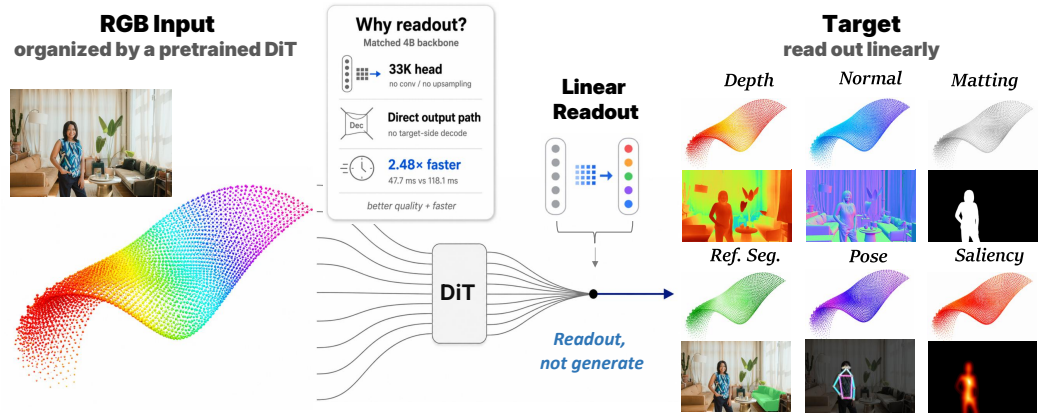

Figure 1: **ReChannel: readout, not generation.** A pretrained DiT organizes RGB inputs into a patch-aligned spatial token field, enabling direct readout of task-native quantities on the same image plane rather than RGB-style target reconstruction. Each output patch acts as a spatial carrier whose channels are reinterpreted from RGB appearance to depth, surface normals, matting, referring segmentation, pose, and saliency, without target-side VAE decoding.

ABSTRACT

Large-scale text-to-image models are attractive backbones for dense prediction because RGB generation pretraining learns rich semantic, structural, and geometric priors. Representative generative and editing approaches reuse these priors by casting dense prediction as target generation: annotations such as depth, normals, alpha mattes, masks, and heatmaps are encoded into an RGB-trained VAE latent space and decoded back as image-like targets. This target-side interface is inherited from RGB synthesis rather than derived from the task. Unlike RGB synthesis, dense prediction asks for pixel-correct, task-native fields on the same image plane, not new RGB content to be rendered. Our starting point is that a pretrained DiT already organizes RGB inputs through a patch→token→patch lattice on the image plane. Each token therefore indexes a fixed output patch whose channels can carry task-native quantities instead of RGB appearance. We instantiate this idea as **ReChannel**. We retain the VAE encoder to preserve the DiT's input distribution, remove the target-side decoder, adapt the frozen DiT with per-task LoRA, and map each token to its $p \times p \times K_t$ pixel-space patch using a shared token-local linear head. The head has about 33K parameters, performs no spatial mixing, and is trained directly with pixel-space, task-native supervision. Using FLUX-Klein, we evaluate ReChannel on six dense prediction tasks and more than a dozen benchmarks. This minimal interface achieves state-of-the-art results on trimap-free matting, KITTI depth, and referring segmentation, while remaining competitive on normals, saliency, and pose. Under a matched 4B backbone, it is more accurate and up to 2.48× faster than latent, VAE-decoded, and edit output interfaces, while larger heads provide no further gains. These results show that dense prediction can inherit the prior of generative pretraining without inheriting its output interface. Code and models will be released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

RGB-based dense prediction begins with a simple constraint: the RGB image is the only visual observation from which pixel-aligned fields such as geometry, masks, mattes, and heatmaps are estimated. The dominant discriminative paradigm addresses this problem by pairing pretrained visual encoders with task-specific decoders, increasingly benefiting from foundation representations such as DINO, DINOv2, and SAM (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025; Kirillov et al., 2023). These models provide strong RGB-conditioned representations and produce task-native dense outputs through learned decoders in pixel space (Ranftl et al., 2021; Cheng et al., 2022; Yang et al., 2024; Xu et al., 2022).

Text-to-image models offer a different source of visual priors through large-scale generative pretraining (Black Forest Labs, 2026; Team et al., 2025; Wu et al., 2025; Seedream et al., 2025; Lin et al., 2026). Their RGB synthesis objective learns rich image-formation priors over appearance, semantics, structure, and layout, and probing studies show these priors already expose dense structure such as depth, saliency, and correspondence to simple readouts (Chen et al., 2023; Baranchuk et al., 2022; Tang et al., 2023; Du et al., 2023). Recent generative dense predictors have reused these priors across depth, surface normals, segmentation, and matting (Ke et al., 2024; Fu et al., 2024; Wang et al., 2025b; He et al., 2025), evolving rapidly from iterative diffusion to deterministic one-step prediction and image-to-image editing (Xu et al., 2025a; He et al., 2025; Wang et al., 2025a; Shi et al., 2025). Yet representative VAE-based generative and editing approaches keep the same target-side output interface: the dense target is represented as an image-like or latent target, passed through the latent space of an RGB-trained VAE, and decoded back to the final pixel-space prediction. This is natural for RGB synthesis, but indirect for dense prediction: task-native targets must be fit into an RGB-trained latent space, while the supervision, prediction, and evaluation of interest are pixel-space fields. As illustrated in Fig. 1, we ask whether dense prediction should inherit this target-side RGB reconstruction path, or whether a DiT at generative scale already organizes the image well enough that task-native fields can be read out directly.

We argue the output interface should reflect the asymmetry between the two problems: RGB generation must render a complex appearance field of color, texture, illumination, and high-frequency detail, whereas dense prediction re-expresses the same visual evidence as task-native pixel-space fields such as geometry, opacity, or heatmaps (Ranftl et al., 2021; Yang et al., 2024; Xu et al., 2022). Our observation is structural. ViTs establish a spatial lattice of image-patch tokens, and DiTs extend it into a patch-to-token-to-patch transformation for generation (Dosovitskiy et al., 2020; Peebles & Xie, 2023). An output patch is thus a spatial carrier: its channels are interpreted as RGB appearance in generation, but the same carrier can hold task-native quantities for dense prediction. The backbone itself already organizes the image into this spatially indexed field, so the remaining output interface need not reconstruct spatial structure; the minimal readout admitted by the lattice is a reinterpretation of each carrier's channels. Section 4.3 tests this view under matched backbones and finds this minimal readout is sufficient: larger heads bring no accuracy gains, while full fine-tuning provides no consistent improvement and incurs substantially higher training cost.

As shown in Fig. 2(b), we instantiate this insight as **ReChannel**. The pretrained VAE encoder is retained to preserve the latent input distribution expected by the DiT, while dense targets never enter the VAE. Lightweight per-task LoRA adapts the generation-pretrained token field toward the target semantics, and a shared token-local linear projection maps each adapted token to its $p \times p \times K_t$ target patch under pixel-space supervision. The adapted token field provides the spatial structure; the readout only converts channels, with no spatial decoder and no target-side reconstruction path.

We validate this recipe on six dense prediction tasks and more than a dozen benchmarks, spanning targets as different as continuous geometry, high-frequency mattes, language-conditioned masks, and multi-channel heatmaps. The same token-local readout, varying only the output channels K_t, matches or surpasses specialized and generative baselines across these tasks (Sec. 4.2).

Our contributions are threefold:

- **Rechanneling, not generating. ReChannel** treats each DiT output patch as a spatial carrier and reinterprets its channels from RGB to task-native fields, with no target-side VAE or spatial decoder.

- **One RGB prior, all dense fields.** Per-task LoRA and a token-local linear readout suffice to serve geometry, mattes, language-conditioned masks, and heatmaps, varying only the output channels K_t.
- **Strong, efficient results.** State-of-the-art on trimap-free matting, KITTI depth, and referring segmentation, up to 2.48× faster than latent, VAE-decoded, and edit interfaces under a matched 4B backbone.

2 Related Work

From task decoders to unified visual backbones. Dense prediction has long followed a feature-to-field paradigm, with task-specific targets, supervision, and output heads for depth, normals, matting, segmentation, and pose. Multi-task studies explore transfer and shared representations across heterogeneous visual tasks (Zamir et al., 2018; Eftekhar et al., 2021; Bachmann et al., 2022). More recent foundation models provide stronger transferable features, including DINO, DINOv2, DINOv3, and SAM (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025; Kirillov et al., 2023). Dense prediction systems such as DPT, MaskFormer, Mask2Former, OneFormer, X-Decoder, Depth Anything, and ViTPose then pair such backbones with learned decoders for pixel-level tasks (Ranftl et al., 2021; Cheng et al., 2021; 2022; Jain et al., 2023; Zou et al., 2023; Yang et al., 2024; Xu et al., 2022). This establishes a common recipe: pretrained representations plus downstream spatial decoding. ReChannel asks whether a generative backbone still needs an explicit spatial decoder after task adaptation.

Generative priors for dense perception. Text-to-image models provide rich visual priors through large-scale RGB synthesis pretraining. Diffusion representations expose semantic, geometric, and spatial structure useful for downstream perception (Chen et al., 2023; Tang et al., 2023; Du et al., 2023; Luo et al., 2023; Zhang et al., 2023; Hedlin et al., 2023; Xu et al., 2023; Zhao et al., 2023; Kondapaneni et al., 2024; Tian et al., 2024). This has motivated direct reuse for dense prediction. Marigold and GeoWizard adapt diffusion priors to monocular geometry (Ke et al., 2024; Fu et al., 2024), while PrimeDepth extracts single-step diffusion representations for depth estimation (Zavadski et al., 2024). GenPercept favors deterministic one-step adaptation over iterative sampling (Xu et al., 2025a), and Lotus further moves toward single-step annotation prediction (He et al., 2025). More recently, FE2E and Edit2Perceive exploit pretrained editing priors for dense geometry and perception (Wang et al., 2025a; Shi et al., 2025). These works simplify generation but still rely on spatial decoders or generative target interfaces. ReChannel instead directly reads task-native fields from the generative prior.

Output interfaces and patch-indexed generation. A complementary direction pushes the generative view further and treats image generation itself as a general-purpose interface for visual understanding, expressing heterogeneous tasks through RGB image outputs (Gabeur et al., 2026). In parallel, the ViT/DiT family represents images as spatially indexed token lattices, and pixel-space generators such as JiT show that such patch-indexed transformers can model image-plane quantities without relying on a VAE tokenizer or decoder (Dosovitskiy et al., 2020; Peebles & Xie, 2023; Li & He, 2026). These developments suggest that the representation learned by a generative model and the interface used to expose its predictions need not be identical. Our work studies this distinction specifically for deterministic dense prediction. Rather than rendering a task target as an RGB-like image or reconstructing it through a generative decoder, ReChannel treats each adapted DiT token as a spatial carrier and directly reinterprets its channels as task-native quantities. In this sense, our question is not whether generative pretraining is useful for perception, but whether inheriting a generative prior also requires inheriting its generative output interface.

3 Method

3.1 Dense Prediction as Field Readout

We study the output interface between a pretrained text-to-image backbone and deterministic dense prediction. Existing generative approaches expose dense targets through an image-generation interface: the target is represented as an image-like output or latent and, when applicable, decoded back to pixels (Fig. 2a). This interface is natural for RGB synthesis, whereas dense prediction is

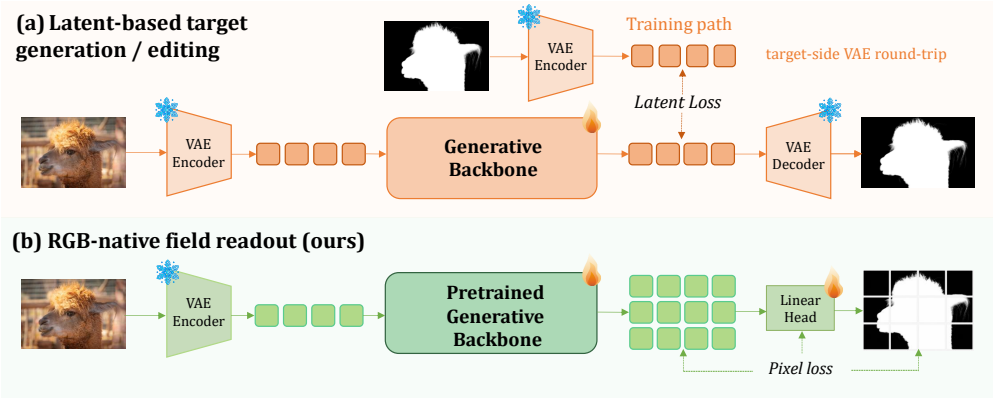

Figure 2: **ReChannel vs. target-side generation.** (a) Existing generative and editing interfaces treat dense targets as image-like outputs: the target is encoded into an RGB-trained VAE latent, supervised there, and decoded back through a target-side reconstruction path. (b) ReChannel instead keeps the RGB input interface but drops the target-side path: the pretrained backbone acts as an RGB-native field organizer, and a lightweight linear readout maps its spatial tokens directly to task-native pixel-space patches under pixel supervision—field readout rather than target generation.

supervised and evaluated as task-native quantities on the image plane. We therefore ask what output interface is actually necessary once a strong generative prior has already organized the input.

Our view separates three roles that are often coupled in generative adaptation: the *input interface*, which maps RGB observations into the representation expected by the pretrained model; the *backbone representation*, which organizes the observation into a spatial token field; and the *output interface*, which converts that field into the task target. ReChannel keeps the first two components inherited from generative pretraining, while replacing only the last. The dense target is therefore not required to imitate an RGB image or pass through the target-side generative path.

Concretely, we treat the spatial token grid of the T2I transformer as the readout substrate for dense prediction. After lightweight adaptation, each token is mapped to its corresponding local pixel-space target patch by a shared linear projection: no generative decoder, no high-capacity spatial head, and no inter-token output-side processing. The token field carries the spatial organization; the readout only converts the representation at each indexed location into task-native channels.

This gives **ReChannel** (Fig. 2b): the pretrained RGB input pathway is preserved, but dense targets never enter the VAE. Each adapted token is read out directly into its pixel-space target patch, making the output interface task-native rather than generative.

3.2 Adapting the RGB Field

Given an input image x, we follow the original input pathway of the pretrained T2I model and encode the RGB image with its VAE encoder. This preserves the latent input distribution and coordinate system on which the generative backbone was pretrained. We then run the DiT backbone in a deterministic zero-noise mode, freezing the pretrained parameters θ and inserting only task-specific LoRA parameters Δ_t:

$$Z_t = F_{\theta+\Delta_t}\big(\mathrm{VAE}_{\mathrm{enc}}(x); \sigma = 0, c_t\big), \tag{1}$$

where F_θ denotes the pretrained T2I transformer, c_t is the task condition, and $Z_t = \{z_{ij}^t\}$ is the resulting spatial token field for task t.

The role of adaptation is deliberately separated from that of the output head. The pretrained field is spatially organized but not necessarily aligned with the semantics of depth, normals, opacity, masks, or heatmaps. LoRA therefore reorients the pretrained representation toward the target semantics, while the final dense prediction is still produced by the minimal readout in Sec. 3.3. This distinction is important: ReChannel is not a frozen linear probe of generative features, but a lightweight adaptation of the pretrained field followed by a minimal task-native output interface.

The VAE is used only on the input side. This keeps the DiT operating on the latent representation for which it was pretrained, while avoiding any requirement that the dense target itself be encoded into or reconstructed from the RGB-trained latent space. Consequently, task supervision remains entirely in the native pixel-space representation.

3.3 Token-Local Pixel Readout

After task adaptation, each spatial token is mapped directly to a local dense patch by a shared token-local linear projection:

$$\hat{Y}^t_{ij} = \mathrm{reshape}\left(W_t z^t_{ij} + b_t\right), \quad \hat{Y}^t_{ij} \in \mathbb{R}^{p\times p\times K_t}. \tag{2}$$

Here p is the output patch size and K_t is the number of task-native output channels. Tiling the predicted patches over the spatial lattice produces the final dense field $\hat{Y}^t$.

Each token z^t_{ij} is spatially indexed to a local image region Ω_{ij}, and the readout predicts the target values for the same region. The spatial correspondence is therefore inherited from the token lattice rather than reconstructed by a learned decoder. In this sense, the token acts as a spatial carrier: generation interprets its channels as appearance to be rendered, whereas ReChannel interprets them as task-native quantities associated with the same location.

The projection W_t is shared across all spatial positions and is applied independently to each token. It contains no convolution, upsampling, cross-token attention, or other learned spatial mixing. The reshape operation only rearranges the projected channels into a $p \times p$ patch and introduces no additional parameters. Any spatial structure in the final prediction must therefore already be present in the adapted token field.

For a training pair (x, Y^t), we optimize the task-specific adapter and readout directly under the native dense-prediction loss,

$$\mathcal{L}_t = \ell_t\left(\hat{Y}^t, Y^t\right), \tag{3}$$

where ℓ_t is the conventional loss for task t. No latent reconstruction objective or target-side RGB reconstruction loss is introduced. This keeps supervision, prediction, and evaluation in the same task-native space.

We call this interface **ReChannel**: it reinterprets the channels of each spatial carrier from RGB appearance to task-native fields while leaving spatial organization to the adapted token field. That such a purely token-local output map is sufficient—without a spatial decoder, full fine-tuning, or a generative target path—is the central empirical claim tested in Sec. 4.2 and isolated in Sec. 4.3.

3.4 Task Instantiations

The same readout form is used across tasks. For each task, we train a task-specific LoRA adapter and readout projection, with the output channel dimension K_t and supervision loss chosen according to the target representation. Referring segmentation additionally uses the referring expression as the text condition. Scalar fields such as depth, alpha, and saliency use $K_t = 1$; surface normals use $K_t = 3$; pose estimation uses multi-channel heatmaps. Continuous regression tasks are trained with their standard pixel-space losses, while binary mask tasks use mask supervision in pixel space. We keep these choices conventional, since the focus of our study is the output interface rather than task-specific loss design.

4 Experiments

4.1 Setup

We use FLUX-Klein (4B and 9B variants) as the pretrained text-to-image backbone and freeze it for all main results, training a single LoRA adapter and a token-local linear head per task with deterministic σ=0 inference (Black Forest Labs, 2026). We use standard task-specific training sets with no additional data and evaluate under each benchmark's standard protocol: depth is trained on Hypersim + Virtual KITTI (the Marigold protocol (Ke et al., 2024)) and evaluated on NYU (Eigen

split), KITTI (Garg crop), and ScanNet val with log-space scale-shift alignment; surface normals are trained on Hypersim and evaluated on the DSINE NYU / ScanNet / iBims split; matting is trained on P3M-10k and evaluated on P3M-500-P, P3M-500-NP, and zero-shot AIM-500; referring segmentation is trained on RefCOCO/ +/ g and evaluated on the LISA/GLaMM 8-split protocol with cumulative cIoU; saliency is trained on DUTS-TR and evaluated on DUTS-TE and ECSSD with the canonical PySODMetrics $F_{\max}$; and pose is trained on COCO train2017 and evaluated on COCO val with the YOLOv11x detection-box protocol.

Baselines. For depth and normals we compare against full-strength specialist systems (Omnidata v2, DAv2, DSINE (Eftekhar et al., 2021; Yang et al., 2024; Bae & Davison, 2024)) and the generative line from multi-step diffusion through one-step prediction to editing-based adaptation (GeoWizard, Marigold, GenPercept, Lotus, Edit2Perc (Fu et al., 2024; Ke et al., 2024; Xu et al., 2025a; He et al., 2025; Shi et al., 2025)); GenPercept also serves as the generative baseline for matting. Trimap-free matting baselines span two-stage cascades, real-time decomposition, and transformer pipelines (SHM, HATT, MODNet, P3M-Net, ViTAE-S (Chen et al., 2018; Qiao et al., 2020; Ke et al., 2022; Li et al., 2021a; Ma et al., 2023)). Referring segmentation covers vision-language fusion (CRIS, PolyFormer (Wang et al., 2022; Liu et al., 2023)), LLM-based segmentation-token pipelines (LISA, GLaMM, OMG-LLaVA, READ, CoPRS (Lai et al., 2024; Rasheed et al., 2024; Zhang et al., 2024a; Qian et al., 2025; Lu et al., 2025)), and text-generation or consistency-learning formulations (PSALM, Text4Seg, FCLM (Zhang et al., 2024b; Lan et al., 2025; Yang et al., 2026)). Pose baselines include HRNet, TokenPose, ViTPose, RTMO, and DynPose (Sun et al., 2019; Li et al., 2021c; Xu et al., 2022; Lu et al., 2024; Xu et al., 2025b); saliency baselines include ICON, MENet, VST++, VSCode, and EVPv2 (Zhuge et al., 2022; Wang et al., 2023; Liu et al., 2024; Luo et al., 2024; Liu et al., 2025).

4.2 Main Results

Table 1: Main results on geometric dense prediction. Depth entries report absRel ↓ / δ_1 ↑. Normal entries report mean angular error ↓ / 11.25° accuracy ↑.

Method	Monocular depth			Surface normals		
	NYU	KITTI	ScanNet	NYU	ScanNet	iBims
Specialized / discriminative methods						
Omnidata v2 (ICCV'21)	–	–	–	17.2 / 55.5	16.2 / 60.2	18.2 / 63.9
DAv2 (NIPS'24)	0.045 / **0.979**	0.074 / 0.946	–	–	–	–
DSINE (CVPR'24)	–	–	–	16.4 / 59.6	16.2 / 61.0	17.1 / 67.4
Generative dense prediction methods						
GeoWizard (ECCV'24)	0.052 / 0.966	0.097 / 0.921	0.061 / 0.953	17.0 / 56.5	15.4 / 61.6	19.3 / 63.0
Marigold (CVPR'24)	0.055 / 0.964	0.099 / 0.916	0.064 / 0.952	16.1 / 60.5	14.5 / 66.1	16.3 / 68.5
GenPercept (ICLR'25)	0.056 / 0.960	0.130 / 0.842	0.062 / 0.961	18.3 / 56.0	17.7 / 58.3	18.2 / 64.0
Lotus-D (ICLR'25)	0.053 / 0.977	0.093 / 0.928	0.060 / 0.963	16.2 / 59.8	14.7 / 64.0	17.1 / 66.4
Edit2Perc (CVPR'26)	**0.044** / 0.976	0.079 / 0.945	0.049 / 0.973	15.7 / **61.6**	14.1 / 66.3	15.1 / 70.9
RGB-native field readout						
ReChannel-4B (Ours)	0.056 / 0.964	0.067 / 0.952	0.058 / 0.966	15.8 / 60.8	14.3 / 64.9	15.5 / 70.0
ReChannel-9B (Ours)	0.051 / 0.974	**0.063 / 0.959**	**0.047 / 0.976**	**15.6** / 61.2	**13.9 / 66.6**	**15.1 / 71.1**

Geometry: depth and normals (Table 1). On monocular depth, ReChannel-9B achieves the best absRel on KITTI (0.063) and ScanNet (0.047), while reaching 0.051 on NYU, behind Edit2Perc (0.044) and DAv2 (0.045). On surface normals, ReChannel-4B is already competitive with the strongest generative baselines on every split, and ReChannel-9B achieves the best mean angular error on all three datasets (NYU 15.6°, ScanNet 13.9°, iBims 15.1°), even though the same token-local linear readout is used with K_t=3.

Trimap-free matting (Table 2). Matting is the most boundary-sensitive task in our study and the one where a target-side VAE round-trip is expected to hurt most. ReChannel-9B sets a new state of the art on both P3M-500-P (SAD 5.69) and P3M-500-NP (SAD 6.67), improving over the strongest specialized matting backbone ViTAE-S by 0.55 and 0.92 SAD respectively, and over the strongest generative-prior baseline GenPercept by 4.06 and 6.10. The gap widens out of distribution:

Table 2: Trimap-free alpha matting on P3M-500. We report SAD ↓, MSE$\times 10^3$ ↓, and Conn ↓. Blank entries indicate that the corresponding metric is not reported in the source paper.

Method	P3M-500-P			P3M-500-NP			AIM-500 (Zero-shot)		
	SAD	MSE	Conn	SAD	MSE	Conn	SAD	MSE	Conn
Trimap-free matting methods									
SHM (ACMMM'18)	21.56	10.0	17.53	20.77	9.3	17.09	-	-	-
HATT (CVPR'20)	25.99	5.4	25.29	30.53	7.2	27.42	-	-	-
MODNet (AAAI'21)	13.31	3.8	10.88	16.70	5.1	13.81	-	-	-
P3M-Net (ACM MM'21)	8.73	2.6	13.88	11.23	3.5	12.51	103.03	54.47	87.29
ViTAE-S (IJCV'23)	6.24	1.5	5.86	7.59	1.9	6.96	112.52	60.20	43.18
Generative dense prediction methods									
GenPercept (ICLR'25)	9.75	2.5	8.84	12.77	2.7	10.46	75.5	24.20	36.74
RGB-native field readout									
ReChannel-4B (Ours)	5.99	1.3	5.73	6.98	1.5	6.55	43.83	19.20	42.19
ReChannel-9B (Ours)	**5.69**	**1.2**	**5.42**	**6.67**	**1.4**	**6.23**	**34.90**	**14.18**	**32.01**

Table 3: Referring segmentation on RefCOCO-family benchmarks (cIoU, %) over the standard 8 splits and their average.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	val	testA	testB	val	testA	testB	val	test	
CRIS (CVPR'22)	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	64.3
PolyFormer (CVPR'23)	74.8	76.6	71.1	67.6	72.9	59.3	67.8	69.1	69.9
LISA-7B (CVPR'24)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	69.9
GLaMM-7B (CVPR'24)	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.6
PSALM-1.3B (ECCV'24)	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	77.1
OMG-LLaVA-7B (NIPS'24)	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9	72.9
Text4Seg-8B (ICLR'25)	79.2	81.7	75.6	72.8	77.9	66.5	74.0	75.3	75.4
READ-7B (CVPR'25)	78.1	80.2	73.2	68.4	73.7	60.4	70.1	71.4	71.9
CoPRS-7B (ICLR'26)	81.6	85.3	79.5	75.9	80.3	69.7	76.2	76.2	78.1
FCLM-7B (CVPR'26)	82.6	83.9	79.9	77.4	80.5	71.3	79.4	80.5	79.4
ReChannel-4B (Ours)	83.8	85.6	81.6	77.9	81.6	72.5	79.2	79.8	80.3
ReChannel-9B (Ours)	**84.9**	**86.5**	**83.7**	**79.5**	**83.0**	**75.0**	**81.2**	**81.9**	**82.0**

on zero-shot AIM-500, ReChannel-9B reaches 34.90 SAD, less than half of GenPercept (75.5), suggesting that removing the target-side VAE helps not only in-domain boundaries but also cross-domain transfer.

Referring segmentation (Table 3). Referring segmentation is natural ground for a T2I backbone: language conditioning is part of its pretraining interface, whereas self-supervised dense encoders such as the DINO family carry no text pathway and enter this task only through assembled pipelines. The same readout, with the referring expression supplied as the backbone's text condition, handles language-conditioned segmentation without any added language branch, mask-proposal head, or task-specific decoder. ReChannel-4B already outperforms 7B–8B LLM-based baselines (LISA, GLaMM, Text4Seg) and the strongest recent method FCLM-7B (80.3 vs. 79.4 average cIoU). ReChannel-9B reaches 82.0 average cIoU and is the best on all 8 splits, indicating that a T2I-pretrained backbone with a token-local readout already supplies the grounding and precision that prior work obtains through dedicated mask heads or LLM decoders.

Pose and saliency (Table 4). Multi-channel keypoint heatmaps test the readout in its most structured regime. ReChannel-9B reaches 79.2 AP, surpassing ViTPose-L without any pose-specific architecture. At the opposite extreme, saliency tests high-contrast binary fields: ReChannel-9B achieves the best reported $F_{\max}$ and MAE on both DUTS-TE and ECSSD.

Table 4: **Additional dense prediction tasks.** (a) COCO human pose estimation under the detection-box protocol (AP / AP_{50}, %). (b) Saliency detection on DUTS-TE and ECSSD ($F_{\max} \uparrow$ / MAE $\downarrow$).

(a) Human pose

Method	AP	AP_{50}
HRNet-W48 (CVPR'19)	76.3	90.8
TokenPose-L (ICCV'21)	75.8	90.3
ViTPose-L (NIPS'22)	78.3	91.4
RTMO-l (CVPR'24)	74.8	–
DynPose (CVPR'25)	75.0	90.6
ReChannel-4B	78.0	91.4
ReChannel-9B	**79.2**	**92.7**

(b) Saliency

Method	DUTS-TE	ECSSD
ICON (TPAMI'22)	0.904 / 0.037	0.954 / 0.032
MENet (CVPR'23)	0.917 / 0.028	0.957 / 0.031
VST++ (TPAMI'24)	0.887 / 0.033	0.949 / 0.029
VSCode-B (CVPR'24)	0.930 / 0.022	0.961 / 0.022
EVPv2 (TPAMI'26)	0.921 / 0.026	0.967 / 0.022
ReChannel-4B	0.939 / 0.018	0.965 / 0.018
ReChannel-9B	**0.944 / 0.018**	**0.968 / 0.017**

Table 5: Diagnostic ablation under a matched 4B backbone (FLUX-Klein 4B, 512^2, σ=0). Each italic block targets one diagnostic question while holding the backbone family, data, resolution, and evaluation protocol fixed. Normal: mean angular error ($^\circ \downarrow$); Matting: SAD ($\downarrow$) on P3M-500-P / -NP. Latency is single-stream L40S GPU time.

Configuration	Surface normal ($^\circ \downarrow$)			Matting (SAD$\downarrow$)		Lat. (ms)$\downarrow$
	NYU	ScanNet	iBims	P3M-P	P3M-NP	
Is anything beyond the frozen body needed?						
Head-only (body frozen)	44.30	43.55	48.16	180.97	170.92	47.7
Prior or architecture?						
Rand-init Full-FT	22.90	26.43	21.73	11.56	12.32	47.7
Given the prior, does extra capacity or a generative interface help?						
4-stage conv head (425K)	15.89	14.64	16.45	6.90	8.30	48.0
Full fine-tuning (4B)	15.87	**14.06**	15.84	6.29	7.43	47.7
Latent target	17.78	16.45	17.45	10.12	11.58	74.4
VAE-frozen (pixel-sup.)	16.30	14.73	16.13	8.38	10.29	74.4
Edit paradigm	15.94	14.37	15.83	6.68	8.00	118.1
ReChannel-4B (Ours)	**15.82**	14.29	**15.54**	**5.99**	**6.98**	**47.7**

4.3 Diagnostic Ablations

Table 5 examines three diagnostic questions under a matched FLUX-Klein 4B backbone at 512^2.

LoRA adaptation is necessary. With the body fully frozen and only the linear head trained, normals collapse to 44.30° / 43.55° / 48.16° on NYU / ScanNet / iBims, and matting plateaus at SAD 180.97 / 170.92 on P3M-P / NP, close to the trivial constant-prediction regime. The pretrained RGB field is rich, but its channels are not aligned with task-native quantities; lightweight LoRA adaptation is needed before a linear readout becomes meaningful.

A strong pretrained prior is necessary. Training the same architecture from random initialization leaves a large gap on both tasks (22.90° on NYU normals, $+7.08^\circ$ over ours; SAD 11.56 / 12.32 on matting, $\sim 2\times$ worse). The readout thus relies on a strong pretrained prior, not on backbone scale or token-grid structure alone. We do not claim generative pretraining is uniquely required; our contribution concerns the output interface, not the pretraining source.

The adapted token field already carries the output; the interface only reads it. Once the DiT organizes the RGB observation into a task-aligned token field, the target's spatial structure is already in place, so the interface only has to read it out. Extra output-side machinery thus does not help: a $13\times$ larger convolutional head and full fine-tuning of the 4B body both fail to beat our token-local linear readout (full fine-tuning brings no consistent gain, at far higher cost). Reintroducing a generative output interface on top of the same LoRA-adapted token field is consistently worse (Table 5), and the three variants differ only in the output path: *Latent target* regresses the target in the VAE latent space (latent-space MSE) and decodes it through the frozen VAE; *VAE-frozen (pixel-sup.)* decodes through the same frozen VAE decoder but supervises directly in pixel space; and *Edit paradigm* produces the target under the backbone's image-editing formulation (noise$\oplus$RGB

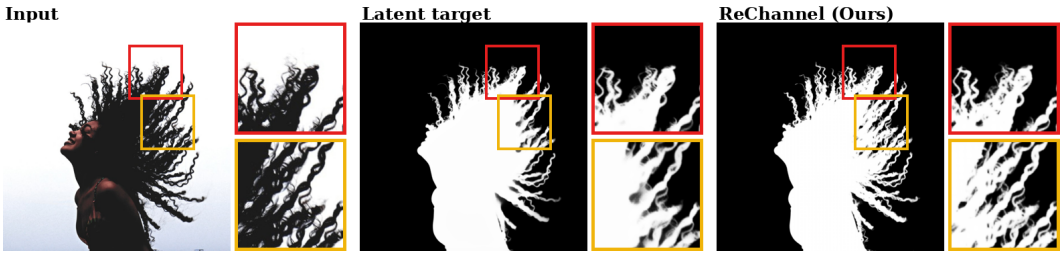

Figure 3: Zero-shot matting on AIM-500 under a matched 4B backbone. The two configurations share the same LoRA-adapted token field and differ only in the output interface: the VAE round-trip (middle) blurs and merges fine hair structure that the token-local readout (right) preserves.

as a $2N$-token input). These rows localize where the cost comes from: moving supervision from the VAE latent space to pixel space (*Latent target*→*VAE-frozen*) recovers part of the gap (SAD $10.12 \rightarrow 8.38$), and dropping the VAE round-trip entirely (*VAE-frozen*→Ours) recovers the rest ($8.38 \rightarrow 5.99$), with the same ordering on all normal splits. Re-rendering the target through a decode path the token field has made unnecessary only adds error and latency; the best overall configuration in our study is thus also the fastest.

Qualitative evidence. Figure 3 further isolates the output-interface effect on zero-shot AIM-500. Given the same LoRA-adapted token field, the latent-target interface merges fine hair strands and removes narrow transparent gaps, whereas the token-local readout preserves them, consistent with the ordering in Table 5.

4.4 Efficiency

On a single L40S GPU at 512^2, ReChannel runs at 47.7 ms per image, identical to a forward pass of the adapted T2I backbone: the token-local linear head adds no measurable cost. The matched-backbone alternatives all pay for a target-side decode. The latent-target and VAE-frozen variants require a VAE decoding pass and run at 74.4 ms ($1.56\times$ slower), while the edit paradigm runs at 118.1 ms ($2.48\times$ slower) under the same backbone, resolution, and LoRA rank. Removing the target-side VAE decode accounts for the $1.56\times$ gap over the latent and VAE-frozen variants; the remaining gap up to $2.48\times$ over the edit paradigm additionally reflects its extra generative passes. These differences arise entirely from the output interface rather than from backbone capacity or input resolution. Unlike these baselines, ReChannel never runs the decoder at all: the target is read out in the same forward pass that produces the token field, so accuracy and speed come from the same design choice rather than trading off against each other. The best overall configuration in our study is thus also the fastest.

5 Conclusion

We revisited the output interface between text-to-image generative priors and dense prediction. ReChannel preserves the pretrained RGB input pathway but removes the target-side generative path: an adapted DiT token field serves as a spatial carrier, and a token-local linear head directly reinterprets its channels as task-native quantities. Across six tasks and more than a dozen benchmarks, this minimal interface achieves state-of-the-art results on trimap-free matting, KITTI depth, and referring segmentation, while remaining competitive on normals, saliency, and pose with only the output channels K_t varying across tasks. Under a matched backbone, larger heads, full fine-tuning, latent-space supervision, and generative decoding provide no consistent overall gain over direct readout, while generative interfaces incur up to $2.48\times$ the latency. These results suggest that generative pretraining and generative output interfaces need not be coupled: once the image is organized into a task-alignable token field, dense prediction can operate directly in its native output space. ReChannel thus preserves the benefits of generative pretraining while avoiding unnecessary output-side reconstruction, offering a simple interface for accurate and efficient pixel-space perception across diverse dense tasks.

ETHICS STATEMENT.

This work does not involve human participants, personally identifiable data, or other research activities that require specific ethical approval.

REPRODUCIBILITY STATEMENT.

The main paper describes the proposed method and evaluation protocols in detail, while implementation details and hyperparameter settings are provided in the appendix. We intend to release the code and configurations upon publication.

AI USE STATEMENT.

Generative AI tools assisted with language refinement and manuscript organization. The authors reviewed all AI-assisted material and take full responsibility for the final manuscript.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022.

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.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Black Forest Labs. FLUX.2 [klein]. `https://huggingface.co/black-forest-labs/FLUX.2-klein-9B`, 2026.

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.

Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 618–626, 2018.

Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023.

Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

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.

Xiaodan Du, Nicholas Kolkin, Greg Shakhnarovich, and Anand Bhattad. Generative models: What do they know? do they know things? let's find out! *arXiv preprint arXiv:2311.17137*, 2023.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pp. 241–258. Springer, 2024.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T Barron, et al. Image generators are generalist vision learners. *arXiv preprint arXiv:2604.20329*, 2026.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4340–4349, 2016.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, volume 2025, pp. 89454–89467, 2025.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9492–9502, 2024.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1140–1147, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In *15th European Conference on Computer Vision, ECCV 2018*, pp. 331–348, 2018.

Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogerio Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13883–13893, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wei Zhang. Text4seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations*, volume 2025, pp. 1634–1661, 2025.

Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3501–3509, 2021a.

Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 800–806, 2021b.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 11313–11322, 2021c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Xin Lin, Haodong Li, Zhifei Zhang, Yutong Yang, Haitian Zheng, Juanxi Tian, Zhe Lin, and Truong Nguyen. Pixelcontrol: Fine-grained condition fidelity in text-to-image diffusion. *arXiv preprint arXiv:2608.15705*, 2026.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18653–18663, 2023.

Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, 2024.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for universal foreground segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1491–1500, 2024.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. Coprs: Learning positional prior from chain-of-thought for reasoning segmentation. *arXiv preprint arXiv:2510.11173*, 2025.

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

Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17169–17180, 2024.

Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. Rethinking portrait matting with privacy preserving. *International journal of computer vision*, 131(8):2172–2197, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

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.

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

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how¡ seg¿ token works. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24722–24731, 2025.

Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13676–13685, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yiqing Shi, Yiren Song, and Mike Zheng Shou. Edit2perceive: Image editing diffusion models are strong dense perceivers. *arXiv preprint arXiv:2511.18673*, 2025.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Fengyi Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3554–3563, 2024.

JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, and Xiangxiang Chu. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025a.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 136–145, 2017.

Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, and Xiangjian He. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10031–10040, 2023.

Zanyi Wang, Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, and Jingdong Wang. Deforming videos to masks: Flow matching for referring video segmentation. *arXiv preprint arXiv:2510.06139*, 2025b.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Guangkai Xu, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, Chunhua Shen, et al. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations*, volume 2025, pp. 6786–6799, 2025a.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955–2966, 2023.

Yalong Xu, Lin Zhao, Chen Gong, Guangyu Li, Di Wang, and Nannan Wang. Dynpose: Largely improving the efficiency of human pose estimation by a simple dynamic framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1160–1169, 2025b.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1155–1162, 2013.

Jing Yang, Sen Yang, Boqiang Duan, Ming Dai, Wei Zhang, Xiao Tan, Kunbin Chen, Wei He, Jingdong Wang, and Hanli Wang. Hugging visual prompt and segmentation tokens: Consistency learning for fine-grained visual understanding in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5175–5186, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Denis Zavadski, Damjan Kalšan, and Carsten Rother. Primedepth: Efficient monocular depth estimation with a stable diffusion preimage. In *Proceedings of the Asian Conference on Computer Vision*, pp. 922–940, 2024.

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

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024a.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5729–5739, 2023.

Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3738–3752, 2022.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023.

A Implementation Details

All tasks share the same architecture family: a frozen FLUX-Klein backbone (4B and 9B) (Black Forest Labs, 2026), a per-task LoRA adapter (Hu et al., 2022), and a token-local linear readout head ($128 \to p^2 K_t$) with a fixed reshape and no inter-token mixing. The same readout form is used across tasks, while the output channel count K_t is selected according to the task-native target. For scalar outputs with p=16 and K_t=1, the head contains approximately 33K parameters; its size scales linearly with K_t. The VAE encoder is retained as the input interface, while dense targets are supervised directly in pixel space and the target-side VAE decoder is removed.

Table 6: Shared training and inference configuration (all tasks).

Setting	Value
Backbone (frozen)	FLUX-Klein 4B / 9B
LoRA rank r / α	64 / 32
Readout patch size p	16
Optimizer	AdamW (β_1=0.9, β_2=0.95, ϵ=10^{-8})
LoRA learning rate	1×10^{-4}
Weight decay	0
Gradient clipping	1.0
Batch size	64
Training steps	30,000
EMA decay	0.999
Training resolution	512^2
Inference	deterministic σ=0, guidance 0

B Task and Dataset Summary

Table 7 summarizes the training data, evaluation benchmarks, and task-native outputs used in the main experiments. No additional task-specific training data are used.

Datasets. The geometry experiments use Hypersim (Roberts et al., 2021), Virtual KITTI (Gaidon et al., 2016), NYU (Silberman et al., 2012), KITTI (Geiger et al., 2012), ScanNet (Dai et al., 2017), and iBims-1 (Koch et al., 2018). The matting experiments use P3M-10k and P3M-500-P/NP (Li et al., 2021a), together with AIM-500 (Li et al., 2021b). Referring segmentation uses RefCOCO/+/g (Yu et al., 2016; Mao et al., 2016). Saliency uses DUTS-TR/TE (Wang et al., 2017) and ECSSD (Yan et al., 2013), while pose estimation uses COCO train2017 and val (Lin et al., 2014).

Table 7: Summary of the tasks and datasets used in the main experiments. The same token-local readout is used throughout; only the output channels K_t and the supervision loss differ across tasks.

Task	Training data	Evaluation benchmarks	Output	K_t	Loss
Depth	Hypersim + Virtual KITTI	NYU, KITTI, ScanNet	Depth	1	SILog
Surface normals	Hypersim	NYU, ScanNet, iBims-1	Normals	3	Cosine + ℓ_1
Matting	P3M-10k	P3M-500-P/NP, AIM-500	Alpha	1	ℓ_1
Ref. segmentation	RefCOCO/+/g	8 standard splits	Mask	1	BCE
Saliency	DUTS-TR	DUTS-TE, ECSSD	Saliency	1	BCE
Pose	COCO train2017	COCO val	Heatmaps	17	MSE

C Additional Qualitative Results

The following figures provide additional predictions on representative benchmarks. All outputs are produced using the model configuration described in the main paper.

C.1 Monocular Depth

Figures 4–6 show additional depth predictions on NYU, ScanNet, and KITTI. The readout yields sharp depth discontinuities at object boundaries while keeping large planar regions smooth, and the same behavior transfers to outdoor KITTI street scenes.

C.2 Surface Normals

Figures 7–9 show additional normal predictions on NYU, ScanNet, and iBims. The predictions recover the piecewise-planar structure of indoor scenes, with crisp edges between orthogonal planes and smooth transitions on curved objects.

C.3 Trimap-Free Matting

Figures 10 and 11 show additional alpha predictions on P3M-500-NP and P3M-500-P. The predictions preserve individual hair strands, fine boundary details, and semi-transparent transition regions without trimap guidance.

C.4 Referring Segmentation

Figure 12 shows additional referring-segmentation results with the expression as the text condition. The masks respect spatial and attribute-based references (“in right”, “red bike”) and adhere closely to object contours.

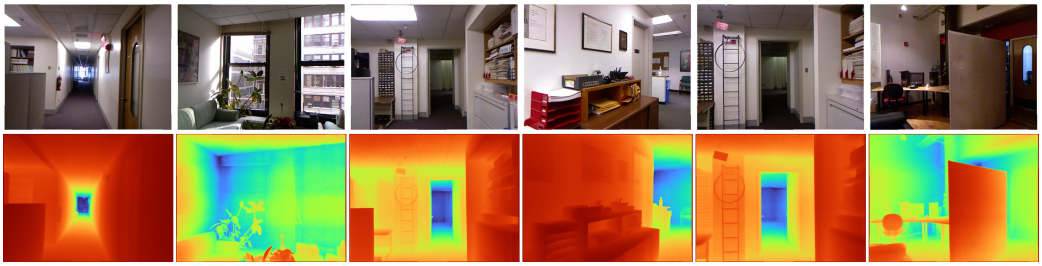

Figure 4: **Monocular depth on NYU.** RGB inputs and the corresponding depth predictions.

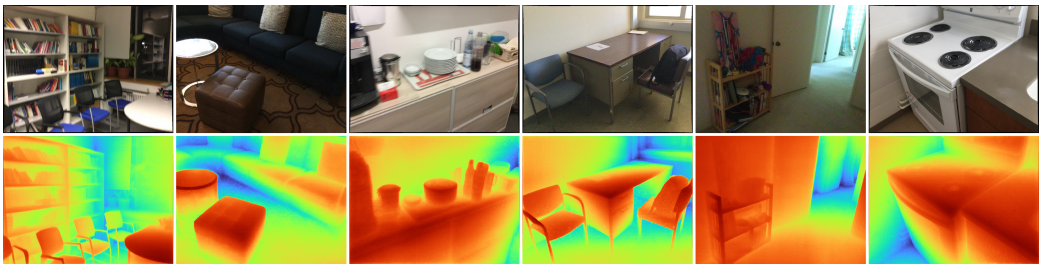

Figure 5: **Monocular depth on ScanNet.** RGB inputs and the corresponding depth predictions.

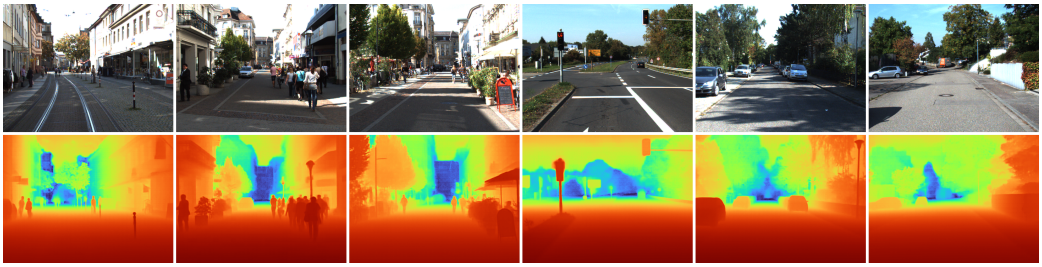

Figure 6: **Monocular depth on KITTI.** RGB inputs and the corresponding depth predictions.

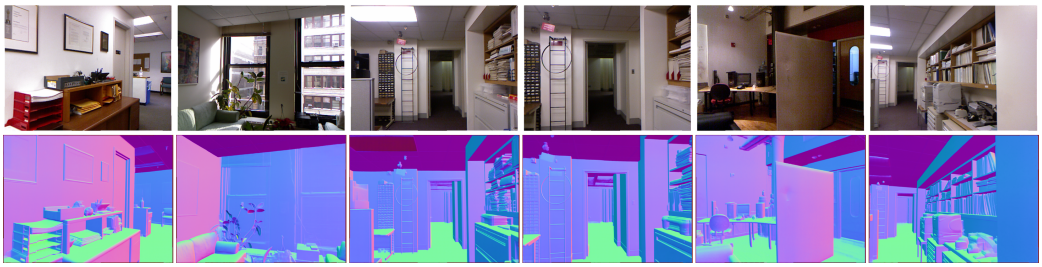

Figure 7: **Surface-normal prediction on NYU.** RGB inputs and normal predictions.



Figure 8: **Surface-normal prediction on ScanNet.** RGB inputs and normal predictions.

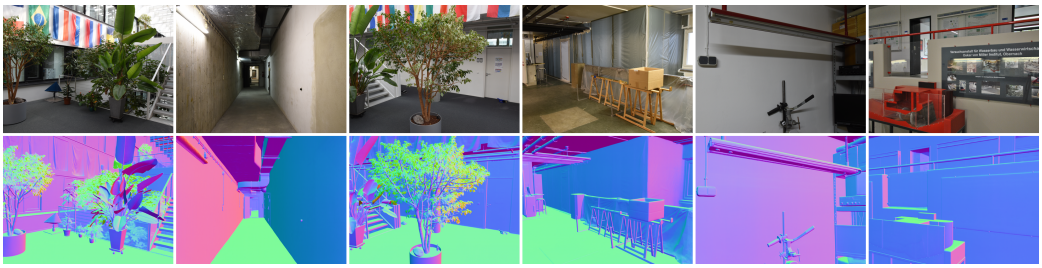

Figure 9: **Surface-normal prediction on iBims.** RGB inputs and normal predictions.



Figure 10: **Trimap-free matting on P3M-500-NP.** RGB inputs and the alpha predictions.



Figure 11: **Trimap-free matting on P3M-500-P.** RGB inputs and the alpha predictions.



Figure 12: **Additional referring-segmentation results.** The referring expression is supplied as the text condition, and the corresponding mask is predicted from the spatial token field.