Probing and Bridging Geometry–Interaction Cues for Affordance Reasoning in Vision Foundation Models

Supplementary Material

1. Additional Experimental Details

Visual Foundation Models (VFMs) have been widely adopted as powerful visual representations [12, 14]. While their performance across various vision tasks is remarkable, few studies have systematically explored their intrinsic capabilities for affordance. Based on the intuition that affordance requires both structural recognition and actionable context, our research hypothesizes that its core capabilities arise from the interplay of geometric perception and interaction perception. Using VFMs as an analytical lens, we first investigate the relationship and emergence of these two capabilities and then conduct simple fusion experiments to validate their composability as the core capacities of affordance.

The main paper provides a concise summary of the experimental setup. Due to space constraints, this supplementary material presents the detailed experimental configurations, specific model checkpoints, and extended results. Given our primary focus on VFMs, the following sections first introduce the specific models used in our main experiments and then explain the execution details for our probing and fusion experiments.

1.1. Visual Foundation Models

For our analysis, we conduct multi-dimensional quantitative and qualitative investigations into both discriminative and generative VFMs to characterize how affordance capabilities are internally represented. We categorize the models into four representative groups based on their training methodologies: **self-supervised models** (DINO/DINOv2 [2, 12]), **vision-language alignment models** (CLIP [14]/SigLiP [17]), **segmentation models** (SAM [8]), and **generative models** (Stable Diffusion [13]/Flux [1]). Importantly, rather than comprehensively benchmarking performance, our approach aims to leverage their inherent representations and visualization capabilities to explore the fundamental components that constitute affordance understanding.

DINO Series (DINO, DINOv2, DINOv3). The **DINO series** models are based on a **self-distillation approach** and utilize the **Vision Transformer (ViT) architecture** [2, 12]. Specifically, DINO (ViT-B/16, 12 layers) and DINOv2 (ViT-B/14, 12 layers) employ standard Transformer depths, while DINOv3 (ViT-B/16) expands to 40 layers [15]. We use the official checkpoints of each model: DINO is trained on **ImageNet-1k**, DINOv2 on the **LVD-142M** dataset, and DINOv3 on the larger **LVD-1689M** dataset.

CLIP, SigLIP. We adopt **CLIP** [14] and **SigLIP** [17] models as representatives of vision-language alignment capabilities. Both are based on contrastive learning frameworks and use the Vision Transformer (ViT) architecture. Specifically, CLIP (ViT-B/16, 12 layers) and SigLIP (ViT-B/16, 12 layers) employ standard Transformer depths. We use the openclip[7] checkpoints for CLIP, which is trained on the large-scale **LAION-400M** dataset, while SigLIP, an offical version, utilizes a similar scale of web image-text pairs but optimizes training efficiency and performance through a novel **sigmoid loss function**.

SAM (Segment Anything Model). We employ the **Segment Anything Model (SAM)** [8] as a representative of general segmentation capability. This model is based on a promptable segmentation framework and utilizes the Vision Transformer (ViT) architecture. Specifically, SAM (ViT-B/16, 12 layers) employs a standard Transformer depth. We use the official checkpoint, which is trained on the **SA-1B** dataset, containing 11 million images and 1 billion masks, endowing it with strong zero-shot segmentation capabilities.

Generative Models (Stable Diffusion, Flux). We adopt **Stable Diffusion (SD)** [13] and **FLUX** [1] as representatives of generative visual models. Both are built upon diffusion model frameworks but employ distinct core architectures: SD primarily utilizes a **U-Net** as the backbone, whereas FLUX adopts the more scalable **DiT (Diffusion Transformer) architecture** [5]. We use the official checkpoints of each model: Stable Diffusion base 2.1 is trained on the large-scale **LAION dataset**, while FLUX.1-dev is trained on its own non-public dataset.

1.2. Implementation Details

This section outlines the technical implementation of our experimental framework. We detail the standard protocols for image preprocessing and feature extraction across diverse Visual Foundation Models (VFMs). Subsequently, we specify the exact configurations for the two primary evaluation pipelines: supervised linear probing for affordance segmentation and zero-shot geometry-interaction fusion for affordance grounding, providing all necessary parameters to ensure the reproducibility of our results.

1.2.1. Feature Extraction Protocol

Image Preprocessing. To preserve the original image composition and structural integrity, we maintained the native resolution of each image. Input images were padded to the nearest integer multiple of the model's patch size to satisfy the architectural constraints of each Vision Transformer (ViT). Subsequently, all images were normalized using the specific mean and standard deviation parameters defined by their respective pre-training recipes.

Feature Extraction. For discriminative models based on Vision Transformers, we adopted a multi-layer feature extraction strategy consistent with the **Probing3D [4]** protocol. For standard 12-layer ViTs (including DINO series, CLIP, SigLIP, and SAM), features were extracted from layers 3, 6, 9, and 12. For the deeper 40-layer DINOv3, features were extracted from layers 10, 20, 30, and 40. This stratified approach captures visual representations across varying levels of abstraction, ranging from local geometric parts to global semantic objects.

For generative models, we adopted architecture-specific extraction methods. For **Stable Diffusion**, we analyzed feature maps from four key blocks within the U-Net decoder to probe spatial representations at different scales. For **Flux**, to capture verb-conditioned interaction priors, we exclusively extracted the cross-attention maps corresponding to verb tokens. These maps were averaged across **all denoising timesteps** and the entire **Dual-Transformer block** to ensure robust and stable spatial localization.

Feature Usage. In all linear probing experiments for affordance segmentation, feature maps from the four selected layers were channel-wise concatenated to construct a comprehensive multi-scale representation before being fed into the linear head. Conversely, for qualitative visualizations and geometric analyses (e.g., PCA projections, cosine similarity), features were analyzed independently layer-by-layer. This separation allows us to isolate and characterize the distinct properties specific to each architectural depth.

1.2.2. Linear Probing On UMD Setup

This section details the setup for the linear probing experiments conducted on the UMD dataset [3, 11] to evaluate the geometric awareness of features for affordance segmentation. The UMD dataset provides dense, pixel-level annotations for 7 affordance categories (e.g., *grasp, cut, support*) across a variety of tool and object categories, making it a standard benchmark for the structural perception of affordances. It contains approximately 11,800 training and 14,020 testing images.

Probe Architecture. We employed a minimal linear probe head, consisting of a single `BatchNorm2d` layer followed by a 1×1 convolutional layer that directly outputs the per-pixel class logits. This architecture intentionally contains no non-linear activation functions or additional hidden layers, ensuring that the resulting performance directly reflects the linear separability of the frozen pre-trained representations.

Optimization and Training. The probe was optimized using the CrossEntropyLoss (with `ignore_index=255`) and the AdamW optimizer. For the results reported in the main paper, we used a fixed learning rate of $1e{-3}$ and a weight decay of $1e{-2}$. The probe was trained for a maximum of 2 epochs, with validation performed after each epoch. We employed early stopping with a patience of 1 epoch. Gradient norms were clipped to a maximum value of 1.0, and all training operations were conducted using `bfloat16` precision to ensure efficiency.

Evaluation. All UMD dataset images were processed at their native resolution of 640×480. We utilized a DataLoader with a batch size of 4. The mean Intersection-over-Union (mIoU) was computed online during training and served as the primary evaluation metric for model selection and reporting.

1.2.3. PCA Subspace Projection for Visualization

Subspace Definition. We define a geometric feature subspace using a single reference image containing the target object. The object's **Region of Interest (ROI)** is identified utilizing either an external segmentation mask or the image's alpha channel. This ROI is aligned to the model's patch feature grid via letterboxing and resizing. The corresponding ROI tokens are selected through thresholding and optional morphological dilation; in the absence of a provided mask, the entire reference image is treated as the ROI.

PCA Fitting. The feature tokens extracted from the reference ROI are centered by subtracting their mean vector μ. We then fit a **3D Principal Component Analysis (PCA)** solely on these centered ROI tokens using randomized Singular Value Decomposition (SVD). We utilize 3 components and 5 iterations with a fixed random seed to ensure reproducibility. This process yields the principal components (a $C \times 3$ matrix) and eigenvalues, which, together with μ, constitute the saved subspace model.

Projection and Coloring. To visualize a target scene, its feature tokens are centered using the reference mean μ and projected into the learned 3D subspace via right-multiplication with the principal components. To ensure consistent visualization contrast, we compute the 1st and 99th percentile thresholds for each component based on the projection scores of the *reference* ROI. These fixed thresholds are applied to clip and linearly normalize the target scene's projection scores to the range $[0, 1]$. The normalized values for the first three principal components are mapped directly to the RGB channels. Finally, the resulting low-

resolution feature map is bilinearly upsampled to the target image resolution for visualization.

Output. The process generates pseudo-color images showing the geometric similarity between the target scene and the reference object, as well as individual heatmaps for each principal component. All visualizations use fixed color mappings and normalization based on the reference ROI to ensure consistent cross-scene comparisons. The subspace can be redefined for different reference objects by repeating this procedure with new reference images.

1.2.4. Cross-Attention Map Setup on AGD20K

This section details the procedure for leveraging generative models to achieve zero-shot affordance estimation on the **AGD20K dataset** [10]. AGD20K is specifically designed for cross-view affordance grounding, comprising over 20,000 exocentric human-object interaction images and 3,700 egocentric object-only images across 36 affordance categories. The ground-truth annotations are provided as probability heatmaps derived from point-level supervision, which naturally align with our objective of predicting spatial likelihoods for verb-based interactions. We evaluate our method under the challenging **'Unseen' setting**, where the object categories in the test set are disjoint from those in the training set.

Interaction Prior Extraction. To obtain verb-conditioned spatial priors, we utilize the official **FLUX.1-dev** and **FLUX.1-Kontext** checkpoints [1]. For a given affordance category and object image, we construct a generative editing prompt following a structured triplet template: `"add [agent part] [affordance label] [object name]"`. Here, the `[affordance label]` and `[object name]` are derived directly from the dataset annotations. The `[agent part]` is inferred based on commonsense knowledge of the interaction (e.g., primarily `hands`, or occasionally `mouth/foot`). We extract the cross-attention maps associated specifically with the `[affordance label]` token (e.g., "hold"). To ensure robustness, these maps are averaged across **all denoising timesteps** and **all Dual-Transformer blocks** within the FLUX architecture, producing a single, stable 2D interaction prior.

Spatial Alignment of Interaction Priors. Since the generative editing process involves latent space decoding that may introduce minor spatial shifts, we employ a rapid affine registration strategy to project the extracted interaction heatmap back onto the original image coordinates. We extract sparse keypoints using the efficient **ORB** detector and establish feature correspondences between the original and generated images via a Brute-Force Matcher with Lowe's ratio test. A robust 2D affine transformation matrix is then estimated using **RANSAC** to account for translation and scaling differences, enabling the precise inverse warping of the attention map. This lightweight alignment pipeline is computationally negligible, executing in **under one second**, ensuring that the interaction cues are strictly synchronized with the input geometry without bottlenecking the inference flow.

Geometry-Interaction Fusion. The fusion process integrates the interaction prior from FLUX with geometric structural cues from **DINOv3** [15]. First, the object's Region of Interest (ROI) is roughly localized by applying a percentile threshold to the object token's attention map from FLUX Kontext. This ROI is mapped to the DINOv3 token grid to extract relevant patch features from the model's final layer. We apply **Principal Component Analysis (PCA)** to these ROI-specific features, retaining the top-5 components to obtain a set of compact, part-level geometric bases. Each PCA component is resized to the original image resolution and undergoes Gaussian smoothing.

To select the geometric component most aligned with the interaction, we compute the **Normalized Scanpath Saliency (NSS)** score between each geometric energy map and the verb attention map. The component with the highest NSS score is selected, representing the strongest spatial consensus between geometric structure and interaction intent.

The final affordance probability map is derived through a **soft fusion** mechanism in the probability domain. The verb attention map is processed through a softmax function with temperature $T = 0.5$ to form a probability distribution P_v. The selected geometric map is normalized to $[0, 1]$, raised to a power $\gamma = 0.7$, and similarly processed to form P_g. The log-probabilities are combined using a weighting factor $\lambda = 0.65$:

$$\log F = \lambda \log P_v + (1 - \lambda) \log P_g$$

The output F is exponentiated and renormalized to produce the final continuous affordance probability map.

Evaluation. The fused output F serves as the zero-shot prediction of the affordance distribution. To comply with the evaluation protocols of AGD20K, F is **normalized to sum to 1**. This probability map is evaluated against the ground-truth heatmaps using the standard metrics of Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS).

1.2.5. Computational Resources and Runtime

All experiments were conducted on a high-performance computing workstation equipped with two **NVIDIA RTX 6000 Ada Generation GPUs** (each with 49 GB VRAM), an **Intel Xeon w7-3445 CPU** (20 physical cores), and **250 GiB of system memory**.

The **linear probing** experiments on the UMD dataset, encompassing training and evaluation for a single model

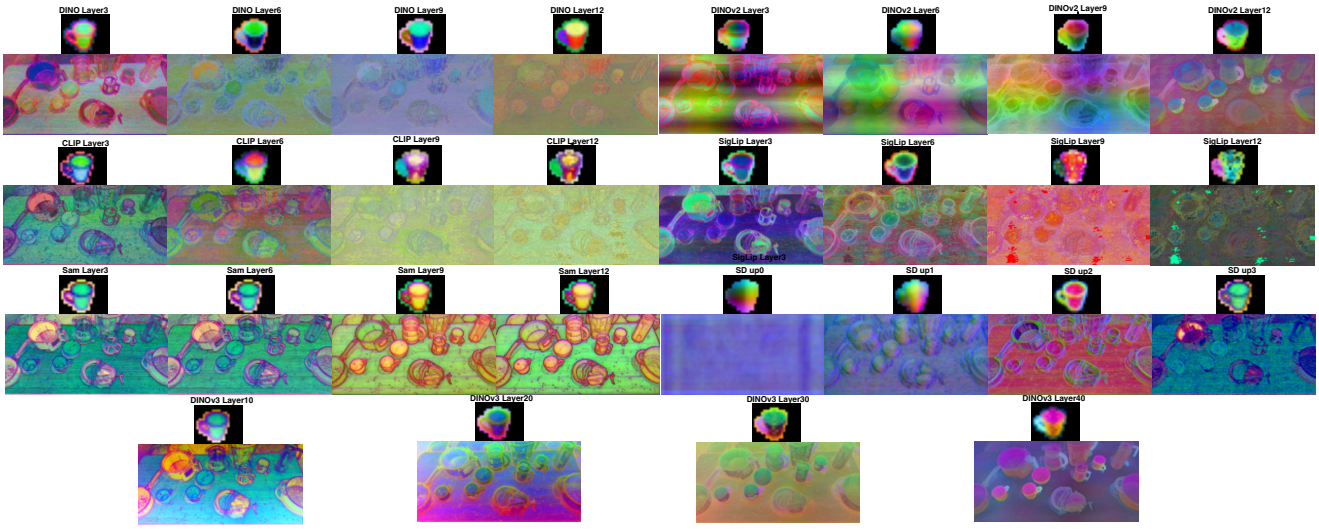

Figure 1. **PCA visualization of seven probed models at different levels of depth**

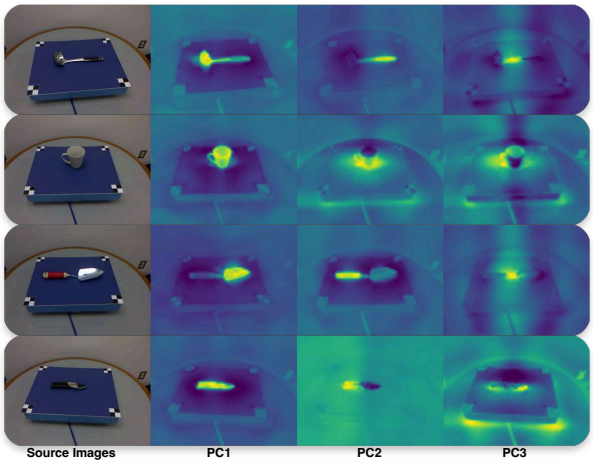

Figure 2. **Stable part-level decomposition in DINO representations.** We visualize the first three principal components (PC1–PC3) derived from PCA applied to DINO feature embeddings on UMD objects. Across diverse objects such as tools and containers, the leading components consistently highlight meaningful structural regions, including handles, functional heads, and graspable parts. This consistency suggests that DINO organizes object representations into stable part-level structures, providing a geometric basis for affordance understanding.

configuration, required approximately **20 to 40 minutes** to complete depending on the backbone size. The most computationally intensive component was the **interaction prior extraction** from generative models. Generating the verb-conditioned cross-attention maps for a single image-verb pair using FLUX required an average of **1.5 to 2 minutes**. In contrast, the subsequent **geometry-interaction fusion** pipeline demonstrated high efficiency, processing a single sample in approximately **2 to 3 seconds**.

Model	Depth RMSE [m]	Normal RMSE [°]
DINO-ViT-B16	0.5071	28.35
DINOv2-ViT-B16	0.3307	22.41
CLIP-ViT-B16	0.9351	34.68
SigLIP-ViT-B16	0.7187	34.96
SAM-ViT-B16	0.5665	26.89
Stable Diffusion-Unet	0.4801	24.68

Table 1. Depth and surface normal RMSE on NYU (lower is better). Units: depth in meters [m], normals in degrees [°]. Data from Proing 3D [4].

2. Additional Results

This section presents comprehensive empirical evidence to further validate our dual-dimensional framework. We organize the findings into two primary parts mirroring our main analysis: first, we deepen the investigation of **geometric perception** by establishing a statistical correlation between intrinsic 3D awareness and affordance capabilities, supported by extended visualizations of internal representations across VFMs. Second, we examine the **interaction perception** dimension, providing a comparative analysis of generative models to demonstrate the superiority and robustness of the interaction priors extracted from FLUX.

2.1. Detailed Linear Probing Results on Geometry

This section quantitatively substantiates our hypothesis that geometric perception forms the foundational substrate of affordance understanding. We establish a mechanistic link by

Figure 3. **Comparison among SD2.1, SDXL, SD3.5, Flux.1-dev.** Prompt: "A person is grasping a knife". SD models lack clear hand details, while Flux captures realistic hand–object interaction.

Figure 4. **Flux encodes interaction priors.** Generated samples show strong understanding of human–object relations.

correlating the intrinsic 3D awareness of Visual Foundation Models (VFMs) [12, 14, 17], measured by depth and normal estimation errors on NYU [4] (Table 1), with their linear probing performance on UMD affordance segmentation (Table 3) [11].

PCA Component Analysis. To further examine how geometric structure emerges in VFM representations, we visualize the principal components derived from PCA on DINO object embeddings (Fig. 2). Across multiple objects, the leading components consistently correspond to meaningful structural regions such as handles, blades, and other graspable parts. Notably, these components remain stable across different object instances and categories, suggesting that DINO organizes visual features into reusable part-level structures rather than object-specific semantics. This observation provides qualitative evidence that DINO's representations encode intrinsic geometric primitives that naturally align with actionable object regions.

Compensatory Effect of Explicit 3D Cues. The impact of explicit 3D augmentation provides further mechanistic insight. As detailed in Table 3, we observe that external depth and normal cues serve as powerful compensatory signals for geometrically weaker models—CLIP and SigLIP see substantial gains when augmented (e.g., CLIP+Normal boosts mIoU from 0.520 to 0.581). In stark contrast, DINOv2 shows negligible improvement or even slight regression, suggesting its representations have already reached a saturation point for structural information. This indicates that explicit 3D cues are most beneficial for models whose internal representations lack a strong geometric prior.

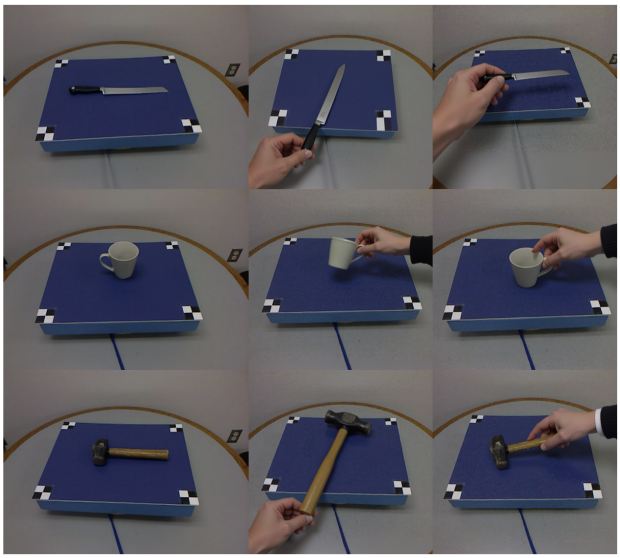

Figure 5. **Flux-Kontext editing.** Editing UMD images by adding grasping hands produces natural interactions and aligned attention.

Category-Specific Dependency. Our analysis also reveals that this geometric dependency is task-specific. Affordances requiring precise spatial localization and interaction with specific parts, such as *cut* or *pound*, benefit significantly more from explicit geometry (e.g., CLIP's IoU on *cut* improves from 0.529 to 0.626 with normal augmentation). Conversely, shape-semantic categories like *contain*, which rely more on global object recognition, are already well-localized by baseline models and gain less from explicit geometric cues.

2.2. Extended Exploration of Geometric Representations

This section presents an extended qualitative analysis of geometric representations across seven Visual Foundation Models, expanding upon the findings in Section 3.1 of the main paper. Utilizing the PCA subspace projection method described in Section 1.4, we visualize the internal representations of DINO series [2, 12, 15], SAM [8], CLIP [14], SigLIP [17], and Stable Diffusion [13] across four distinct network depths. These visualizations, shown in Figure 1, reveal how geometric structures are encoded and transformed as information propagates through the layers.

Our observations confirm a fundamental divergence in representation strategies. The **DINO series** demonstrates remarkable consistency in part-level decomposition, with deeper layers cleanly isolating functional components (e.g., handles, rims). **SAM** maintains a strong focus on object

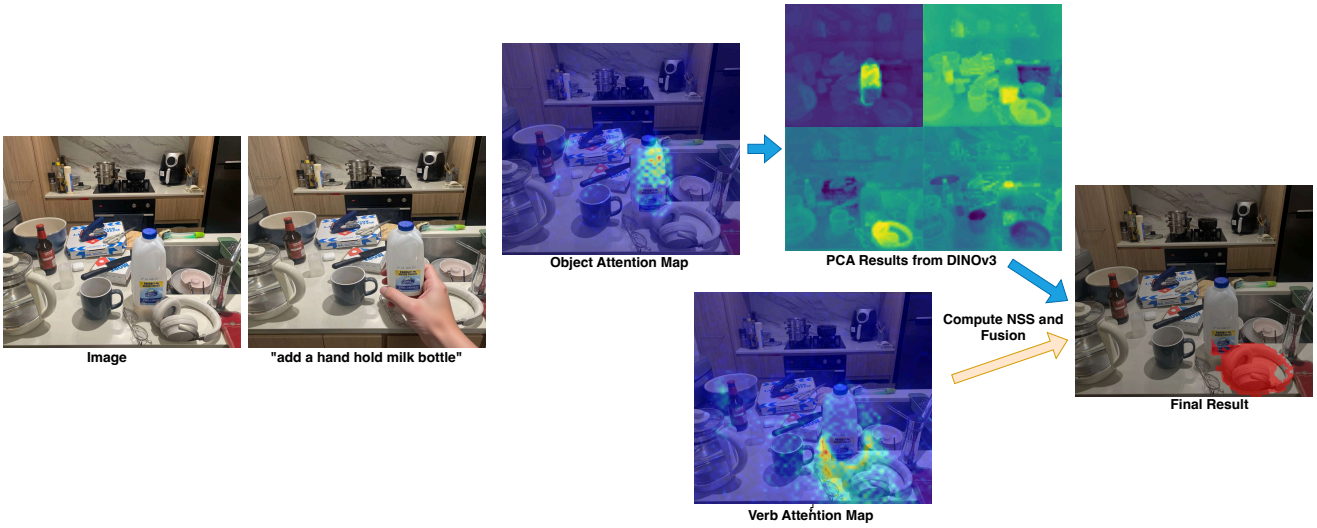

Figure 6. **Failure Case: ROI Contamination in Complex Scenes.** In cluttered scenes, the Flux object attention map is overly coarse, bleeding into adjacent objects (kettle, cups). The contaminated ROI corrupts the DINOv3 PCA geometric features, leading to inaccurate NSS-based selection and fusion failure.

boundaries and holistic masks throughout all layers, reflecting its segmentation-specific training objectives. In contrast, vision-language models like **CLIP** and **SigLIP** exhibit a marked transition: while mid-level layers capture some geometric features, the final layers collapse into semantically clustered representations where categorical information dominates over structural details. **Stable Diffusion 2.1** reveals a similar pattern, processing surface-like geometric cues in intermediate layers that largely dissipate in the deeper, more abstract stages of the U-Net.

2.3. Comparison of Interaction Priors Across Generative Models

Model Selection through Comparative Analysis. We perform a comparative analysis of leading generative models—Stable Diffusion 2.1, SDXL [13], SD3.5 [5], and **FLUX.1-dev** [1]—using the prompt *"A person is grasping a knife"*. As visualized in Figure 3, while previous iterations of diffusion models frequently generate anatomically implausible interactions (e.g., hands grasping the blade or exhibiting biomechanically invalid poses), FLUX.1-dev consistently produces physically coherent grasps aligned with the knife's handle. This superior performance in generating spatially correct interactions establishes FLUX as our primary source of interaction cues.

Interaction Generation Capability Validation. We further examine FLUX's capacity to generate diverse human-object interactions beyond the initial knife-grasping scenario. As shown in Figure 4, the model demonstrates a remarkable understanding of various interaction types, consistently producing physically plausible relations across a wide range of object categories and action verbs. This robustness confirms that FLUX embeds comprehensive knowledge of how humans engage with objects in three-dimensional space, validating its suitability for extracting generalized interaction priors.

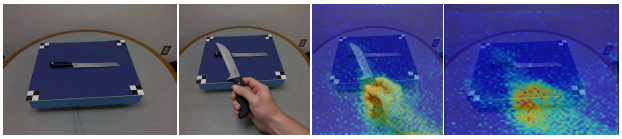

Figure 7. **Failure Case: Generative Object Duplication.** The editing model generates a duplicate object (second knife), causing the interaction prior to focus on the hallucinated instance. This results in critical spatial misalignment and failure to project the affordance cue onto the original object.

Practical Validation through Image Editing. We validate the functional utility of these interaction priors using **FLUX-Kontext** [1] for practical image editing tasks. When editing images from the **UMD dataset** [11] by adding grasping hands to various objects (Figure 5), the model generates natural interactions with attention patterns precisely aligned to affordance-bearing regions. This editing capability demonstrates that the extracted interaction priors are not merely analytical artifacts but functionally meaningful guides for spatial reasoning in real-world applications.

2.4. Generalization Across Geometric Backbones

To verify that the proposed framework does not depend on a specific geometric backbone, we replace DINOv3 with

Method	KLD ↓	SIM ↑	NSS ↑
DINOv3 + Flux	1.493	0.326	1.090
DINOv2 + Flux	1.495	0.322	1.105

Table 2. **Generalization across geometric backbones.** Replacing DINOv3 with DINOv2 yields nearly identical performance when fused with Flux.

DINOv2 while keeping the interaction prior extracted from Flux unchanged. Table 2 shows that the resulting performance remains nearly identical across all evaluation metrics. This observation indicates that the effectiveness of the proposed geometry–interaction fusion arises from part-centric geometric representations rather than a particular model checkpoint.

3. Limitation

Generative Object Duplication. One limitation of the current generative editing framework is the potential for semantic drift or object duplication during the synthesis process. As illustrated in Figure 7, when prompted to modify the scene (e.g., "add a hand holding the knife"), the model occasionally generates a duplicate instance of the target object rather than synthesizing an interaction with the existing one.

This results in a critical spatial misalignment: the verb-conditioned interaction attention map correctly localizes the "holding" action, but it does so on the *newly generated* object. Since our pipeline projects this attention map back onto the original image's coordinate system, the resulting affordance prediction fails to align with the original object. This issue highlights the challenge of maintaining strict structural consistency in generative editing. Future work could mitigate this by employing **negative prompting** (e.g., "no extra objects," "duplicate") or advanced attention-injection techniques [6, 16] to rigidly anchor the generation to the original object's layout.

ROI Contamination in Complex Scenes. In complex, cluttered environments, relying solely on Flux's object attention map for Region of Interest (ROI) extraction can be insufficient. As shown in Figure 6, the attention map for the target object (e.g., "milk bottle") is relatively coarse and often exhibits spatial leakage, bleeding into adjacent objects such as the kettle and cups.

Consequently, when this noisy ROI is used to mask DINOv3 features, the subsequent PCA dimensionality reduction incorporates geometric cues from unrelated objects. In such high-entropy scenarios, the simple Normalized Scanpath Saliency (NSS) metric struggles to differentiate the true functional part of the target object from the geometric noise of the background. This suggests that while attention maps are effective for isolated objects, complex scenes require tighter spatial grounding. Integrating explicit segmentation models, such as the **Segment Anything Model (SAM)** [8], or leveraging more precise VLM grounding [9] to extract cleaner ROIs would likely improve the robustness of geometric decomposition in crowded scenes.

Fusion Hyperparameter Sensitivity. While our primary objective is to demonstrate the *composability* of geometry and interaction cues using the simplest possible framework, we acknowledge that our fusion stage employs several hyper-parameters, including the Gaussian smoothing factor, the geometric power γ, and the probability weighting factor λ. The current quantitative analysis does not include an extensive hyper-parameter search or a comprehensive ablation study for these settings. Instead, the final parameters (e.g., $\lambda = 0.65, \gamma = 0.7$) were fixed during the qualitative development phase based on empirical stability. This approach, while prioritizing simplicity over exhaustive optimization, intentionally maintains the focus of our objective: to extract and combine the most potent geometric and interaction representations from leading VFMs [1, 15] using a straightforward protocol. Consequently, our results should be interpreted as a strong **proof-of-concept** for the dual-dimensional framework, rather than a claim for an optimally tuned fusion strategy.

References

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1, 3, 6, 7

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 5

[3] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 2

[4] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 2, 4, 5

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 6

Table 3. Affordance segmentation results (IoU per affordance and mean IoU). Gray rows denote base models, white rows denote variants with additional depth/normal cues. All results are aggregated across multiple layers.

Model	grasp	cut	scoop	contain	pound	support	wrap-grasp	mIoU
DINO-ViT-B16	0.365	0.505	0.290	0.801	0.300	0.467	0.610	0.477
+Depth	0.387	0.643	0.270	0.818	0.506	0.370	0.742	0.534↑
+Normal	0.397	0.640	0.327	0.799	0.540	0.449	0.689	0.549↑
+Both	0.386	0.579	0.283	0.793	0.485	0.428	0.671	0.518↑
DINOv2-ViT-B14	0.513	0.608	0.526	0.808	0.766	0.675	0.792	0.670
+Depth	0.503	0.729	0.465	0.808	0.747	0.600	0.784	0.662
+Normal	0.462	0.728	0.439	0.788	0.767	0.592	0.779	0.651
+Both	0.446	0.715	0.465	0.785	0.754	0.639	0.781	0.655
CLIP-ViT-B16	0.403	0.529	0.290	0.756	0.542	0.543	0.574	0.520
+Depth	0.418	0.621	0.364	0.770	0.533	0.435	0.603	0.535↑
+Normal	0.437	0.626	0.325	0.790	0.648	0.538	0.701	0.581↑
+Both	0.313	0.627	0.396	0.780	0.651	0.545	0.700	0.573↑
SigLIP-ViT-B16	0.295	0.399	0.385	0.736	0.521	0.586	0.699	0.517
+Depth	0.356	0.563	0.419	0.729	0.507	0.560	0.660	0.542↑
+Normal	0.359	0.561	0.416	0.712	0.584	0.593	0.663	0.556↑
+Both	0.360	0.571	0.373	0.718	0.526	0.551	0.717	0.545↑
SAM-ViT-B16	0.398	0.632	0.345	0.790	0.475	0.452	0.729	0.546
+Depth	0.421	0.657	0.357	0.766	0.413	0.598	0.594	0.544
+Normal	0.402	0.663	0.356	0.768	0.373	0.534	0.722	0.545
+Both	0.447	0.655	0.375	0.771	0.480	0.591	0.708	0.575↑
SD2.1-Unet	0.443	0.591	0.372	0.797	0.632	0.478	0.781	0.585
+Depth	0.370	0.566	0.337	0.788	0.645	0.483	0.745	0.562
+Normal	0.483	0.636	0.340	0.805	0.632	0.503	0.776	0.596↑
+Both	0.428	0.576	0.367	0.793	0.594	0.477	0.790	0.575

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 7

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 5, 7

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 7

[10] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 3

[11] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1374–1381. IEEE, 2015. 2, 5, 6

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 5

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 5, 6

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 5

[15] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 3, 5, 7

[16] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven

image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023. 7

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 5