

MTV: Revisiting Multi-Task Visual Representation Learning

Shangzhe Di^{1,2}, Zhonghua Zhai², and Weidi Xie¹

¹ SAI, Shanghai Jiao Tong University

² ByteDance Seed

<https://github.com/Becomebright/MTV>

Abstract. Current visual representation learning remains bifurcated: vision-language models (e.g., CLIP) excel at global semantic alignment but lack spatial precision, while self-supervised methods (e.g., MAE, DINO) capture intricate local structures yet struggle with high-level semantic context. We argue that these paradigms are fundamentally complementary and can be integrated into a principled multi-task framework, further enhanced by dense spatial supervision. We introduce **MTV**, a multi-task visual pretraining framework that jointly optimizes a shared backbone across vision-language contrastive, self-supervised, and dense spatial objectives. To mitigate the need for manual annotations, we leverage high-capacity “expert” models—such as Depth Anything V2 and OWLv2—to synthesize dense, structured pseudo-labels at scale. Beyond the framework, we provide a systematic investigation into the mechanics of multi-task visual learning, analyzing: (i) the individual and (ii) marginal gain of each objective, (iii) task synergies versus interference, and (iv) scaling behavior across varying data and model scales. Our results demonstrate that MTV achieves “best-of-both-worlds” performance, significantly enhancing fine-grained spatial reasoning without compromising global semantic understanding. Our findings suggest that multi-task learning, fueled by high-quality pseudo-supervision, is a scalable path toward more general visual encoders.

Keywords: Visual representation learning · Multi-task training

1 Introduction

Recent advances in foundation models have reshaped visual representation learning by scaling pretraining on web-scale data. Currently, two paradigms prevail: *Vision-language contrastive learning* [19, 36, 53] aligns global image and text embeddings via instance-level discrimination, enabling open-vocabulary recognition and zero-shot transfer. *Self-supervised learning*, such as masked image modeling [15, 55], contrastive instance discrimination [7, 16], and student-teacher matching [5, 34], enforces invariance and regularities without manual labels. Despite their success, these paradigms remain bifurcated, primarily focusing on either global semantics or local regularities while leaving fine-grained spatial and geometric reasoning as under-constrained emergent properties. Given the

inherent diversity of visual tasks—ranging from low-level geometry to high-level reasoning—learning truly universal visual representations that bridge these multifaceted requirements remains an open and fundamental challenge [10, 27, 41, 42].

In contrast, large language models (LLMs) scale effectively by absorbing multifaceted supervision—such as summarization, translation, and question answering—within a unified training interface [1, 4, 21]. Admittedly, unlike natural language, visual tasks are inherently heterogeneous. While this diversity makes a single, universal objective more elusive in vision, we argue that visual pretraining can still embrace the LLM philosophy by simultaneously optimizing across a diverse suite of tasks to capture the full spectrum of information.

To bridge this gap, we introduce **MTV**, a multi-task visual pretraining framework that jointly optimizes a shared backbone across vision-language contrastive, self-supervised, and dense spatial objectives, as shown in Fig. 1. To overcome the scarcity of human-labeled data, we leverage high-capacity “expert” models—such as Depth Anything V2 [49] for monocular depth estimation and OWLv2 [30] for open-vocabulary grounding—to synthesize dense, structured pseudo-labels at scale. By integrating these expert-guided targets, we enrich traditional objectives with explicit geometric and spatial priors.

Besides, we go beyond simple multi-tasking to provide a systematic investigation into the mechanics of multi-task visual learning. We create a controlled setting to study how heterogeneous signals cooperate. Specifically, we examine: (i) the individual and (ii) marginal gain of each objective, (iii) the dynamics of task synergies versus interference, and (iv) scaling behavior across varying data and model scales. Empirically, MTV achieves “best-of-both-worlds” performance, yielding representations that are more general and data-efficient than any single objective alone. For instance, our **MTV-Base** model trained on 100M samples achieves 69.5% ImageNet zero-shot accuracy—outperforming **CLIP-Base** [36] trained on 400M samples—while delivering substantial gains across retrieval, depth estimation, segmentation, and correspondence benchmarks.

Taken together, these findings suggest a clear design principle: expanding conventional visual pretraining with complementary tasks—even when supervised by imperfect pseudo labels—shifts learning toward a unified, multi-granularity representation space. This unification, analogous in spirit to the convergence of

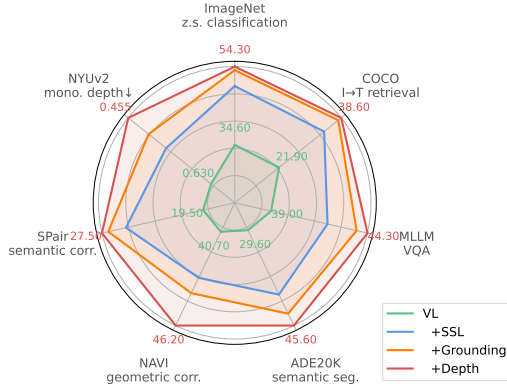


Fig. 1: Multi-task supervision effects on a ViT-L model with 10M training samples. We integrate VL, SSL, and pseudo-labeled grounding and depth estimation into a unified training framework, which leads to strong and consistent performance gains across diverse tasks.

diverse language tasks for LLM, offers a scalable and data-efficient pathway for building versatile visual encoders capable of anchoring a wide range of perception and multimodal reasoning tasks.

2 Related Work

Self-supervised Learning (SSL). Self-supervised visual pretraining aims to learn transferable representations without manual annotations. Contrastive and distillation-based methods, such as DINO [5], iBOT [55], and I-JEPA [2], encourage feature invariance across augmented views, whereas reconstruction-based approaches, like MAE [15], rely on masked image modeling. Recent large-scale efforts such as DINOv2 [34] demonstrate that scaling data and model size leads to robust features for recognition, segmentation, and correspondence tasks. However, these methods remain limited to unimodal visual signals and lack semantics from natural language or high-level reasoning [41].

Vision-Language Contrastive Learning (VL). Weakly-supervised vision-language pretraining has emerged as a dominant paradigm for learning semantic-aligned visual encoders. CLIP [36] and ALIGN [19] show that large-scale vision-language contrastive learning yields powerful zero-shot transfer. Subsequent works such as Florence [51] and SigLIP [53] further improve alignment quality and scalability through better data curation and sigmoid-based objectives. Despite their impressive semantic generalization, these models are optimized purely for global alignment and often underperform on spatially structured or geometric tasks such as depth and correspondence estimation [10].

Self-Training (ST). Self-training uses a supervised model to generate pseudo labels on unlabeled data, followed by training a student model on these pseudo-labeled samples [47], expanding supervision beyond human-annotated datasets. For VL contrastive learning, several approaches [11, 23, 26, 50] generate synthetic captions or rewrite noisy web captions to enhance CLIP-style training. However, textual descriptions alone cannot fully capture the rich visual content. Beyond VL pretraining, self-training has also demonstrated remarkable success in other visual domains. For instance, Depth Anything [48, 49] trains a teacher on synthetic depth data to pseudo-label real images, leading to state-of-the-art monocular depth estimation. Similarly, OWLv2 [30] employs a supervised open-vocabulary detector to produce billion-scale pseudo labels for scaling. We build upon this trend by extending self-training to multi-task settings that combine VL and SSL objectives, enabling more general and scalable representation learning.

Multi-task Visual Representation Learning. Multi-task learning optimizes shared representations to capture cross-domain structures and enhance generalization. Early work, such as Cross-stitch [31], MTAN [25], and MTI-Net [44], focuses on architecture design for sharing features across a few supervised tasks. However, such methods target task-specific performance on small labeled datasets, rather than learning general-purpose representations through large-scale pre-

training. Methods such as SLIP [32], SILC [33], and TIPS [28] show that integrating VL and SSL objectives yields complementary supervision, improving both semantic alignment and spatial understanding. In the context of ST, MuST [14] pioneered the idea that combining multiple self-training tasks can yield stronger representations than purely supervised or self-supervised learning. We advance this trajectory by integrating VL, SSL, and ST under a unified framework. Notably, while SigLIP2 [43] recently explored a similar combination of VL, SSL, and grounding-based ST [45], it lacks a systematic analysis of how task composition and scaling impact the learned feature space. Our work diverges by providing rigorous studies across task diversity and data/model scales, and crucially, by introducing depth supervision into the multi-task objective. We demonstrate that this geometric cue not only enhances dense spatial understanding but also unexpectedly bolsters global semantic alignment.

3 Preliminary

Our goal is to learn unified visual representations that support a broad spectrum of visual understanding tasks via large-scale, multi-source supervision. We begin by introducing the key components used throughout the paper, covering both the shared architecture and the heterogeneous annotations that supervise it.

Architecture. At the core of our framework is a shared ViT [9] encoder f_θ that maps an image $v_i \in \mathbb{R}^{H \times W \times 3}$ to multi-layer patch features:

$$\{\mathbf{z}_i^{(1)}, \dots, \mathbf{z}_i^{(L)}\} = f_\theta(v_i), \quad \mathbf{z}_i^{(l)} \in \mathbb{R}^{N \times D},$$

where L is the number of layers, N the number of patches, and D the feature dimension. For brevity, we denote the final layer representation as $\mathbf{z}_i := \mathbf{z}_i^{(L)}$. For tasks requiring a global image-level representation, we apply attentive pooling [53] to the final patch features: $\mathbf{v}_i = \text{Pool}(\mathbf{z}_i) \in \mathbb{R}^D$.

Simultaneously, a text encoder g_ϕ processes natural language inputs. To enable cross-modal alignment, g_ϕ maps text t_i into a joint latent space shared with the pooled visual features, producing the text embedding $\mathbf{t}_i = g_\phi(t_i) \in \mathbb{R}^D$.

Training Data. We consider a pretraining corpus $\mathcal{D} = \{(v_i, a_i)\}$, where each image v_i is paired with automatically acquired annotations a_i that provide semantic, spatial, and geometric signals (Fig. 2a): (i) A **caption** t_i capturing the global semantics of the image, used for vision-language contrastive learning; (ii) **Region-text pairs** $\{(b_{ij}, r_{ij})\}$ where b_{ij} and r_{ij} denote the j -th bounding box and its entity name. These pairs are obtained by extracting salient entities using RAM++ [17], and then prompting the open-vocabulary detector OWLv2-Base [30] to localize the referred regions; (iii) A **relative depth map** $d_i \in \mathbb{R}^{H \times W}$ that encodes per-pixel depth, produced by the Depth-Anything-V2-Large model [49]. In the supplementary material, we also provide results using a smaller teacher model and show that our method is robust to teacher capacity.

These heterogeneous annotations, derived entirely from web data and off-the-shelf specialist models, provide a rich supervisory signal for studying the effectiveness of large-scale multi-task representation learning.

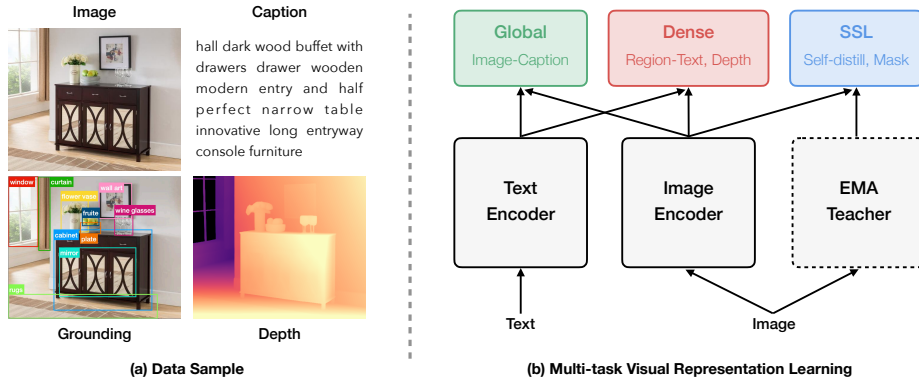


Fig. 2: Overview of our MTV framework. (a) Each image is paired with a web-crawled caption, and augmented with pseudo region-text pairs and relative depth maps generated by teacher models. (b) MTV jointly learns from three complementary supervision types: **Global** (image-caption contrast), **Dense** (region-text alignment, depth), and **SSL** (self-distillation, masked feature prediction). A shared image encoder is optimized together with a text encoder and an EMA teacher.

4 Multi-task Visual Representation Learning

This section details **MTV**, a principled multi-task visual representation learning framework. As shown in Fig. 2b, MTV optimizes a shared encoder across three complementary supervision paradigms: (i) vision-language contrastive learning for global semantic alignment, (ii) self-supervised learning to induce local structural invariance, and (iii) dense supervision to capture fine-grained spatial structures. We elaborate on each supervisory component below.

4.1 Global Semantic Supervision

We adopt vision-language contrastive learning to provide global semantic supervision by aligning image and text embeddings at the instance level. This objective enables the model to learn high-level semantic representations from noisy, large-scale web data.

Given a batch of image-text pairs $\mathcal{B} = \{(v_i, t_i)\}_{i=1}^B$, the visual and textual embeddings $\mathbf{v}_i$ and $\mathbf{t}_i$ are obtained using the visual encoder and text encoder, as defined in Section 3. We adopt the sigmoid-based objective from SigLIP [53], which treats the alignment task as a dense pairwise binary classification problem:

$$\mathcal{L}_{\text{VL}} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B \log \frac{1}{1 + e^{y_{ij}(-\tau \mathbf{v}_i \cdot \mathbf{t}_j + \beta)}}, \quad (1)$$

where $y_{ij} = 1$ if (v_i, t_j) is a matched pair and -1 otherwise. The logit scale τ and bias β are learnable parameters initialized from $\log 10$ and -10 , respectively. Unlike the softmax-based contrastive loss [36], the sigmoid objective decouples the batch size from the normalization, leading to more stable gradients at scale.

To support large-scale distributed training, SigLIP originally employs a cyclic rotation-based protocol to avoid large collective communication. However, in our framework, we observe that the communication overhead of text embeddings is negligible relative to the computational cost of the shared multi-task backbone. Therefore, we utilize a more direct and implementation-efficient alternative: we synchronize textual embeddings across all devices using a differentiable `all_gather`³ operation. This preserves the mathematical correctness of Eq. 1 while facilitating easier integration within a multi-task pipeline.

4.2 Self-Supervised Objectives

While VL supervision captures global semantics, it provides limited guidance for learning locally consistent features. To complement it, we incorporate two self-supervised objectives within a teacher-student distillation framework: (i) local-to-global self-distillation [5] and (ii) masked feature prediction [55].

Following the architecture in DINOv2 [34] and TIPS [28], our student network comprises the visual encoder f_θ and separate MLP projection heads for each SSL task to mitigate gradient interference. The teacher network is maintained as an Exponential Moving Average (EMA) of the student, providing stable targets throughout pretraining.

Local-to-Global Self-distillation encourages spatial consistency by aligning the representations of local and global views. For each image v_i , the student processes $M = 6$ local crops to produce their global embeddings $\{\mathbf{v}_{i,m}\}_{m=1}^M$, which are then projected into “prototype scores” $\{\mathbf{p}_{i,m}\}_{m=1}^M$ by the student MLP. Simultaneously, the teacher processes a larger, global crop to produce the target scores $\hat{\mathbf{p}}_i$. The student is trained to match the teacher’s distribution using a temperature-scaled KL divergence:

$$\mathcal{L}_{\text{distill}} = - \sum_i^B \sum_m^M \text{KL} \left(\text{softmax} \left(\frac{\hat{\mathbf{p}}_i - \mathbf{c}}{\tau_t} \right) \parallel \text{softmax} \left(\frac{\mathbf{p}_{i,m}}{\tau_s} \right) \right),$$

where τ_t and τ_s are the teacher and student temperatures for score sharpening, and $\mathbf{c}$ is an EMA-updated centering variable that prevents collapsing to a trivial solution by encouraging a uniform distribution across prototypes.

Masked Feature Prediction focuses on capturing fine-grained spatial dependencies. We randomly mask 50% of the patches in v_i and replace them with a learnable mask token. The student encodes the masked input into prototype scores $\{\mathbf{p}_i^j\}$, while the teacher processes the original unmasked image to provide target scores $\{\hat{\mathbf{p}}_i^j\}$ for the corresponding masked locations j . The student is trained to match the teacher’s distributions at these masked locations, thereby

³ `all_gather` is a collective communication primitive that synchronizes and collects tensors from all devices. We use a differentiable version of it that supports gradient backpropagation through $\mathbf{t}_j$.

forced to model long-range spatial regularities from incomplete inputs:

$$\mathcal{L}_{\text{mask}} = - \sum_i^B \sum_{j \in \text{masked}} \text{KL} \left(\text{softmax} \left(\frac{\hat{\mathbf{p}}_i^j - \mathbf{c}}{\tau_t} \right) \parallel \text{softmax} \left(\frac{\mathbf{p}_i^j}{\tau_s} \right) \right).$$

Following [34], we further include the KoLeo regularizer [39] to promote a more uniform feature distribution within each batch. The total SSL objective is:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{distill}} + \lambda_{\text{mask}} \times \mathcal{L}_{\text{mask}} + \lambda_{\text{koleo}} \times \mathcal{L}_{\text{koleo}}, \quad (2)$$

where $\lambda_{\text{mask}}=2$ and $\lambda_{\text{koleo}}=0.1$ as in [34]. By integrating $\mathcal{L}_{\text{SSL}}$ with $\mathcal{L}_{\text{VL}}$, the model learns representations that are both semantically rich and spatially grounded.

4.3 Dense Structured Supervision

While VL and SSL provide semantic and structural regularities, they often lack the explicit spatial precision required for dense downstream tasks. To bridge this gap, we incorporate dense structured supervision by distilling knowledge from high-capacity expert models into our shared encoder. Concretely, we introduce two complementary objectives: (i) region-level grounding for object-level semantics and (ii) per-pixel depth regression for fine-grained geometry.

Region-level Grounding. To enforce spatial grounding, we incorporate regional-textual correspondences. For each image v_i , we sample up to four region-text pseudo-labels to balance supervision density and computational cost. The visual and textual inputs are encoded into patch-level features $\mathbf{z}_i$ and region text embeddings $\mathbf{t}_{i,j}$, respectively. Following CLOC [6], we utilize the Soft Region Aggregation mechanism to handle potential misalignments in pseudo-labels. Specifically, a lightweight Transformer layer processes the patch-level features $\mathbf{z}_i$, conditioned on the positional encodings derived from each bounding box. This learnable attention allows the model to adaptively integrate context beyond rigid box boundaries, mitigating noise in pseudo labels. The resulting soft region embeddings $\mathbf{z}_{i,j}$ are aligned with their corresponding text embeddings $\mathbf{t}_{i,j}$ using the sigmoid contrastive loss (Eq. (1)), yielding the grounding loss $\mathcal{L}_{\text{ground}}$.

Pixel-level Geometric Supervision. To capture dense geometric priors, we incorporate monocular depth estimation supervised by high-fidelity pseudo labels. We attach a lightweight DPT head [37] to the shared visual encoder, which aggregates multi-layer features $\{\mathbf{z}_i^{(l)}\}_{l \in \mathcal{S}}$ from selected ViT layers to predict a dense depth map $\hat{d}_i \in \mathbb{R}^{H \times W}$. The prediction is supervised using a combination of a scale- and shift-invariant loss alongside a multi-scale gradient loss from MiDaS [3], jointly denoted as $\mathcal{L}_{\text{depth}}$. To mitigate the influence of potential artifacts in pseudo labels, we adopt a top- K denoising strategy, discarding the highest 10% of per-pixel losses within each sample during training. This encourages the encoder to prioritize high-confidence geometric structures, providing a structural anchor that complements the global semantic objectives. Detailed loss formulations are provided in the supplementary material A.2.

By unifying region-level semantics and pixel-level geometry, the total dense supervision loss is defined as:

$$\mathcal{L}_{\text{Dense}} = \mathcal{L}_{\text{ground}} + \mathcal{L}_{\text{depth}}. \quad (3)$$

4.4 Joint Optimization

We optimize the shared visual encoder by jointly minimizing a multi-task loss that integrates the VL, SSL, and dense signals introduced in Eqs. (1)–(3):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VL}} + \mathcal{L}_{\text{SSL}} + \mathcal{L}_{\text{Dense}}.$$

We provide a detailed gradient analysis in the supplementary material B.3. Empirically, we observe that uniform loss weighting yields stable convergence and competitive performance. As the constituent objectives naturally align in both magnitude (1–10) and gradient scale, the optimization process is not dominated by any individual task. Consequently, our framework remains robust without requiring exhaustive hyperparameter tuning. Furthermore, since every training image is fully equipped with all task annotations, we circumvent the need for complex task-specific sampling or data balancing. This unified training interface enables the shared encoder to efficiently internalize a comprehensive spectrum of visual information within a single forward-backward pass.

5 Experiments

Table 1: Training Cost for a ViT-B/16 model under different training task combinations. (Numbers in parentheses) indicate the increase over the previous row. *The EMA teacher used in SSL is frozen during training.

Training Tasks	Trainable Parameters	Training Time
VL	203.2 M	1.0×
VL + SSL*	232.6 M (+29.4 M)	1.2× (+20%)
VL + SSL* + Ground.	239.7 M (+7.1 M)	1.5× (+25%)
VL + SSL* + Ground. + Depth	250.3 M (+10.6 M)	1.7× (+13%)

5.1 Training Details

Architecture. We evaluate our MTV framework using ViT-B/16-224 and ViT-L/16-256 backbones, following the SigLIP [53] architecture and its text tokenizer. To accommodate the multi-task nature of MTV, we attach lightweight heads for SSL (MLP heads), grounding (Transformer aggregator), and depth estimation (DPT head). Table 1 summarizes the incremental cost of each task, while architectural specifications are detailed in the supplementary material.

Training Data Curation. All experiments are conducted on the DataComp dataset [12]. We crawled 200 million raw image–text pairs and applied a filtering pipeline. We discard images with a shorter side of < 224 pixels and downsample

high-resolution images to a maximum dimension of 1024 pixels while maintaining aspect ratios. This filtering removes low-information outliers and ensures that the “expert” models—OWLv2 and Depth Anything V2—operate within their optimal receptive fields. We then apply the pseudo-labeling procedure described in Section 3, resulting in 100 million “fully-labeled” samples, each containing a caption, region–text pairs, and a relative depth map. More data curation details can be found in the supplementary material A.1.

Optimization Configurations. We pretrain our models from scratch, and investigate the scaling behavior of MTV at three data scales: 10M, 50M, and 100M samples. The 10M and 50M scales are used for ablation and are trained for 20 and 10 epochs, respectively, with a global batch size of 4k. Our full-scale 100M model is trained for 32 epochs with a global batch size of 8k, following the CLIP [36] protocol for fair comparison. We use the Adam optimizer with a learning rate of 5×10^{-4} , and a linear warmup over the first 1k steps. Training uses BF16 mixed precision for efficiency, except for the DPT depth head, which is kept in FP32 to ensure numerical stability when regressing continuous geometric values. Experiments are conducted on 8-64 NVIDIA H200 GPUs. The final MTV-B model takes 4.4k GPU hours, and MTV-L costs 14.4k GPU hours.

5.2 Evaluation Details

We evaluate the learned representations across a broad range of downstream tasks, spanning global semantic understanding and fine-grained spatial reasoning. As summarized in Table 2, each task is assessed on one or more representative benchmarks. Below, we outline the setup for each task.

Zero-shot Image Classification. We report Top-1 accuracy on the ImageNet-1k [38] validation set using the standard prompt-based zero-shot protocol.

Zero-shot Retrieval. For cross-modal alignment, we evaluate image-to-text (I→T) and text-to-image retrieval (T→I) on COCO [8], reporting Recall@1.

Table 2: Summary of evaluation benchmarks.

Task	Metric	Benchmark
<i>Global Semantic Understanding</i>		
Zero-shot Classification	Top-1 Acc.	ImageNet-1k
Zero-shot Retrieval	Recall@1	COCO
Visual Question-Answering	Score	MMVP, CVBench, RealWorldQA
<i>Fine-grained Spatial Reasoning</i>		
Semantic Segmentation	mIoU	ADE20k
Geometric Correspondence	Avg. Recall	NAVI
Semantic Correspondence	Avg. Recall	SPair
Monocular Depth	RMSE	NYUv2

MLLM VQA. We integrate the frozen visual encoder into the TinyLLaVA framework [20] with a Qwen2.5-3B LLM [35]. Following the standard two-stage recipe, we train the model on LLaVA-558K, followed by supervised fine-tuning on LLaVA-NEXT-779K [24]. Evaluation is conducted on three vision-centric VQA benchmarks categorized by [41], including MMVP [42], CVBench [41], and RealWorldQA [46]. We report the average score across these benchmarks.

Table 3: Individual Task Contributions (50M Data). 10M results are omitted due to space constraints, but show similar trends.

ViT	Training Tasks	IN-1k	COCO _I	ADE20k	NAVI	SPair	NYUv2 ↓
B-16	VL	55.7	43.0	32.2	39.8	18.0	0.604
	VL + SSL	61.4 (+5.7)	48.0 (+5.0)	39.9 (+7.7)	40.4 (+0.6)	22.7 (+4.7)	0.556 (-.048)
	VL + Ground.	59.5 (+3.8)	46.8 (+3.8)	37.6 (+5.4)	41.8 (+2.0)	21.2 (+3.2)	0.541 (-.063)
	VL + Depth	57.7 (+2.0)	44.0 (+1.0)	36.0 (+3.8)	42.2 (+2.4)	21.3 (+3.3)	0.511 (-.093)
L/16	VL	60.8	46.3	37.3	39.9	19.8	0.575
	VL + SSL	65.4 (+4.6)	48.0 (+1.7)	45.5 (+8.2)	43.2 (+3.3)	27.3 (+7.5)	0.501 (-.074)
	VL + Ground.	64.1 (+3.3)	50.5 (+4.2)	42.1 (+4.8)	43.3 (+3.4)	25.4 (+5.6)	0.508 (-.067)
	VL + Depth	63.6 (+2.8)	48.8 (+2.5)	41.3 (+4.0)	43.6 (+3.7)	22.7 (+2.9)	0.467 (-.108)

Semantic Segmentation. We evaluate on ADE20k [54] using a frozen visual encoder and a linear classification head to predict per-patch semantic labels. Training runs for 40k steps with batch size 32, learning rate $4e-5$, and input resolution 576×576 . Performance is measured by mIoU.

Geometric and Semantic Correspondence. To investigate the model’s 3D awareness, we utilize the Probe3D framework [10]. We assess *geometric correspondence* on NAVI [18], which requires matching pixels originating from the same 3D point across different viewpoints. We further evaluate *semantic correspondence* on SPair [29], testing the alignment of semantically similar parts across diverse instances. Both tasks are measured using Average Recall across various viewpoints or distances.

Monocular Depth Estimation. Following [34, 43, 52], we concatenate the global token to each patch token, and train a linear head to predict depth across 256 discretized bins. A classification loss is used during training on NYUv2 [40], with evaluation on its test set. See [52] for more details.

5.3 Ablations and Analysis

We organize our analysis around four core questions:

Q1: Individual Task Contributions.

We first break down the individual contribution of each task to isolate the sources of the gains. Table 3 reveals three observations: *(i) Universal improvement*: adding any single objective on top of VL pretraining yields measurable gains on every benchmark, indicating that the auxiliary signals provide non-redundant supervision rather than task-specific overfitting. Surprisingly, the depth supervision is beneficial to semantic tasks (*e.g.*, +2.8 IN-1k and +2.5 COCO^{I→T} for ViT-L at 50M), even though the depth signal itself carries no semantic information. We conjecture that regressing per-pixel depth forces the encoder to delineate object boundaries, yielding more object-centric features that in turn benefit instance-level recognition and alignment. *(ii) Specialized gains*: each objective excels on tasks aligned with its inductive bias. SSL primarily boosts semantic recognition and alignment (IN-1k, COCO, ADE20k, SPair), Depth contributes vital geometric priors (NAVI, NYUv2), and Grounding serves

Table 4: Marginal Task Gains across data and model scales. (Numbers in parentheses) indicate the incremental increase over the previous row, while **Absolute Gain Δ** denotes the overall improvement of the full multi-task model over the VL-only baseline.

ViT Data Training Tasks	IN-1k	COCO		VQA	ADE20k	NAVI	SPair	NYUv2
	Acc.	T→I	I→T	Score	mIoU	Recall	Recall	RMSE ↓
B/16 10M	VL	36.2	14.8	21.9	41.1	27.5	39.5	17.0
	VL + SSL	43.7(+7.5)	19.7(+4.9)	28.6(+6.7)	43.1(+2.0)	36.2(+8.7)	41.5(+2.0)	21.1(+4.1)
	VL + SSL + Ground.	49.0(+5.3)	23.5(+3.8)	33.9(+5.3)	43.2(+0.1)	39.5(+3.3)	43.3(+1.8)	22.4(+1.3)
	VL + SSL + Ground. + Depth	49.7(+0.7)	23.9(+0.4)	34.1(+0.2)	43.4(+0.2)	41.7(+2.2)	43.6(+0.3)	22.8(+0.4)
	Absolute Gain Δ	+13.5	+9.1	+12.2	+2.3	+14.2	+4.1	+5.8
B/16 50M	VL	55.7	28.2	43.0	42.2	32.2	39.8	18.0
	VL + SSL	61.4(+5.7)	32.4(+4.2)	48.0(+5.0)	44.7(+2.5)	39.9(+7.7)	40.4(+0.6)	22.7(+4.7)
	VL + SSL + Ground.	62.9(+1.5)	34.7(+2.3)	50.7(+2.7)	45.1(+0.4)	43.0(+3.1)	43.8(+3.4)	25.4(+2.7)
	VL + SSL + Ground. + Depth	62.9	35.0(+0.3)	50.5(-0.2)	45.8(+0.7)	43.7(+0.7)	44.1(+0.3)	24.5(-0.9)
	Absolute Gain Δ	+7.2	+6.8	+7.5	+3.6	+11.5	+4.3	+6.5
L/16 10M	VL	35.0	14.9	21.6	39.0	30.0	40.1	18.1
	VL + SSL	49.4(+14.4)	23.7(+8.8)	34.0(+12.4)	42.1(+3.1)	40.4(+10.4)	43.4(+3.3)	25.6(+7.5)
	VL + SSL + Ground.	53.4(+4.0)	25.6(+1.9)	37.8(+3.8)	43.7(+1.6)	43.6(+3.2)	44.3(+0.9)	27.0(+1.4)
	VL + SSL + Ground. + Depth	54.3(+0.9)	26.6(+1.0)	38.6(+0.8)	44.3(+0.6)	45.6(+2.0)	46.2(+1.9)	27.5(+0.5)
	Absolute Gain Δ	+19.3	+11.7	+17.0	+5.3	+15.6	+6.1	+9.4
L/16 50M	VL	60.8	31.1	46.3	43.3	37.3	39.9	19.8
	VL + SSL	65.4(+4.6)	34.5(+3.4)	48.0(+1.7)	44.9(+1.6)	45.5(+8.2)	43.2(+3.3)	27.3(+7.5)
	VL + SSL + Ground.	67.6(+2.2)	37.4(+2.9)	51.5(+3.5)	45.9(+1.0)	47.2(+1.7)	43.7(+0.5)	28.1(+0.8)
	VL + SSL + Ground. + Depth	67.9(+0.3)	37.3(-0.1)	52.3(+0.8)	45.9	49.1(+1.9)	45.9(+2.2)	29.7(+1.6)
	Absolute Gain Δ	+7.1	+6.2	+6.0	+2.6	+11.8	+6.0	+9.9

as a critical bridge that benefits both high-level semantics and fine-grained spatial structures. (iii) *Cross-task spillover*: despite these specializations, we observe meaningful improvements outside each task’s primary domain (e.g., Depth mostly helping semantic understanding), suggesting that the shared encoder is learning transferable primitives rather than siloed features.

Taken together, the ablations indicate that **each objective contributes a distinct yet complementary signal**, laying the foundation for the monotonic gains observed in the subsequent multi-task expansion of Q2.

Q2: Marginal Task Gains.

We then examine whether integrating heterogeneous objectives is consistently beneficial. Table 4 presents the incremental performance along a principled task expansion path: VL → +SSL → +Ground. → +Depth.

Across data and model scales, performance almost always improves monotonically at every step. **Absolute Gain Δ** highlights that the full multi-task framework is strictly better than the VL-only baseline. This demonstrates that **rather than interference, these objectives provide complementary supervision that enriches the shared representation**.

We also observe distinct roles for each objective: (i) SSL acts as a powerful general-purpose enhancer, significantly lifting both semantic (+14.4% on IN-1k for ViT-L/10M) and spatial benchmarks (+10.4 on ADE20k). This suggests that local structural regularities provide a necessary geometric scaffold for global semantics. (ii) Since grounding serves as a fine-grained vision-language signal, it not only boosts spatial tasks but also provides benefits to global multimodal understanding, such as the +3.8 R@1 gain in COCO I→T retrieval. (iii) Adding depth supervision consistently provides the final refinement, yielding maximum gains of 0.9 on IN-1k, 2.2 on ADE20k, and 0.051 on NYUv2, *etc.*

Table 5: Task Synergy (%) computed as Eq. (4), measuring the relative gain from combining two tasks beyond the larger of their individual improvements over the VL-only baseline. *COCO is short for COCO I→T retrieval.

ViT	Data	Baseline	TaskA	TaskB	Synergy (%)						
					IN-1k	COCO*	ADE20k	NAVI	SPair	NYUv2	Average
B/16	10M	VL	SSL	Ground.	62.0	64.4	37.9	72.7	31.7	40.8	51.7
			SSL	Depth	50.7	41.8	31.0	-10.0	26.8	56.6	32.9
			Depth	Ground.	22.8	27.4	33.8	36.4	51.5	56.2	38.1
B/16	50M	VL	SSL	Ground.	26.3	54.0	40.3	100.0	57.4	44.4	53.7
			SSL	Depth	8.8	4.0	27.3	75.0	61.7	29.0	34.3
			Depth	Ground.	2.6	26.3	40.7	87.5	78.8	19.4	42.6
L/16	10M	VL	SSL	Ground.	27.8	30.6	30.8	27.3	18.7	38.9	29.0
			SSL	Depth	17.4	21.8	27.9	24.2	24.0	53.6	28.1
			Depth	Ground.	9.6	3.5	32.0	42.9	0.0	12.5	16.7
L/16	50M	VL	SSL	Ground.	47.8	33.8	30.7	21.8	10.7	39.2	30.7
			SSL	Depth	43.5	92.0	19.5	9.1	-34.7	4.6	22.3
			Depth	Ground.	45.5	33.3	33.3	29.4	17.9	18.5	29.7

While the gains exhibit diminishing marginal returns as the representation nears saturation, the large initial jump from SSL should be interpreted with caution: this is a consequence of the chosen expansion order and the inherent spatial weakness of the VL-only baseline. Nevertheless, the consistent positive deltas in Table 3 and 4 verify the robustness of the joint optimization.

Q3: Synergy *vs.* Interference among Objectives.

We further investigate how heterogeneous objectives interact during joint optimization. To quantify their complementarity, we define *Task Synergy* as:

$$\text{Synergy}(A, B) = \frac{\Delta AB - \max(\Delta A, \Delta B)}{\max(\Delta A, \Delta B)}, \quad (4)$$

where ΔA and ΔB denote the individual performance gains over the VL baseline, and ΔAB is the gain achieved by their combination. Intuitively, a positive score indicates that the tasks are mutually beneficial, providing gains beyond those attainable by the superior individual task alone.

As shown in Table 5, all task pairs exhibit positive average synergy across data and model scales, typically ranging from 20% to 50%. This empirically confirms that **the objectives are fundamentally complementary rather than interfering**. Task pairs involving grounding show particularly robust synergy, reinforcing our argument that region-text alignment serves as a critical bridge between high-level semantics and fine-grained spatial structures.

We also observe distinct scaling trends between model and data scales: For the B/16 model, increasing data from 10M to 50M samples leads to more stable and stronger synergy. While synergy remains positive for the larger L/16 model, the relative scores are generally lower compared to B/16. This is expected as higher-capacity baselines naturally leave less relative headroom for additional supervision, even as absolute performance continues to rise.

Q4: Scaling Effects across Data and Model Capacity.

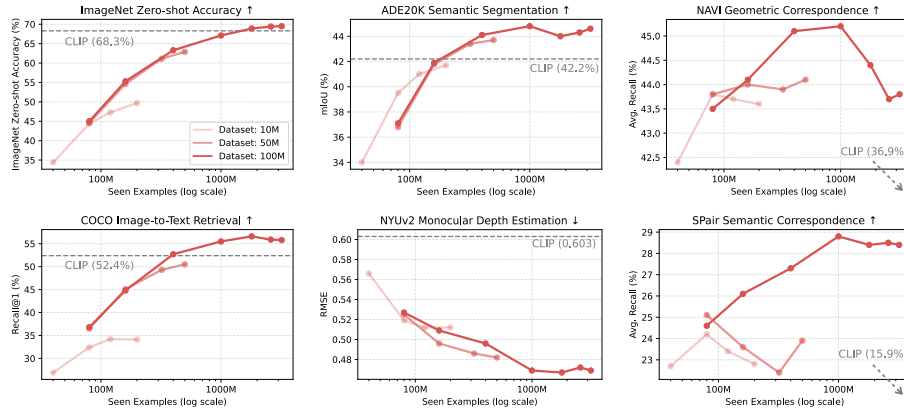


Fig. 3: Data scaling behavior of multi-task visual learning. ViT-Base models are trained with 10M, 50M, and 100M samples under our multi-task setting. CLIP-Base [36] is shown as gray dashed line. Lower $\downarrow$ is better for NYUv2; higher $\uparrow$ is better elsewhere.

We finally investigate how MTV scales with data size and model capacity. As illustrated in Fig. 3, scaling the data yields broadly monotonic improvements. Notably, our ViT-Base model trained on only 100M samples consistently outperforms the CLIP-Base baseline, which was trained on 400M samples. This demonstrates that **multi-task supervision can be more data-efficient than pure VL pretraining for building versatile visual encoders**.

However, the scaling behavior is not uniform across tasks. Geometric and semantic correspondence saturates and begins to decline once the number of seen samples exceeds 1B. We explain this from three aspects: (i) *Task disparity*: correspondence tasks require multi-view consistency. While depth helps, it remains a monocular proxy. The observed saturation reflects the limits of monocular supervision for multi-view tasks. Future work may benefit from incorporating multi-view consistency supervision. (ii) *Data quality bottleneck*: DataComp contains many geometrically “flat” images that hinder geometric learning. Preliminary experiments on a 10M subset show that filtering by depth variance significantly alleviates saturation (Fig. 4). More dedicated filtering methods could further improve performance, which we leave for future work. (iii) *Relative efficiency*: despite saturation, our model still outperforms CLIP/SigLIP on these tasks using much less training data (Table 6).

Moreover, model scaling from B/16 to L/16 (Table 3, 4, 6) improves performance across tasks under the same data scale, narrowing the gap to larger VL-only baselines while preserving strong spatial understanding.

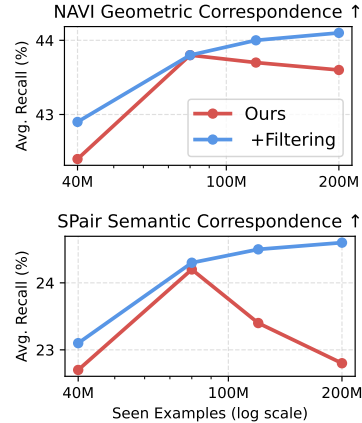


Fig. 4: Data filtering effects (ViT-B, 10M data $\times$ 20 epochs).

Table 6: Comprehensive comparisons with large-scale VLMs.

ViT	Model	Data	IN-1k	COCO		VQA	ADE20K	NAVI	SPair	NYUv2
			Acc.	T→I	I→T	Score	mIoU	Recall	Recall	RMSE ↓
B/16	CLIP [36]	400M	68.3	33.1	52.4	44.8	42.2	36.9	15.9	0.603
	SigLIP [53]	10B	76.2	47.2	64.5	47.7	45.1	38.5	17.3	0.615
	SigLIP2 [43]	10B	78.2	52.1	68.9	48.1	46.0	38.0	20.8	0.562
	MTV	100M	69.5	41.2	57.1	45.6	44.6	43.8	28.4	0.469
L/14	CLIP [36]	400M	75.5	36.5	56.3	46.5	46.2	36.0	20.5	0.588
	SigLIP [53]	10B	80.5	51.2	69.6	48.3	47.7	38.8	19.7	0.550
L/16	SigLIP2 [43]	10B	82.5	54.7	71.5	52.5	51.6	43.2	29.9	0.484
	MTV	100M	75.2	43.0	58.1	47.6	48.2	45.0	30.1	0.414

5.4 Comparison with Large-scale VLMs

Previous ablations (Table 3–5) are strictly controlled under an identical data budget and training setup to ensure clear causal attribution. Notably, our VL+SSL configuration serves as a representative proxy for several existing multi-task frameworks [28, 32, 33], while the further integration of dense supervision yields substantial gains beyond them.

Table 6 compares MTV with established VLMs, including CLIP as a widely used reference with comparable data scale, and SigLIP / SigLIP2 as stronger baselines trained on significantly larger datasets.

On **global semantic tasks** such as classification, retrieval, and VQA, our models consistently outperform CLIP, despite using only one-quarter of its training data. As expected, MTV lag behind SigLIP/SigLIP2, which are trained on 100× more data, but the gap remains modest given the vast scale difference. On **fine-grained spatial tasks**, the advantages of MTV are most prominent. Ours tail SigLIP2 on semantic segmentation, while achieving clearly better performance on correspondence and depth estimation.

Taken together, these results demonstrate that multi-task training offers a compelling and scalable alternative to purely vision–language or self-supervised pretraining for general-purpose visual representation learning.

6 Conclusion

In this work, we revisit large-scale multi-task training as a unified framework for learning general-purpose visual representations. By jointly optimizing vision–language contrastive, self-supervised, and pseudo-labeled dense objectives, our model acquires strong semantic and geometric understanding within a single encoder. Across extensive experiments, we demonstrate that multi-task supervision leads to steadily improving performance as more tasks are incorporated, exhibiting clear task synergy and favorable data- and model-scaling behavior. These findings suggest that multi-task pretraining, fueled by high-quality pseudo-supervision, provides an effective and data-efficient route toward universal visual foundation models for broad visual perception and reasoning tasks.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR (2023)
3. Birkel, R., Wofk, D., Müller, M.: Midas v3. 1—a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
6. Chen, H.Y., Lai, Z., Zhang, H., Wang, X., Eichner, M., You, K., Cao, M., Zhang, B., Yang, Y., Gan, Z.: Contrastive localized language-image pre-training. In: ICML (2025)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: CVPR (2024)
11. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. In: NeurIPS (2023)
12. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. In: NeurIPS (2023)
13. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
14. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. In: ICCV (2021)
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
16. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
17. Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L.: Open-set image tagging with multi-grained text supervision. arXiv preprint arXiv:2310.15200 (2023)
18. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin Brualla, R., Patel, K., et al.: Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In: NeurIPS (2023)

19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
20. Jia, J., Hu, Y., Weng, X., Shi, Y., Li, M., Zhang, X., Zhou, B., Liu, Z., Luo, J., Huang, L., Wu, J.: Tinyllava factory: A modularized codebase for small-scale large multimodal models. arXiv preprint arXiv:2405.11788 (2024)
21. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
22. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024)
23. Lai, Z., Zhang, H., Zhang, B., Wu, W., Bai, H., Timofeev, A., Du, X., Gan, Z., Shan, J., Chuah, C.N., et al.: Vecclip: Improving clip training via visual-enriched captions. In: ECCV (2024)
24. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024)
25. Liu, S., Johns, E., Davison, A.J.: End-to-end multi-task learning with attention. In: CVPR (2019)
26. Liu, Y., Li, X., Wang, Z., Zhao, B., Xie, C.: Clips: An enhanced clip framework for learning with synthetic captions. arXiv preprint arXiv:2411.16828 (2024)
27. Liu, Y., Zhang, Y., Ghosh, D., Schmidt, L., Yeung-Levy, S.: Data or language supervision: What makes CLIP better than DINO? In: EMNLP Findings (2025)
28. Maninis, K.K., Chen, K., Ghosh, S., Karpur, A., Chen, K., Xia, Y., Cao, B., Salz, D., Han, G., Dlabal, J., et al.: Tips: Text-image pretraining with spatial awareness. In: ICLR (2025)
29. Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
30. Minderer, M., Gritsenko, A., Hounsby, N.: Scaling open-vocabulary object detection. In: NeurIPS (2023)
31. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: CVPR (2016)
32. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: ECCV (2022)
33. Naeem, M.F., Xian, Y., Zhai, X., Hoyer, L., Van Gool, L., Tombari, F.: Silc: Improving vision language pretraining with self-distillation. In: ECCV (2024)
34. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
35. Qwen: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
38. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
39. Sablayrolles, A., Douze, M., Schmid, C., Jégou, H.: Spreading vectors for similarity search. In: ICLR (2019)

40. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: ECCV (2012)
41. Tong, P., Brown, E., Wu, P., Woo, S., IYER, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In: NeurIPS (2024)
42. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: CVPR (2024)
43. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
44. Vandenhende, S., Georgoulis, S., Van Gool, L.: MTI-Net: Multi-scale task interaction networks for multi-task learning. In: ECCV (2020)
45. Wan, B., Tschannen, M., Xian, Y., Pavetic, F., Alabdulmohsin, I.M., Wang, X., Susano Pinto, A., Steiner, A., Beyer, L., Zhai, X.: Locca: Visual pretraining with location-aware captioners. In: NeurIPS (2024)
46. xAI: Grok-1.5 vision preview (April 2024)
47. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR (2020)
48. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
49. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024)
50. Yu, Q., Sun, Q., Zhang, X., Cui, Y., Zhang, F., Cao, Y., Wang, X., Liu, J.: Caps-fusion: Rethinking image-text data at scale. In: CVPR (2024)
51. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
52. Yue, Y., Das, A., Engelmann, F., Tang, S., Lenssen, J.E.: Improving 2d feature representations by 3d-aware fine-tuning. In: ECCV (2024)
53. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)
54. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
55. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: ICLR (2022)

A Additional Implementation Details

The main paper provides a high-level summary of our experimental setup while omitting certain lower-level details for clarity. In this section, we offer a more comprehensive account of the implementation and training configurations.

A.1 Data Curation

We adopt DataComp [12] as the source for our pre-training data, which consists of 1.1 billion image-text pairs. Due to limitations in download bandwidth, we collected a subset of 200 million images from the full dataset. The download took around 40 days and resulted in nearly 21 TB of image-text data. We then applied the filtering strategy detailed in Section 5.1, obtaining a final corpus of 100 million image-text pairs used in our experiments.

We employ Depth Anything V2 (DA2-**Large**) [49] to compute a relative depth map for each image. Our depth-processing pipeline attains a throughput of approximately 120K images per GPU-hour, so producing depth maps for 100M images on 64 GPUs takes about 13 hours. This high throughput allows depth supervision to be easily scaled to much larger datasets.

For grounding supervision, we first use RAM++ [17] to extract salient entity names, and then apply OWLv2-**Base** [30] to localize the corresponding regions. This pipeline can handle roughly 18K images per GPU-hour, leading to a total of 86 GPU-hours to produce region-text pairs for 100M images on 64 GPUs. While this is slower than depth estimation, the throughput is still high enough to feasibly generate pseudo-labels at the billion-sample scale.

A.2 Additional Training Details

SSL Supervision. The projection heads for self-distillation and masked prediction follow the same architecture but do not share parameters. Each head is a 3-layer MLP with L2 normalization, followed by a weight-normalized projection layer that maps to a 32k-dimensional prototype space.

The teacher network is updated as the exponential moving average (EMA) of the student using a fixed momentum of 0.994. To avoid representation collapse, we apply both centering and sharpening after the projection head. For centering, we maintain a running mean of student logits (momentum 0.9) and subtract this mean during training. For sharpening, we use a fixed student temperature of $\tau_s = 0.1$ and linearly warm up the teacher temperature τ_t from 0.04 to 0.07.

Grounding Supervision. As described in Section 4.3, we follow the CLOC approach [6] to utilize the grounding supervision. CLOC introduces a lightweight Transformer encoder, referred to as the **Prompter**, which extracts regional visual features conditioned on bounding box locations. The **Prompter** takes three inputs: (i) the visual encoder output $\mathbf{z}_i^*$ before final normalization and attentive pooling, (ii) a location embedding l_i derived from the bounding box coordinates (top-left and bottom-right), each encoded with positional embeddings, and (iii)

a learnable query token $\mathbf{o}$. Given these inputs, the **Prompter** produces a regional visual feature $\mathbf{o}_i = \text{Prompter}(\mathbf{z}_i^*, l_i, \mathbf{o})$, which is then aligned to the regional text feature through the same contrastive formulation used in Eq. (1).

Depth Supervision. The depth head follows the DPT [37] decoder, a convolutional module that fuses features from four intermediate layers of the visual encoder: $\{3, 6, 9, 12\}$ for ViT-B/16 and $\{6, 12, 18, 24\}$ for ViT-L/16.

As described in Section 4.3, we adopt the MiDaS objective [3], combining a scale- and shift-invariant loss with a multi-scale gradient matching loss.

Regarding the scale- and shift-invariant loss, the pseudo depth d and predicted depth $\hat{d}$ are first normalized by:

$$d^* = \frac{d - t(d)}{s(d)}, \quad \hat{d}^* = \frac{\hat{d} - t(\hat{d})}{s(\hat{d})},$$

where $t(d) = \text{median}(d)$, $s(d) = \frac{1}{HW} \sum_{j=1}^{HW} |d - t(d)|$. The scale- and shift-invariant (ssi) loss is then defined as:

$$\mathcal{L}_{\text{ssi}} = \frac{1}{2HW} \sum_{j=1}^{HW} |d_j^* - \hat{d}_j^*|.$$

Following MiDaS [3], we trim the 10% largest residuals to mitigate pseudo-label noise: Let $r_j = |d_j^* - \hat{d}_j^*|$ sorted in ascending order, we keep only the smallest $U = 0.9HW$ residuals and define $\mathcal{L}_{\text{ssitrim}} = \frac{1}{2HW} \sum_{j=1}^U r_j$.

In addition, the gradient matching (gm) loss penalizes the depth discrepancies between the ground truth and the rescaled prediction across $K = 4$ scales:

$$\mathcal{L}_{\text{gm}} = \sum_{k=1}^K \sum_{j=1}^{HW} \left(|\nabla_x^k d_j - s \nabla_x^k \hat{d}_j| + |\nabla_y^k d_j - s \nabla_y^k \hat{d}_j| \right),$$

where ∇_x^k and ∇_y^k denote finite differences along the horizontal and vertical directions at scale k .

The final depth loss is a weighted combination:

$$\mathcal{L}_{\text{depth}} = \mathcal{L}_{\text{ssitrim}} + 2 \times \mathcal{L}_{\text{gm}}.$$

Computational Cost. For the largest 100M setting, our **MTV-Base** is trained on 32 GPUs for 5.7 days, while **MTV-Large** requires 64 GPUs for 9.4 days.

B Additional Experiments

B.1 Zero-shot Relative Depth Estimation

Our model inherently supports zero-shot depth estimation through its pretraining phase. In Table 7, we evaluate MTV against state-of-the-art depth specialists on KITTI [13] and NYUv2 [40]. Notably, despite being trained in a multi-task framework and operating at a significantly lower input resolution than dedicated models (e.g., Depth Anythings v1/v2 at 518^2), MTV achieves competitive zero-shot performance across both benchmarks.

Table 7: Zero-shot relative depth estimation, measured by AbsRel $\downarrow$.

Model	Input	KITTI	NYUv2
Marigold [22]	480×640	9.9	5.5
DAv1-B [48]	518×518	8.0	4.6
DAv2-B [49]	518×518	7.8	4.9
MTV-B	224×224	8.9	6.0
MiDaS v3.1 [3]	518×518	12.7	4.8
DAv1-L [48]	518×518	7.6	4.3
DAv2-L [49]	518×518	7.4	4.5
MTV-L	256×256	8.2	5.2

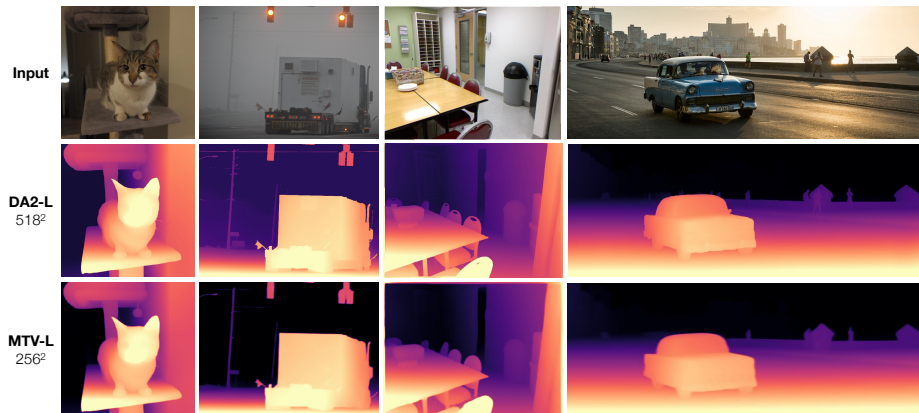


Fig. 5: Visualizations of zero-shot relative depth estimation.

Beyond the quantitative metrics, Figure 5 also presents qualitative comparisons. Our model generates depth maps that capture the correct global geometry and maintain a coherent scene structure across a wide variety of images. Relative to the depth-specialized DA2 model, some fine-scale details (*e.g.*, the cat’s whiskers or the chair back) appear less crisp, primarily due to the difference in resolution. Despite this, the predictions remain generally stable and accurate, indicating that low-resolution, pseudo-labeled depth supervision, when incorporated into multi-task pretraining, can still produce strong and well-generalized geometric representations.

B.2 Ablation on the Depth Teacher

To evaluate our method’s sensitivity to teacher capacity and the quality of pseudo-labels, Table 8 compares configurations using **Depth-Anything-V2-Small** (25M parameters, denoted as “DA2-S”), **Depth-Anything-V2-Large** (335M parameters, “DA2-L”), and a baseline without depth supervision (“N/A”).

Our analysis reveals several key insights:

Table 8: Ablation on Depth Teacher with ViT-B and 10M Data.

Teacher	IN-1k	COCO _{T→I}	COCO _{I→T}	ADE20k	NAVI	SPair	NYUv2 ↓
N/A	49.0	23.5	33.9	39.5	43.3	22.4	0.537
DA2-S	49.5	23.7	34.2	41.1	42.1	21.7	0.527
DA2-L	49.7	23.9	34.1	41.7	43.6	22.8	0.512

- **Stability in Semantic Tasks:** Performance on high-level semantic benchmarks remains remarkably stable as teacher capacity scales (e.g., IN-1k: 49.5 → 49.7; COCO_{T→I}: 23.7 → 23.9; COCO_{I→T}: 34.2 → 34.1). The marginal fluctuations suggest that distilling geometric priors from depth teachers primarily enriches structural understanding without compromising or significantly altering the semantic feature space.
- **Sensitivity of Spatial Understanding:** In contrast, tasks requiring fine-grained spatial reasoning exhibit a pronounced correlation with teacher capacity. Upgrading from DA2-S to DA2-L leads to consistent improvements across ADE20k (41.1 → 41.7), NAVI (42.1 → 43.6), SPair (21.7 → 22.8, and NYUv2 (0.527 → 0.512). This indicates that higher-fidelity pseudo-depth labels are crucial for capturing fine-grained structural details.
- **Efficacy of Structural Priors:** Notably, even the lightweight DA2-S teacher outperforms the “N/A” baseline across most benchmarks (IN-1k, COCO, ADE20k, and NYUv2). This suggests that the inclusion of depth as an auxiliary objective introduces beneficial structural priors, and these gains are largely orthogonal to the absolute precision of the teacher model itself.

B.3 Gradient Analysis

We measure pairwise gradient cosine similarity on the shared ViT encoder (Table 9). All pairs show positive alignment with moderate conflict rates and balanced norm ratios, confirming that no objective monopolizes the encoder. These observations suggest limited optimization conflict: gradients are positively aligned on average, added objectives do not induce clear trade-offs on our benchmarks, and task pairs show positive empirical synergy.

Table 9: Gradient analysis. Conflict: % batches with negative cosine. Norm ratio: larger/smaller.

Task pair	Grad. cos. ↑	Conflict ↓	Norm ratio
VL / SSL	0.18	21.4	1.23
VL / Ground.	0.24	17.8	1.96
SSL / Depth	0.14	24.1	1.84
Ground. / Depth	0.27	15.9	1.18

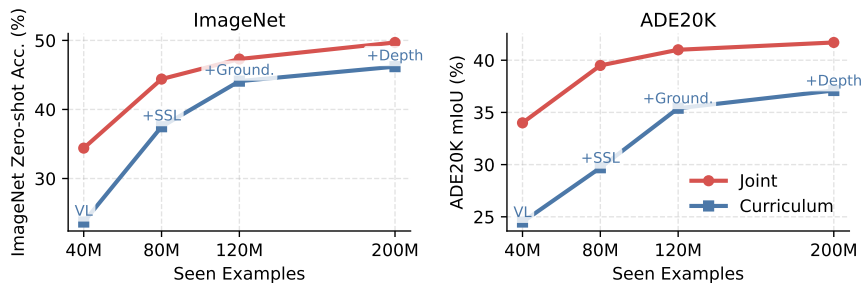


Fig. 6: Joint vs. curriculum learning.

B.4 Joint vs. curriculum learning.

Under the same 10M data budget, we compare joint training with an additive VL→SSL→Grounding→Depth curriculum. Joint training uses all objectives from the start and consistently performs better (Fig. 6), with final gains of +3.5 IN-1k and +4.6 ADE20K. This suggests that early co-training better aligns semantic and spatial signals.

C Limitations and Future Work

While our study provides a systematic examination of multi-task visual pretraining, several avenues for further investigation remain.

More Dedicated Data Filtering. Our pseudo labels were generated automatically without exhaustive filtering. As a result, some images exhibit low-quality captions, noisy grounding annotations, or unreliable depth estimates. While our analysis suggests the model is robust to these imperfections, developing more sophisticated data-cleaning pipelines or noise-adaptive training objectives could further refine representation quality. Exploring the trade-off between "data quantity" and "label fidelity" in multi-task settings is a promising direction.

Expanding the Modality Horizon. Our current framework focuses on single-view image pretraining. Extending this multi-task formulation to temporal or multi-view data—such as video sequences or stereo imagery—could unlock richer supervision signals. Integrating objectives like optical flow, 4D consistency, or video-language alignment would be a natural progression toward learning truly universal and physically-grounded visual representations.