Visual Bridge: Universal Visual Perception Representations Generating

Yilin Gao[1*], Shuguang Dou[2], Junzhou Li[3], Zhiheng Yu[2], Yin Li[2], Dongsheng Jiang[2†], Shugong Xu[4†]

[1]Shanghai University
[2]Huawei Technologies Co., Ltd.
[3]University of Science and Technology of China
[4]Xi'an Jiaotong-Liverpool University
gaoyilin@shu.edu.cn, doushuguang@huawei.com, ljz0824@mail.ustc.edu.cn, yuzhiheng8@huawei.com, sherrylee90@163.com, dongsheng_jiang@outlook.com, shugong.xu@xjtlu.edu.cn

Abstract

Recent advances in diffusion models have achieved remarkable success in isolated computer vision tasks such as text-to-image generation, depth estimation, and optical flow. However, these models are often restricted by a "single-task-single-model" paradigm, severely limiting their generalizability and scalability in multi-task scenarios. Motivated by the cross-domain generalization ability of large language models, we propose a universal visual perception framework based on flow matching that can generate diverse visual representations across multiple tasks. Our approach formulates the process as a universal flow-matching problem from image patch tokens to task-specific representations rather than an independent generation or regression problem. By leveraging a strong self-supervised foundation model as the anchor and introducing a multi-scale, circular task embedding mechanism, our method learns a universal velocity field to bridge the gap between heterogeneous tasks, supporting efficient and flexible representation transfer. Extensive experiments on classification, detection, segmentation, depth estimation, and image-text retrieval demonstrate that our model achieves competitive performance in both zero-shot and fine-tuned settings, outperforming prior generalist and several specialist models. Ablation studies further validate the robustness, scalability, and generalization of our framework. Our work marks a significant step towards general-purpose visual perception, providing a solid foundation for future research in universal vision modeling.

Introduction

In recent years, diffusion models have achieved remarkable success in text-to-image generation, depth estimation, and optical flow estimation. However, these models are typically trained under a "single-task-single-model" paradigm, limiting their generalizability. In contrast, large-scale language models (LLMs) like GPT-4 (OpenAI 2024) and DeepSeek-R1 (DeepSeek-AI 2025) exhibit strong cross-domain zero-shot transfer capabilities, enabled by universal architectures and large-scale pretraining—demonstrating the clear advantages of general-purpose modeling. Motivated by this observation, the paper explores: *Is it possible to achieve a universal generation of the diverse visual perception representation by diffusion models?*

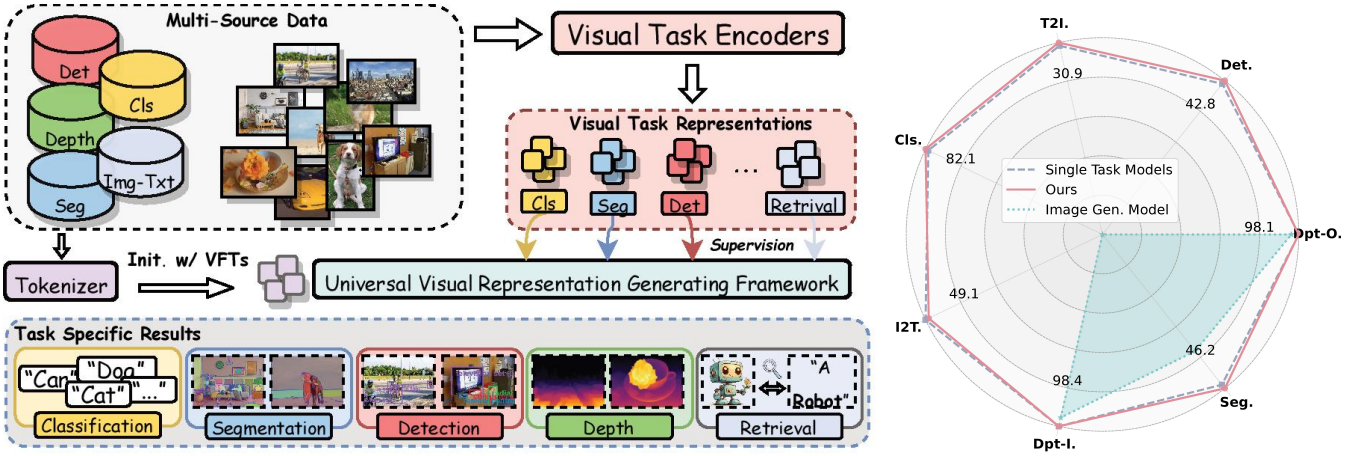
Generating representations for diverse visual perception tasks within a universal diffusion-based framework presents significant challenges. Classification models (e.g. ResNet (He et al. 2016), ViT (Dosovitskiy et al. 2020)) are designed to encode global semantic features, detection models (e.g., Faster R-CNN (Ren et al. 2016)) depend on multi-scale feature pyramid networks (FPNs) to localize and recognize objects at various scales, and segmentation models (e.g., U-Net (Ronneberger, Fischer, and Brox 2015)) rely on encoder–decoder architectures to recover fine-grained spatial details. These structural differences highlight the difficulty of designing a single framework capable of effectively handling the heterogeneous requirements of multiple tasks.

Drawing inspiration from the cross-domain generalization capabilities exhibited by LLMs, we posit that a universal visual representation generating framework should possess the following essential characteristics:

- **Task Generality**: The capacity to generate representations suitable for a variety of visual tasks without requiring separate models or extensive fine-tuning.
- **Architectural Flexibility**: The ability to adapt seamlessly to different network architectures and scales, supporting both global semantic reasoning and fine-grained spatial understanding.
- **Model Efficiency**: The ability to approximate the representations of larger models using compact architectures, enabling efficient deployment without significant performance loss.

To address these challenges, we propose a novel **universal flow-matching framework** for multi visual perception tasks. In contrast to previous works that treat perception tasks as an independent generation or regression prob-

[*]This work was completed during an internship at Huawei Technologies Co., Ltd.
[†]Both Shugong Xu and Dongsheng Jiang are corresponding authors.



(a) One Universal Visual Representation Generating Framework for Multi-Tasks

(b) Performance on Five Task Datasets

Figure 1: (a) We introduce *Vision Bridge*, a universal framework that bridges image patch tokens and fundamental vision tasks through flow-based modeling. The architecture supports diverse downstream tasks, including classification, detection, segmentation, depth estimation, and image-text retrieval. *Vision Bridge* enables task-agnostic representation learning and task-specific adaptation without introducing external data. (b) Radar plot comparing performance across five core vision tasks. Our method consistently outperforms strong baselines on ImageNet-1K (Deng et al. 2009), COCO (Lin et al. 2014), ADE20K (Zhou et al. 2019), and NYUv2 (Nathan Silberman and Fergus 2012).

lem (Zhao et al. 2025; Le et al. 2024), our approach formulates multi-vision perception tasks as *flow matching problem from image patch tokens to task-specific representations*, as shown in Fig. 1. By leveraging powerful self-supervised models such as DINOv2 (Oquab et al. 2023) as the foundation and learning a universal velocity field conditioned on task embeddings, we enable smooth and efficient representation transfer across various vision tasks.

Our key contributions can be summarized as follows:

- We introduce a general and scalable *flow-matching paradigm* that unifies visual perception tasks by aligning foundation model tokens with task-specific representations, moving beyond traditional image-to-image translation or single-stage distillation methods.
- We design a multi-scale, circular task embedding scheme that enables flexible encoding and interpolation across heterogeneous vision tasks, supporting both dense and sparse prediction scenarios.
- Extensive experiments of our model on five vision tasks demonstrate competitive performance in both zero-shot and fine-tuned settings, outperforming prior generalist models and task-specific baselines.

In summary, our work takes an important step towards **general-purpose visual perception**, bridging the gap between task-specific and universal vision models. We believe our approach can serve as a solid foundation for future research in multi-task and foundation vision modeling.

Related Work

Vision Foundation Models

Vision foundation models (VFMs) such as DINOv2 (Oquab et al. 2023), CLIP (Radford et al. 2021), and MAE (He et al. 2022a) learn powerful general-purpose representations through large-scale pretraining. While they exhibit strong cross-task transferability, task-specific architectures—e.g., ViT (Dosovitskiy et al. 2020) for classification, DETR (Carion et al. 2020) for detection, and Mask2Former (Cheng et al. 2022) for segmentation—still achieve superior performance on individual tasks. Recent efforts to adapt VFMs to downstream tasks (Wang, Ye, and Zhang 2024; Khanna et al. 2024; Li et al. 2024; Rajič et al. 2023; Zhong et al. 2024; Zhu et al. 2024; Gao et al. 2025) often rely on custom adapters, specialized heads, or extensive fine-tuning, which compromise flexibility and scalability. Although universal representation learning methods (Ren et al. 2024a; He et al. 2022b; Oquab et al. 2023; Caron et al. 2021) aim to close this gap, they typically lag behind specialist models in task-specific accuracy. In contrast, our framework directly bridges VFM patch tokens to diverse task representations via a unified flow-matching formulation, enabling strong performance without architectural redesign.

Multi-task Generalist Models

Multi-task vision systems generally adopt either shared encoder-decoder frameworks (Lu et al. 2022; Bachmann et al. 2024; Mizrahi et al. 2023) or vision-language models (VLMs) (Bai et al. 2023; Chen et al. 2024; Lu et al. 2024; Ren et al. 2024b). Encoder-decoder approaches use task-specific decoders but require large labeled datasets per task; VLMs excel in multimodal reasoning yet underperform

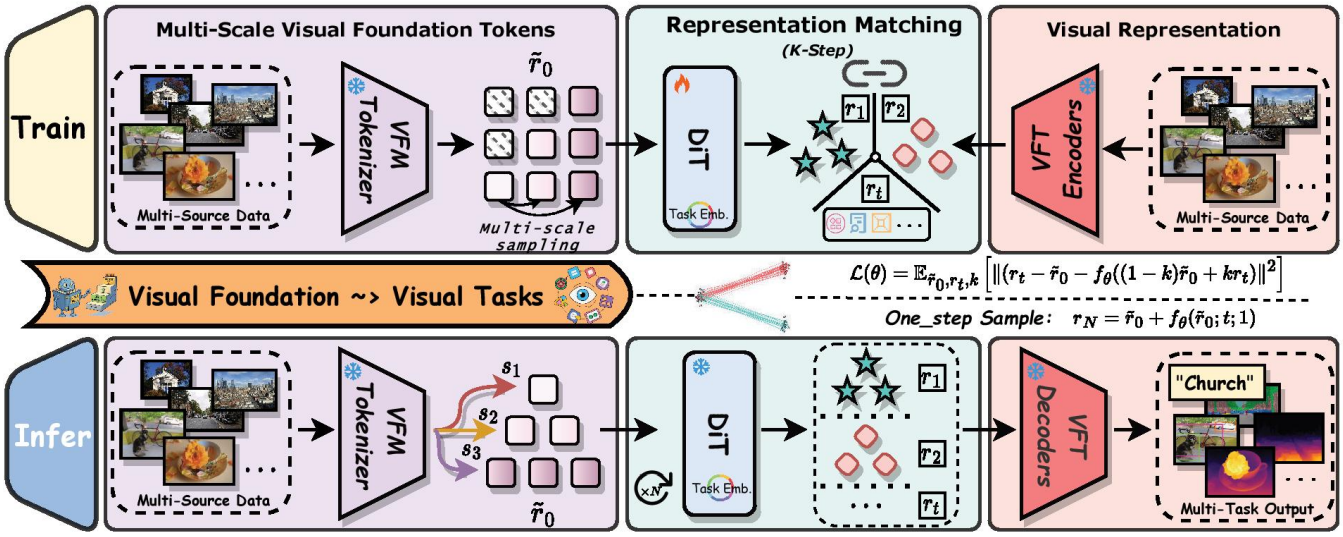


Figure 2: Overview of the proposed Visual Bridge. During training, tokens from the foundation model are sampled and interpolated with task-specific representations at multiple scales. A universal velocity field, conditioned on circular task embeddings and learnable scale embeddings, models the dynamics at each step. During inference, the learned flow is integrated to generate task-specific outputs (e.g., bounding boxes, labels, segmentation masks) using dedicated decoders. The proposed architecture enables efficient and flexible unification of a wide range of visual perception tasks.

on dense perception tasks like depth estimation or segmentation. More recently, diffusion-based unification methods (Le et al. 2024; Zhao et al. 2025) repurpose generative models for multi-task understanding. However, by framing tasks as image-to-image generation, they prioritize pixel-level reconstruction over semantic representation quality, leading to suboptimal results in non-generative tasks such as classification or retrieval.

Our work departs from these paradigms by formulating multi-task perception as a *flow matching problem from image patch tokens to task-specific representations*. This token-to-representation perspective establishes a task-agnostic bridge that supports seamless switching across heterogeneous tasks with minimal architectural changes and efficient sampling-based inference.

Method

Objective. We formulate the process of multi-vision perception tasks as a *universal flow matching* problem from image tokens to **Vision Foundation Task (VFT)** representations, as shown in Fig. 2. Specifically, we treat established task-specific models (e.g., DeiT for classification, DETR for detection) as target representations and utilize powerful self-supervised vision models such as DINOv2 (Oquab et al. 2023) as the initial flow anchor, as opposed to VAE-based latent representations in generative models. This enables us to effectively exploit the intrinsic visual reasoning capabilities of vision foundation models, thereby advancing universal learning across diverse vision tasks.

Mathematically, let $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ denote a raw image sampled from a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$, where N is the number of training samples. We extract image tokens $\mathbf{r}_0$ from a self-supervised VFM, such as DINOv2 (Oquab et al. 2023), and task-specific representations $\mathbf{r}_t$ from corresponding encoders. Our goal is to learn a velocity field f_θ that maps $\mathbf{r}_0$ to $\mathbf{r}_t$ under a flow-matching objective, conditioned on task embeddings e_t, which captures the dynamics of task-specific transformations. This formulation enables efficient multi-task learning by aligning intermediate representations via a universal flow field.

Training as Universal Flow-matching. Our training pipeline is based on a universal flow-matching paradigm, as detailed in Algorithm 1. At each training step, we sample a batch $\mathbf{x}$ from the dataset $\mathcal{D}$ and extract its VFM tokens $\mathbf{r}_0$ using a frozen DINOv2 tokenizer. Simultaneously, we extract task-specific representations $\mathbf{r}_t$ using a task-specific encoder E_{task}. In particular, to accommodate VFTs involving multi-scale features, we sample r_0 to match the resolutions present in $\mathbf{r}_t$, constructing multi-scale tokens $\tilde{\mathbf{r}}_0$ to prepare for subsequent multi-scale modeling.

These tokens are interpolated at discrete steps k to generate intermediate states $\mathbf{r}_k$, which serve as inputs to the velocity field f_θ. To ensure smooth transitions between $\tilde{\mathbf{r}}_0$ and $\mathbf{r}_t$, we employ a linear interpolation scheme:

$$\mathbf{r}_k = \left(1 - \frac{k}{K}\right)\tilde{\mathbf{r}}_0 + \frac{k}{K}\mathbf{r}_t \tag{1}$$

The velocity field f_θ predicts the dynamics of these interpolations, and the ground-truth velocity is computed as the difference between task and foundation representations:

$$v_k^{\text{true}} = \mathbf{r}_t - \tilde{\mathbf{r}}_0 \tag{2}$$

Algorithm 1: Universal Flow-matching Approach

Require: Training dataset $\mathcal{D} = \{(x^{(i)})\}_{i=1}^{N}$, where $x^{(i)}$ represents an input image
Model parameters θ, learning rate η, epochs E
Ensure: Trained model parameters θ

```
Randomly initialize θ
for e = 1 to E do
    for each batch x ~ D do
        r_0 ← DINOV2_Tokenizer(x)
        r_t ← E_task(x)
        Generate multi-scale tokens by sampling r_0 to various scales: r̃_0 = MultiScalesample(r_0).
        Sample k ~ Discrete(0, K)
        r_k ← (1 − k/K) r̃_0 + (k/K) r_t
        Predict velocity field: v̂_k = f_θ(r_k, k, s, e_t)
        Compute ground-truth velocity: v_k^true = r_t − r̃_0
        Compute loss: L(θ) = ‖v̂_k − v_k^true‖^2
        θ ← θ − η ∇_θ L(θ)
    end for
end for
```

To distinguish across different scales and tasks, we augment $\tilde{\mathbf{r}}_0$ with both scale and task embeddings. For scale embeddings, we introduce the learnable scale embedding module $\mathbf{s} \in \mathbb{R}^{L\times d}$, where L denotes the number of pre-defined scale levels and d is the embedding dimension. Each scale level corresponds to a specific resolution in the multi-scale token hierarchy. During training, the scale embedding s is indexed according to the current token scale and concatenated with the input representation $\tilde{\mathbf{r}}_0$, enabling the model to condition its predictions on the spatial resolution. Formally, we define:

$$s = \text{ScaleEmbed}(l), \tag{3}$$

where $l \in \{0, ..., L-1\}$ indicates the current scale level. For task embeddings, instead of adopting the text-prefix-style used in T2I tasks (Le et al. 2024), we propose a *circular task embedding* to map discrete task identities into continuous representations. Given T tasks, each task index $t \in \{0, ..., T-1\}$ is mapped onto the unit circle by defining its angular position as:

$$\theta_t = \frac{2\pi t}{T}. \tag{4}$$

This angle is then expanded into a vector of frequencies and used to construct a d-dimensional embedding via sinusoidal functions:

$$\begin{aligned}\mathbf{e}_t = \big[&\cos(\theta_t),\ \sin(\theta_t),\\ &\cos(2\theta_t),\ \sin(2\theta_t),\\ &\dots,\\ &\cos\left(\tfrac{d}{2}\theta_t\right),\ \sin\left(\tfrac{d}{2}\theta_t\right) \big]^\top.\end{aligned} \tag{5}$$

This design ensures that task representations reside on a smooth, periodic manifold, facilitating task variation and supporting multi-scale modeling within multi-scale vision foundation tasks.

The training objective aims to minimize the discrepancy between predicted and ground-truth velocities:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|f_\theta(\mathbf{r}_k, k, s, e_t) - v_k^{\text{true}}\|^2\right] \tag{6}$$

This flow-matching objective encourages f_θ to capture task-specific dynamics while preserving the semantic structure of the foundation model.

Inference as Flow Integration. During inference, we employ a deterministic integration process to evolve $\mathbf{r}_0$ into $\mathbf{r}_t$ for a given task t. Starting from the VFM representation $\mathbf{r}_0$, we sample r_0 to generate multi-scale tokens $\tilde{\mathbf{r}}_0$ and iteratively update the state using the learned velocity field f_θ. The integration is implemented using Euler discretization:

$$r_{n+1} \leftarrow r_n + f_\theta(r_n, n, s, e_t)/N \tag{7}$$

where N denotes the total step size and n represents the current sampling steps. The final state $\mathbf{r}_N$ is decoded into the target output (e.g., bounding boxes, class labels) through a task-specific decoder D_{task}. This integration procedure is detailed in Algorithm 2.

Algorithm 2: Inference Algorithm (Euler Method)

Require: Input image x, target task t, number of integration steps N
Ensure: Sampled output $y \in \mathcal{Y}$ (e.g., bounding box, class label)

```
r_0 ← DINOV2_Tokenizer(x)
Generate multi-scale tokens by sampling r_0 to different scales: r̃_0 = MultiScalesample(r_0).
Initialize state: r ← r̃_0
for n = 0 to N − 1 do
    Predict velocity at current state: v = f_θ(r_n, n, s, e_t)
    Update state using Euler's method: r_{n+1} ← r_n + v/N
end for
y = D_task(r_N)
return y
```

Experiments

Implementation Details

Data. We evaluate our method on diverse vision tasks: image classification, object detection, semantic segmentation, depth estimation, and image-text retrieval. We use standard benchmarks without introducing external data: **ImageNet-1K** for classification, **COCO** for detection and retrieval, **ADE20K** for segmentation, and **NYUv2/KITTI** for indoor/outdoor depth estimation. All datasets are used in their standard splits, consistent with recent foundation model evaluations, following conventions in recent multi-task learning and vision foundation model studies.

Training. Our framework is built on the **DINOv2-base** vision foundation model (Oquab et al. 2023). Task-specific heads are based on standard architectures: DeiT (ViT-B) (Touvron et al. 2021), DETR (Carion et al. 2020), Mask2Former (Cheng et al. 2022), CLIP (Radford et al.

Method	Acc@1 (%)	Patch Size
DeiT-B Baseline	**81.8**	14×14
MAE (He et al. 2022a)	79.8	14×14
CLIP (Radford et al. 2021)	78.1	7×7
OSD	81.2	16×16
Noise†	0.1	16×16
Ours (Zero-shot)	81.5	16×16
Ours (Fine-tuned)	**82.1**	16×16

Table 1: Zero-shot and fine-tuned performance (Top-1 Accuracy %) on ImageNet-1K. (†) denotes using noise as input with DINOV2 embedding as a condition.

2021), and Depth Anything (Yang et al. 2024a). The diffusion component uses **DiT-B** (Peebles and Xie 2023). We use a detection-style data pipeline: images in a batch are padded to the maximum resolution to preserve aspect ratios. Models are trained for 300 epochs with AdamW (1×10^{-4} learning rate, 0.01 weight decay) and a cosine scheduler. For fair evaluation of representation, we keep backbone fixed during fine-tuning and only update task heads. Ablation studies explore different DiT scales to assess model capacity.

Results on Visual Perception Tasks

Classification. We evaluate the quality of visual representations on ImageNet-1K (Deng et al. 2009) using DeiT-B (Touvron et al. 2021) as the target architecture. Our goal is to assess whether our framework learns transferable features that match or exceed those from task-specific pre-training. As shown in Table 1, our method achieves a zero-shot accuracy of **81.5%**, closely matching the supervised baseline (81.8%). Fine-tuning only the classification head further improves performance to **82.1%**, demonstrating strong representation quality and adaptability.

An intuitive attempt is to fully follow the pipeline of the flow-matching model, which uses random noise as r_0 and the embedding extracted by the vision foundation models as the condition. However, this results in poor performance (0.1%), likely due to the lack of meaningful initialization and insufficient conditioning for fine-grained semantics.

To better understand the impact of different vision foundation models, we compare our approach against MAE (He et al. 2022a) and CLIP (Radford et al. 2021). When transferred to DeiT-B without fine-tuning, MAE obtains 79.8%, while CLIP yields a significantly lower accuracy of 78.1%. The performance gap may be due to CLIP's extremely low-resolution representation: it uses a 32× downsampling strategy (from 224×224 input to 7×7 patch tokens), which severely limits its ability to capture fine-grained visual details. In contrast, our method, built upon DINOv2 (Oquab et al. 2023), benefits from a dense and high-resolution tokenization scheme (with a patch size of 16×16) and self-distillation during training, resulting in richer and more localized visual representations. These advantages enable our model to outperform both MAE and CLIP under identical downstream settings, validating the effectiveness of DINOv2 as a strong foundation for multi-task learning.

Method	mAP (%)	Encoder
DETR Baseline	**41.9**	ResNet-50
Ours (Zero-shot)	39.2	DiT-B
Ours (Fine-tuned)	**42.8**	DiT-B

Table 2: Comparison of object detection performance (mAP %) on COCO.

Method	PQ	AP	mIoU
Mask2Former (ResNet50)	39.7	**26.5**	46.1
Ours (Zero-shot)	37.9	24.5	44.6
Ours (Fine-tuned)	**40.0**	26.4	**46.2**

Table 3: Semantic segmentation performance on ADE20K (PQ / AP / mIoU).

We also compare our method with a One-Step Distillation (OSD) variant under the same network architecture and training setup. This alternative strategy achieves an accuracy of 81.2%, which is slightly lower than ours. This suggests that generative learning approach still retains strong representational power and outperforms distillation-based methods that rely on a single-stage knowledge transfer.

Detection. We evaluate our method on the object detection task using the COCO dataset (Lin et al. 2014), with DETR (Carion et al. 2020) as the target architecture equipped with a ResNet-50 backbone.

Our approach demonstrates strong zero-shot transfer capability, achieving an mAP of **39.2** without any task-specific adaptation. This result indicates that the representations learned by our foundation model can also capture rich semantic and spatial structures, enabling effective deployment to downstream architectures even in the absence of fine-tuning. When fine-tuning only the detection head, we further improve the performance to **42.8**, surpassing the original DETR baseline by a clear margin.

Notably, this cross-backbone compatibility highlights the **universality and robustness** of our learned representations and framework, making our approach a promising foundation for multi-task and multi-architecture vision systems.

Segmentation. We evaluate our framework on the ADE20K (Zhou et al. 2017), comparing our *DiT-B* against a ResNet50-based *Mask2Former* (Cheng et al. 2022) with **multi-scales**. As shown in Table 3, our method achieves competitive performance in a zero-shot manner, attaining 37.9 PQ, 24.5 AP, and 44.6 mIoU. Although the performance is slightly insufficient, our method obtains a more fine-grained segmentation result.

Furthermore, with fine-tuning, our method demonstrates clear improvements across all metrics, confirming its strong adaptability to downstream tasks.

Image-Text Retrieval. We further evaluate the cross-modal alignment capability of our framework on the image-

Method	Text-to-Image (T2I)			Image-to-Text (I2T)		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-B	30.4	54.8	66.1	**50.0**	**74.9**	83.2
Ours	**30.9**	**56.2**	**67.4**	49.1	**74.9**	**83.6**

Table 4: Zero-shot image-text retrieval performance (Recall@K, %) on COCO.

text retrieval task using the COCO dataset (Lin et al. 2014). Following standard evaluation protocols, we report Recall at K (R@K) for K=1, 5, and 10 in both text-to-image (T2I) and image-to-text (I2T) settings.

Our method demonstrates strong zero-shot transfer performance when compared to CLIP-B (Radford et al. 2021), a widely recognized foundation model for vision-language learning. As shown in Table 4, our approach achieves T2I scores of **30.9** (R@1), **56.2** (R@5), and **67.4** (R@10), outperforming CLIP-B across all metrics. In the I2T direction, our results are highly competitive, with scores of 49.1 (R@1), 74.9 (R@5), and 83.6 (R@10), closely matching those of CLIP-B.

Depth. We evaluate our framework on monocular depth estimation using KITTI (Geiger et al. 2013) and NYUv2 (Nathan Silberman and Fergus 2012), reporting the standard metrics AbsRel (↓) and δ_1 (↑). As shown in Table 5, our method achieves strong results: on KITTI, we obtain AbsRel = 0.048 and δ_1 = 0.981—approaching Depth Anything (Yang et al. 2024b) and Depth Anything v2 (Yang et al. 2024c), and outperforming the dedicated model Depth Pro (Bochkovskii et al. 2024); on NYUv2, we reach AbsRel = 0.056 and δ_1 = 0.984.

Our approach learns a flow-based mapping to mimic Depth Anything's in-domain outputs, inheriting its geometric priors while enabling efficient deployment. It also significantly outperforms generalist vision models—including 4M-XL (Mizrahi et al. 2023), Painter (Wang et al. 2023), Unified-IO (Lu et al. 2022), DICEPTION (Zhao et al. 2025), and OneDiffusion (Le et al. 2024)—e.g., gaining +8.0% δ_1 over OneDiffusion and +4.8% over DICEPTION on KITTI. This highlights our framework's superior preservation of spatial and geometric structure for dense prediction.

For completeness, Table 5 also includes comparisons with established depth estimators: MiDaS (Ranftl et al. 2020), Omnidata (Eftekhar et al. 2021), Metric3D v2 (Hu et al. 2024), DiverseDepth (Yin et al. 2020), LeReS (Yin et al. 2021), HDN (Yu et al. 2024), GeoWizard (Fu et al. 2024), DepthFM (Gui et al. 2024), Marigold (Ke et al. 2024), and GenPercept (Xu et al. 2024).

Method	KITTI		NYUv2	
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
MiDaS	0.236	0.630	0.111	0.885
Omnidata	0.149	0.835	0.074	0.945
Metric3D v2	0.052	0.979	0.039	0.979
DiverseDepth	0.190	0.704	0.117	0.875
LeReS	0.149	0.784	0.090	0.916
HDN	0.115	0.867	0.069	0.948
GeoWizard	0.097	0.921	0.052	0.966
DepthFM	0.083	0.934	0.065	0.956
Marigold	0.099	0.916	<u>0.055</u>	0.964
GeoWizard	0.129	0.851	0.059	0.959
DepthFM	0.174	0.718	0.082	0.932
Genpercept	0.094	0.923	0.091	0.932
Depth Pro	0.055	0.974	**0.042**	<u>0.977</u>
DepthAnything	0.046	0.982	0.056	**0.984**
DepthAnything v2	**0.045**	**0.983**	0.056	**0.984**
4M-XL	0.105	0.896	0.068	0.951
Painter	0.324	0.393	0.046	0.979
Unified-IO	0.188	0.699	0.059	0.970
OneDiffusion	0.101	0.908	0.087	0.924
DICEPTION	0.075	0.945	0.072	0.939
Ours	<u>0.048</u>	<u>0.981</u>	0.056	**0.984**

Table 5: Quantitative comparison of depth estimation on KITTI and NYUv2 datasets.

Method	Acc@1 (%)	Flow Steps
DeiT-S (Target)	79.8	-
Ours	77.3	1
	77.8	2
	77.9	10

Table 6: Ablation study on the number of flow sampling steps. Top-1 accuracy (%) on ImageNet using DiT-S as the student model to simulate DeiT-S.

Ablation Study

Ablation on Flow Sampling We evaluate how the number of flow sampling steps affects classification accuracy using DiT-S (Peebles and Xie 2023) on ImageNet (Deng et al. 2009). As shown in Table 6, increasing the number of steps leads to consistent performance gains.

With a single forward pass (1-Flow), our method achieves 77.3% top-1 accuracy, showing strong mimicry of DeiT-S without explicit classification training. Using two steps (2-Flow), the accuracy improves to 77.8%, narrowing the gap to the baseline (79.8%). With ten steps (10-Flow), we reach 77.9%, demonstrating that iterative refinement better recovers the target representation.

This indicates that more sampling steps enhance feature reconstruction through progressive updates, without requiring fine-tuning or access to training data — highlighting the expressive power of our flow-based framework.

Ablation on Model Capacity To assess our framework's ability to emulate stronger vision transformers, we use a lightweight *DiT-S* to simulate larger models (*DeiT-B*, *DeiT-H*), as shown in Table 7. Despite its simplicity, our method achieves competitive zero-shot accuracy — 79.46% and 83.92% Top-1 for DeiT-B and DeiT-H targets — outperforming existing one-step distillation (OSD).

Notably, without access to training data or teacher gradients, our method aligns well with complex targets using only a single flow step. Fine-tuning further improves performance

Method	Acc@1 (%)	Target Model
DeiT-B	81.8	-
DeiT-H	85.2	-
OSD (Zero-shot)	83.79	DeiT-H
Ours (Zero-shot)	79.46	DeiT-B
Ours (Fine-tuned)	**81.78**	DeiT-B
Ours (Zero-shot)	**83.92**	DeiT-H

Table 7: Comparison of zero-shot and fine-tuned classification performance (Top-1 Accuracy %) on ImageNet. DiT-S is used to simulate increasingly powerful targets.

to 81.78% on DeiT-B. This demonstrates that our approach enables effective knowledge, even with a smaller backbone.

Visualizations and Analysis

Latent Similarity and Variance To better understand the relationship between generated and target latents, we analyze their feature similarity (measured by cosine similarity) and distribution variance across samples. As shown in Fig. 3, higher similarity between *Flow Z* (generated latent) and *Flow B* (target latent) generally correlates with improved zero-shot performance.

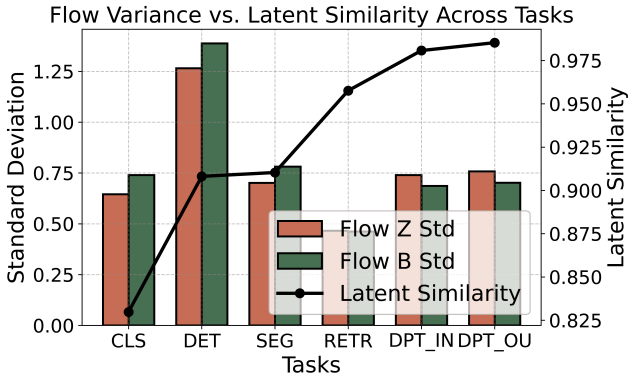


Figure 3: Similarity (line plot) and variance (bar plot) of generated (Flow z) and target (Flow b) latents across different models. Each bar represents the average standard deviation of features per dimension; the line denotes cosine similarity between Flow z and Flow b.

Interestingly, we observe that models with lower latent variance often exhibit worse zero-shot accuracy, even when similarity is high. We hypothesize that flow-matching favors global feature consistency over preserving rare or extreme patterns, resulting in more compact latent distributions. While this may limit expressiveness in the zero-shot setting, it encourages robust feature learning that adapts well under fine-tuning — consistent with our observation of improved performance after adaptation.

Feature Evolution Analysis We analyze feature evolution using PCA visualization as shown in Fig. 4. The generative

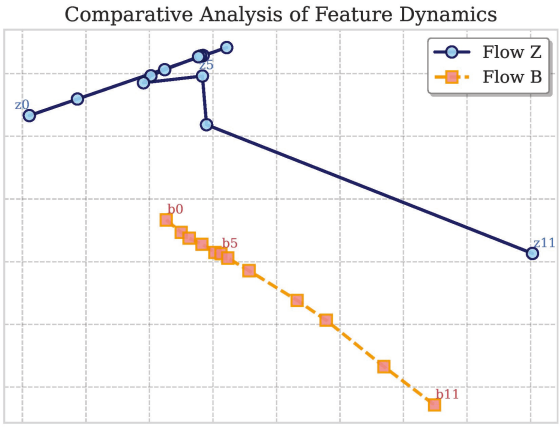


Figure 4: Comparative Analysis of Feature Evolution Dynamics. Flow Z demonstrates an exploratory strategy characterized by early-stage feature exploration followed by rapid convergence. Flow B exhibits a progressive refinement pattern with incremental adjustments toward the target.

Flow Z exhibits a two-phase pattern: initial exploration of the feature space through iterations, followed by sharp convergence after locating the target region. In contrast, the supervised Flow B maintains steady incremental adjustments throughout, resulting in smoother but slower progress. The distinct trajectories reveal fundamental differences in optimization strategies—generative methods prioritize global exploration while supervised approaches focus on local refinement. These observations highlight the strengths of different learning paradigms in representation learning.

Discussion and Conclusion

We present a universal visual perception generating framework based on flow matching, enabling a single framework to generate task-specific representations across diverse vision tasks. Our method achieves strong performance in both zero-shot and fine-tuned settings on classification, detection, segmentation, depth estimation, and image-text retrieval — demonstrating the effectiveness and versatility of the proposed flow-matching paradigm for general-purpose visual representation.

Despite these promising results, our current approach relies on task-specific encoders for supervision during training, and scaling to a large number of tasks remains nontrivial. In future work, we aim to explore fully end-to-end training to reduce dependency on external models, extend our framework to open-world and continual learning scenarios, and investigate universal modeling across modalities beyond vision. In conclusion, our work advances the pursuit of general-purpose vision models and provides a foundation for future research in scalable and universal visual perception systems.

Acknowledgments

This work was supported in part by the 6G Science and Technology Innovation and Future Industry Cultivation Spe-

cial Project of Shanghai Municipal Science and Technology Commission under Grant 24DP1501001, in part by the National High Quality Program under Grant TC220H07D.

References

Bachmann, R.; Kar, O. F.; Mizrahi, D.; Garjani, A.; Gao, M.; Griffiths, D.; Hu, J.; Dehghan, A.; and Zamir, A. 2024. 4M-21: An Any-to-Any Vision Model for Tens of Tasks and Modalities. *arXiv preprint arXiv:2406.09406*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bochkovskii, A.; Delaunoy, A.; Germain, H.; Santos, M.; Zhou, Y.; Richter, S. R.; and Koltun, V. 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eftekhar, A.; Sax, A.; Malik, J.; and Zamir, A. 2021. 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*, 10786–10796.

Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, 241–258. Springer.

Gao, Y.; Chen, K.; Peng, Z.; Lu, H.; and Xu, S. 2025. Knowledge Transfer from Interaction Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3585–3595.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Gui, M.; Schusterbauer, J.; Prestel, U.; Ma, P.; Kotovenko, D.; Grebenkova, O.; Baumann, S. A.; Hu, V. T.; and Ommer, B. 2024. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022a. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022b. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, M.; Yin, W.; Zhang, C.; Cai, Z.; Long, X.; Chen, H.; Wang, K.; Yu, G.; Shen, C.; and Shen, S. 2024. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-shot Metric Depth and Surface Normal Estimation. *arXiv preprint arXiv:2404.15506*.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9492–9502.

Khanna, S.; Irgau, M.; Lobell, D. B.; and Ermon, S. 2024. ExPLoRA: Parameter-Efficient Extended Pre-Training to Adapt Vision Transformers under Domain Shifts. *arXiv preprint arXiv:2406.10973*.

Le, D. H.; Pham, T.; Lee, S.; Clark, C.; Kembhavi, A.; Mandt, S.; Krishna, R.; and Lu, J. 2024. One Diffusion to Generate Them All. arXiv:2411.16318.

Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; and Yu, F. 2024. Matching Anything by Segmenting Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*.

Mizrahi, D.; Bachmann, R.; Kar, O.; Yeo, T.; Gao, M.; Dehghan, A.; and Zamir, A. 2023. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36: 58363–58408.

Nathan Silberman, P. K., Derek Hoiem; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rajič, F.; Ke, L.; Tai, Y.-W.; Tang, C.-K.; Danelljan, M.; and Yu, F. 2023. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1137–1149.

Ren, T.; Chen, Y.; Jiang, Q.; Zeng, Z.; Xiong, Y.; Liu, W.; Ma, Z.; Shen, J.; Gao, Y.; Jiang, X.; et al. 2024a. DINO-X: A Unified Vision Model for Open-World Object Detection and Understanding. *arXiv preprint arXiv:2411.14347*.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024b. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6830–6839.

Wang, X.; Ye, F.; and Zhang, Y. 2024. Task-Aware Low-Rank Adaptation of Segment Anything Model. *arXiv preprint arXiv:2403.10971*.

Xu, G.; Ge, Y.; Liu, M.; Fan, C.; Xie, K.; Zhao, Z.; Chen, H.; and Shen, C. 2024. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024c. Depth Anything V2. *arXiv preprint arXiv:2406.09414*.

Yin, W.; Wang, X.; Shen, C.; Liu, Y.; Tian, Z.; Xu, S.; Sun, C.; and Renyin, D. 2020. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*.

Yin, W.; Zhang, J.; Wang, O.; Niklaus, S.; Mai, L.; Chen, S.; and Shen, C. 2021. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 204–213.

Yu, Q.; Zhao, X.; Pang, Y.; Zhang, L.; and Lu, H. 2024. Multi-view Aggregation Network for Dichotomous Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3921–3930.

Zhao, C.; Liu, M.; Zheng, H.; Zhu, M.; Zhao, Z.; Chen, H.; He, T.; and Shen, C. 2025. Diception: A generalist diffusion model for visual perceptual tasks. *arXiv preprint arXiv:2502.17157*.

Zhong, Z.; Tang, Z.; He, T.; Fang, H.; and Yuan, C. 2024. Convolution meets lora: Parameter efficient finetuning for segment anything model. *arXiv preprint arXiv:2401.17868*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhu, M.; Liu, Y.; Luo, Z.; Jing, C.; Chen, H.; Xu, G.; Wang, X.; and Shen, C. 2024. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *arXiv preprint arXiv:2410.02369*.