2D-LFM: Lifting Foundation Model without 3D Supervision

Mosam Dabhi[1*] Irhas Gill[2*] László A. Jeni[1] Simon Lucey[2]

[1]Carnegie Mellon University [2]Adelaide University

2dlfm.github.io

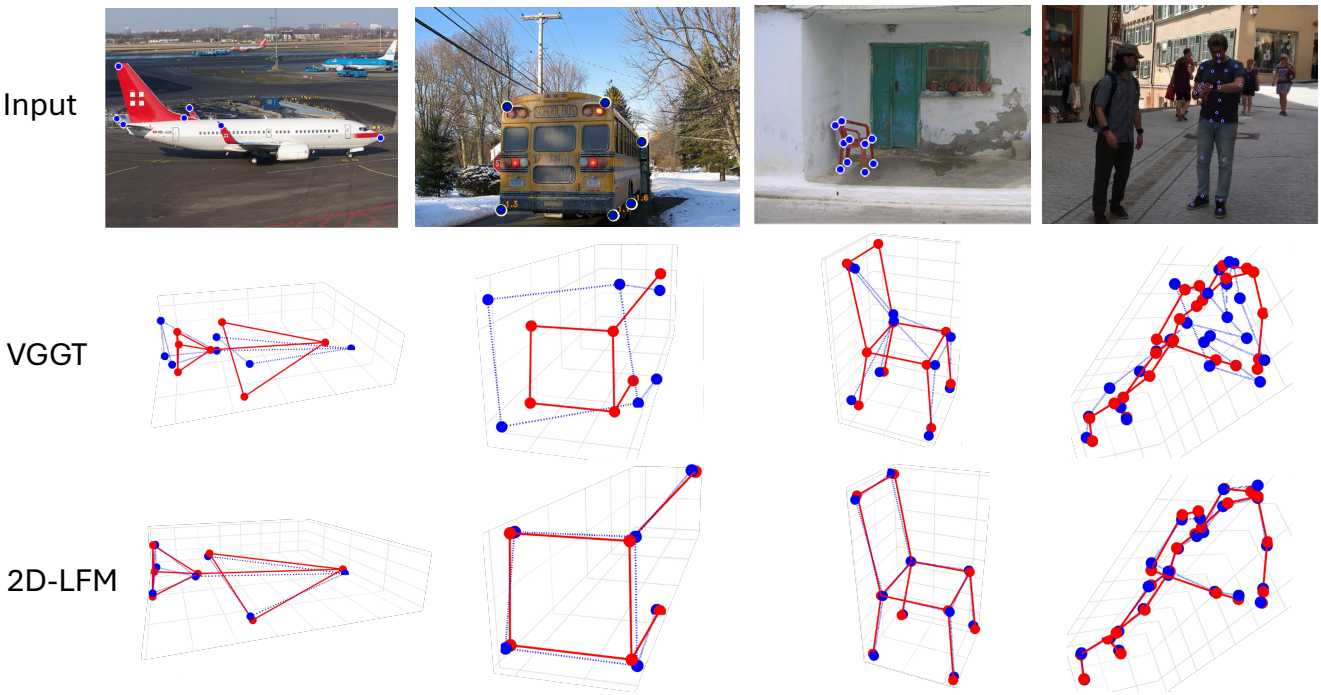

Figure 1. **2D landmark lifting without 3D supervision matches supervised methods**. Given 2D keypoints from single images (top), we compare against VGGT [19] which back-projects through predicted depth (middle) versus our 2D-LFM trained with 2D supervision only (bottom). VGGT produces distorted geometry despite accurate scene depth ($>$ 100mm MPJPE), while 2D-LFM recovers correct object structure (8.1mm). Ground truth: red, Predictions: blue.

Abstract

Recent vision foundation models give the impression that 3D reconstruction from RGB is largely solved. Yet these systems struggle with object-specific 3D structure: the fine-grained geometry implied by an object's landmarks or skeleton. In this paper, we show that when a model is given only 2D landmarks, it can recover more accurate 3D structure than state-of-the-art depth-from-RGB foundation models. Classical lifting approaches such as PAUL demonstrate this principle but do not scale beyond single categories, while methods like 3D-LFM scale but require extensive 3D supervision. We present the first lifting foundation model that learns object-specific 3D geometry using only 2D supervision. The key idea is to inject correspondence structure into the model via a positional encoding inspired by classical structure-from-motion. This simple inductive bias enables robust, object-agnostic 3D lifting that rivals or exceeds recent 3D-supervised approaches, revealing that landmark-based lifting remains a powerful and under-exploited paradigm for 3D understanding.

*Equal contribution.

1. Introduction

Reconstructing 3D structure from 2D observations has been a central challenge in computer vision since its inception. Recent progress in scene-centric monocular depth estimation - most notably models such as VGGT [19], Depth Anything [24], and MiDaS [1] has strengthened the belief that 3D vision from a single RGB image is largely solved. These models produce accurate dense depth maps across diverse natural images, and have become widely adopted as generic 3D perception backbones.

We show that this progress does not extend to object-specific 3D structure. Figure 1 highlights a consistent and surprising failure mode: when provided with 2D semantic landmarks from a single image and asked to reconstruct object-specific geometry, state-of-the-art depth-from-RGB models produce distorted, implausible, or anatomically inconsistent shapes. Back-projecting landmarks through predicted depth yields collapsed bottle necks, warped limbs, inconsistent cloth folds, and incorrect skeletal proportions across humans, hands, animals, vehicles, and articulated objects. Scene-level depth estimation and object-level structure reconstruction rely on fundamentally different inductive biases; success in one does not imply success in the other.

Before proceeding, we clarify the task examined in this paper. **3D lifting** refers to the reconstruction of a non-rigid object category from *atemporal* 2D observations: a single still image with no access to motion, multi-view geometry, or physical priors. Although an individual instance (e.g., a particular chair) may be rigid, the *category* is non-rigid: its members vary in shape, proportions, and articulation. A lifting model must therefore infer a non-rigid shape prior directly from data, enabling plausible 3D reconstruction from only a sparse 2D landmark configuration. The problem is strictly harder than classical structure-from-motion (SfM), since no temporal coherence or rigidity constraints are available.

More surprisingly, we find that a model trained solely on 2D landmarks from *single* images, without using RGB appearance or temporal cues - can recover more accurate object-specific 3D structure than state-of-the-art depth models applied to the same images. This suggests that, for semantic 3D reconstruction, explicit geometric reasoning from sparse 2D measurements often exceeds the structural fidelity obtainable from dense appearance-based depth prediction.

A rich literature has explored this setting. Early neural approaches, including Deep NRSfM, C3DPO [12], PAUL [17], and related methods demonstrated that the inductive bias of neural networks can serve as a powerful *non-rigid shape prior*, enabling single-view 3D lifting from only 2D supervision. By constraining the space of allowable 3D shapes through low-rank or bottleneck representations, these models recover plausible geometry without relying on motion. However, these architectures scale poorly: their priors depend on manually designed category-specific structures (e.g., fixed joint orderings, symmetry assumptions, or landmark conventions), preventing a single model from generalizing across diverse object types.

Transformer-based lifting models [3, 9] address scalability through *permutation equivariance*, treating 2D landmarks as an unordered set and enabling weight sharing across categories with different landmark definitions. Yet this advantage introduces a fundamental limitation: **permutation equivariance eliminates correspondence**. Without knowing that token i consistently represents "left shoulder" or "right elbow," the network cannot form stable non-rigid priors from 2D observations alone. Existing transformer lifters therefore require *3D supervision* to re-establish token identity, making 3D labels an essential training signal.

Classical SfM offers a conceptual clue. Tomasi–Kanade factorization [15] succeeds with only 2D inputs because it preserves *correspondence* across views, enabling recovery of 3D structure without depth annotations. While our setting is strictly single-frame, the underlying principle remains: geometric reasoning from 2D inputs requires correspondence, whether provided by multi-view tracking, architectural design, or learned positional structure.

Core Insight

By restoring correspondence through positional encodings, we reactivate the powerful non-rigid neural prior within transformers, enabling single-frame 3D lifting at category scale without any 3D supervision.

We instantiate this insight in **2DLFM**, the first lifting foundation model trained exclusively from 2D supervision. By injecting correspondence-aware positional encodings at every layer, 2DLFM learns meaningful non-rigid structure directly from single-frame 2D landmarks, while retaining the scalability advantages of transformer architectures. As a result, a single model reconstructs object-specific 3D structure across over 45 diverse categories: including humans, hands, faces, animals, vehicles, furniture, and deformable objects without 3D labels or category-specific architectural design. The model exhibits strong cross-category transfer, with low-data categories benefiting from priors learned across the entire taxonomy.

Beyond practical performance, our results offer conceptual clarity: the ability to recover 3D structure from 2D projections fundamentally depends on correspondence. Whether implemented explicitly (SfM), architecturally (Deep NRSfM, C3DPO, PAUL), supervised (3D-labeled transformers), or encoded (our method), correspondence is the unifying principle connecting classical geometry to modern transformer-based perception. Our work revives this principle within

scalable architectures, showing that correspondence-aware transformers can unlock non-rigid neural priors for single-frame 3D lifting at foundation-model scale.

2. Related Work

The field of 2D-to-3D lifting has evolved through distinct paradigms, each making tradeoffs between supervision requirements, correspondence handling and scalability. We trace this evolution to identify the core tradeoff that motivates our approach.

2.1. Classical Structure Recovery

The foundational work of Tomasi and Kanade [15] established that 3D structure can be recovered from 2D observations through matrix factorization, given reliable correspondences across views. Non-Rigid Structure from Motion (NRSfM) extended this to deformable objects [2], typically using trajectory bases or shape priors. These classical methods demonstrated the sufficiency of 2D supervision but relied critically on explicit correspondence tracking, failing when correspondences are unreliable due to occlusion or appearance changes.

2.2. Deep 2D→3D Lifting

Modern deep learning transformed the landscape by learning correspondence implicitly. C3DPO [12] introduced canonical lifting with 2D-only supervision through carefully designed bottlenecks. PAUL [17] improved this with Procrustean alignment to focus on non-rigid deformations. Deep NRSfM++ [18] handled missing data and perspective effects. More recently, Maiti et al. [10] borrows from [18] by using local low rank priors in their unsupervised loss function along with an MLP-Mixer network for better scalability. However, all these methods share a critical limitation: they require object-specific architectural constraints (bottleneck dimensions, local priors, rank specifications) that must be manually tuned per category. This prevents training a single unified model across diverse object classes, fundamentally limiting their scalability.

2.3. Foundation Models for 3D

Recent work has pursued foundation model capabilities for 3D tasks. 3D-LFM [3] achieved remarkable scalability with a single transformer handling 30+ categories, leveraging permutation equivariance to share knowledge across different keypoint configurations. NRSfM-Transformer [8] and SimpleBaseline3D [11] similarly demonstrated transformer effectiveness. Yet all transformer approaches thus far require 3D supervision. The field has accepted this as fundamental: permutation equivariance is considered necessary for scalability but incompatible with 2D-only learning. Our work challenges this assumption.

2.4. Positional Encoding in Transformers

Positional encoding provides spatial context to attention mechanisms. While extensively studied in NLP and vision [4, 16], its role in geometric reasoning and 2D→3D lifting remains underexplored. Zheng et al. [25] revisit positional encodings for implicit neural representations and propose Token Positional Encoding (TPE), which injects coordinate information directly into token embeddings. However, when such encodings are added only once at the input, deep transformers can gradually wash out positional structure through attention mixing, allowing permutation equivariance to effectively re-emerge in later layers.

Beyond sinusoidal encodings, various spatial priors have been proposed. Random Fourier Features (RFF)[13] and graph Laplacian eigenvectors[5] encode coordinates or topology, but they rarely analyze positional encodings with permutation equivariance in unsupervised 2D→3D lifting.

Our approach reinterprets positional encoding as a mechanism for restoring *correspondence*. We adapt coordinate-based encodings in the spirit of TPE and RFF to define a correspondence positional encoding (CPE) tailored to landmarks, and we combine this with per-layer injection (Sec. 3.3) so that spatial identity is reinforced throughout the network.

3. Method

Our goal is to learn a 2D→3D lifting model that scales across diverse object categories using only 2D supervision. The key challenge is that transformers are permutation equivariant, which eliminates the correspondence signal necessary to learn non-rigid 3D structure from single-frame landmarks. We first formalize the problem, then show why 2D-only lifting is impossible without correspondence, and finally present a simple architectural change that restores it.

3.1. Problem Formulation

Given N 2D keypoints $\mathbf{X} \in \mathbb{R}^{N\times 2}$ from a single image, we aim to recover their 3D positions $\mathbf{Y} \in \mathbb{R}^{N\times 3}$ under weak-perspective projection:

$$\mathbf{X} = s\,\mathbf{P}\mathbf{R}\mathbf{Y}^\top, \qquad \mathbf{P} = [\mathbf{I} \mid \mathbf{0}],$$

with unknown scale s and rotation $\mathbf{R}$. The network predicts $\hat{\mathbf{Y}} = f_\theta(\mathbf{X})$ and is trained using only 2D reprojection consistency:

$$\mathcal{L}_{2D} = \min_{s,\mathbf{R}} \|\mathbf{X} - s\,\mathbf{P}\mathbf{R}\hat{\mathbf{Y}}^\top\|_F^2. \tag{1}$$

We recover $(s, \mathbf{R})$ via differentiable Procrustes alignment, enabling the network to focus solely on non-rigid shape.

3.2. Why Transformers Fail

Transformers satisfy permutation equivariance:

$$f(\mathbf{\Pi}\mathbf{X}) = \mathbf{\Pi} f(\mathbf{X}),$$

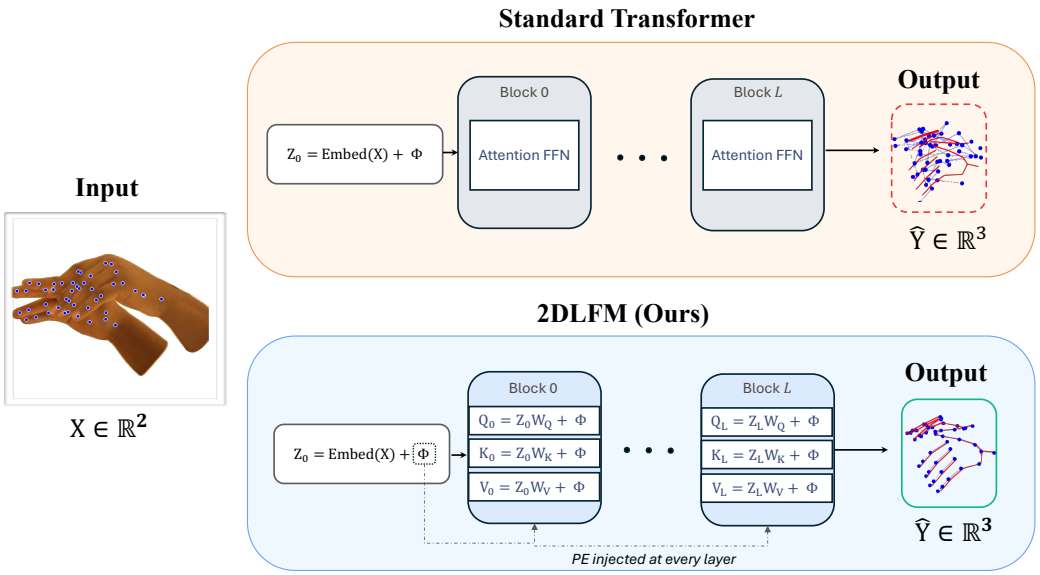

Figure 2. **Overview**: Breaking permutation equivariance preserves correspondence. Standard transformers (with baked-in permutation equivariance) (top) inject PE once, losing spatial identity through attention propogation. 2DLFM (bottom) injects PE at every attention layer, maintaining correspondence necessary for 2D→3D lifting.

which allows a single model to handle many categories with different keypoint layouts. However, this also destroys stable token identity: tokens no longer represent consistent anatomical parts across examples. Under the 2D loss (1), all permutations of landmarks produce identical training signals. The following result formalizes this.

Proposition 1 (Non-identifiability under permutation equivariance)**.** *Let f_θ be permutation equivariant and assume both the data distribution and the 2D loss $\mathcal{L}_{2D}$ are invariant to permutations of keypoint indices. If $\theta^\star$ minimizes $\mathcal{L}_{2D}$, then for any permutation $\mathbf{\Pi}$ there exists $\tilde{\theta}$ such that*

$$\mathcal{L}_{2D}(\tilde{\theta}) = \mathcal{L}_{2D}(\theta^\star), \qquad f_{\tilde{\theta}}(\mathbf{X}) = \mathbf{\Pi} f_{\theta^\star}(\mathbf{X}).$$

Thus token identity is unidentifiable: 2D-only supervision provides no information about which token corresponds to which semantic part.

Sketch. Permutation equivariance permutes outputs when inputs are permuted; the 2D loss and data distribution do not distinguish such permutations. Any permuted solution therefore achieves the same loss. □

This explains the catastrophic failure of standard transformers in Fig. 1: 2D-only lifting requires correspondence, which permutation equivariance removes.

3.3. Correspondence via Per-Layer PE

Standard transformers add positional encodings once at input:

$$\mathbf{Z}_0 = \mathrm{Embed}(\mathbf{X}) + \mathbf{\Phi}, \tag{2}$$

but positional information is rapidly washed out by attention mixing, reinstating permutation symmetry and preventing learning of non-rigid structure.

We restore correspondence by injecting positional encoding into *every* attention layer:

$$\begin{aligned} \mathbf{Q}_\ell &= \mathbf{Z}_\ell \mathbf{W}_Q + \mathbf{\Phi}, \\ \mathbf{K}_\ell &= \mathbf{Z}_\ell \mathbf{W}_K + \mathbf{\Phi}, \\ \mathbf{V}_\ell &= \mathbf{Z}_\ell \mathbf{W}_V . + \mathbf{\Phi} \end{aligned} \tag{3}$$

This ensures that attention scores remain spatially aware thus maintaining token identity and circumventing the non-identifiability in Proposition 1.

3.4. Positional Encoding for Correspondence

A common approach for injecting positional information into transformer architectures is to add a fixed sinusoidal encoding to the input tokens [16]. In its most general form, the i-th token receives a deterministic encoding

$$\begin{aligned} \mathbf{\Phi}_{2k,i} &= \sin(\omega_k \cdot i), \\ \mathbf{\Phi}_{2k+1,i} &= \cos(\omega_k \cdot i), \end{aligned} \tag{4}$$

where ω_k denotes the frequency associated with the k-th embedding dimension.

ViT PE. The standard ViT implementation corresponds to choosing

$$\omega_k = 10000^{-2k/D}$$

where the distribution of ω_k controls the spatial selectivity of the positional encoding.

Analytical RFF PE. Inspired by Random Fourier Features [13], we adopt an *analytical* (deterministic) variant that is more sample–efficient in low-dimensional settings. Rather than drawing frequencies at random, we deterministically cover the Gaussian spectral density by inverting its CDF,

$$\omega_k = \sigma \cdot \mathrm{erf}^{-1}(2k/D),$$

This yields a structured set of Fourier modes with significantly reduced variance compared to Monte-Carlo RFF sampling, allowing us to obtain richer positional encodings with fewer features. Empirically (see Sec. 4) we found the above strategy to significantly outperform naive ViT style positional encoding. After some experimentation, a σ of 2.5 was found to deliver the best results. In lifting tasks the number of keypoints is small (typically $N \in [10, 25]$), so the canonical ViT frequency schedule yields insufficient frequency resolution to reliably distinguish nearby tokens.

Graph PE. Graph based positional encoding has also become increasingly popular [14]. The approach takes the graph Laplacian of the observation and extracts the singular vectors and values from which the positional encoding is then formed (see supplementary for more details). A full ablation can be found in Sec. 4.3.2.

3.5. Architecture and Training

Embedding. We project 2D keypoints to a D-dimensional embedding and add a learnable category encoding:

$$\mathrm{Embed}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathrm{proj}} + \mathbf{b}_{\mathrm{proj}} + \mathrm{TPE}(\mathbf{X}) + \mathbf{e}_c, \quad (5)$$

where $\mathbf{W}_{\mathrm{proj}} \in \mathbb{R}^{2\times D}$ and $\mathbf{b}_{\mathrm{proj}} \in \mathbb{R}^D$ are learned linear projection parameters, and $\mathbf{e}_c \in \mathbb{R}^D$ is a learned category embedding used in multi-category training. The term $\mathrm{TPE}(\mathbf{X})$—following 3D-LFM [3]—is a Random Fourier Feature (RFF)–based positional encoding applied directly to each keypoint's *2D coordinates*, rather than to its token index, ensuring the embedding preserves spatial geometry.

Backbone. We use L transformer layers with standard multi-head attention and FFN blocks, modified only by correspondence per layer PE (see Sec. 3.3) injection.

Output. A linear projection yields $\hat{\mathbf{Y}} \in \mathbb{R}^{N\times 3}$.

Multi-category lifting. We pad all categories to $N_{\max}$ with mask $\mathbf{M}_c$ and apply the 2D loss only to valid keypoints:

$$\mathcal{L}_{2D} = \min_{s,\mathbf{R}} \|\mathbf{M}_c \odot (\mathbf{X} - s\mathbf{P}\mathbf{R}\hat{\mathbf{Y}}^\top)\|_F^2.$$

Optimization. We use Adam (lr=10^{-4}, wd=10^{-4}), batch size 64, and balanced category sampling. Single-category models converge in 50–100 epochs; the 45+ category model in 100–150 epochs.

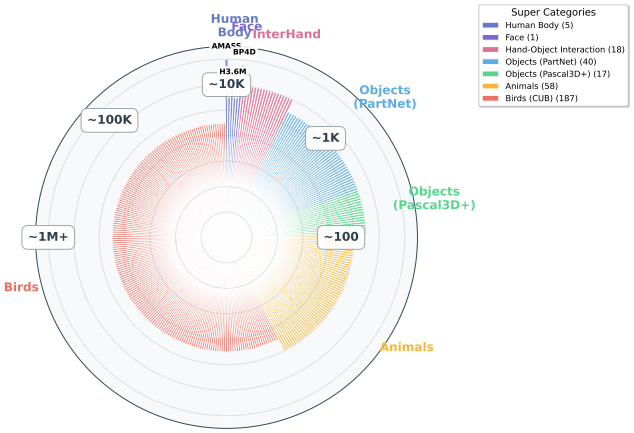

Figure 3. **(b) Training breadth and long-tail.** 2DLFM is trained with 45+ object categories spanning humans, hands/face, animals, birds, and rigid objects (e.g., Pascal3D+). The radial chart shows per-category sample counts on a logarithmic scale (rings: ≈100, ≈1K, ≈10K, ≈100K, ≈1M+), revealing a pronounced long-tail yet broad coverage across super-categories.

Efficiency. PE injection adds $< 2\%$ FLOPs and $< 3\%$ training time. Our full model has ∼25M parameters, comparable to 3D-LFM.

Summary. A single modification - *inject correspondence positional encoding at every layer* breaks permutation symmetry, restores token identity, and enables accurate single-frame 2D-only lifting at foundation scale.

4. Experiments

We validate our hypothesis that **correspondence is a necessary condition** for 2D-only 3D lifting at scale. Without correspondence, the lifting problem becomes degenerate, as shown in classical structure from motion and transformer-based methods. We address four key questions: **Q1**: Can transformers learn 3D lifting from 2D supervision? With proper per-layer PE, they match 3D-supervised methods (Sec.4.1). **Q2**: Does this scale across categories? A single model tackles 45+ categories with strong low-data transfer (Sec.4.2). **Q3**: What architectural choices matter? Through ablations, we show that PE injection strategy (where to add PE) matters more than PE type choice (what to add, *i.e.*, type of PE) (Sec. 4.3). **Q4**: How does landmark-based lifting compare to scene-centric depth? We evaluate against RGB-based reconstruction methods (VGGT [19], DUSt3R [20]) to validate that semantic correspondence outperforms dense scene understanding for object-specific geometry in Sec. 4.4. We begin by describing our experimental setup: datasets, baselines, metrics, and implementation details and then systematically address each question in Sec. 4.1-4.4.

Datasets. We evaluate on diverse benchmarks: Pascal3D+ [21] (12 rigid object categories, $30,000$ instances); Human3.6M [7] (3.6M human motion frames, 17 keypoints); Animal3D [22] (quadrupeds); ARCTIC [6] (hand-object interactions). For foundation model evaluation, we construct a combined dataset merging all sources (45+ categories, heavily imbalanced: 37-1M+ samples per class, Fig. 3).

Baselines. We compare three paradigms: (i) *MLP-based 2D-only:* PAUL [17], C3DPO [12] - learn from 2D supervision but require per-category architectures with manually tuned bottlenecks; (ii) *Transformer-based 3D-supervised:* 3D-LFM [3]: scales across categories but needs 3D labels; (iii) *Scene-level reconstruction:* VGGT [19], DUSt3R [20]—estimate dense depth from RGB. For scene-level methods, we evaluate only visible keypoints (masking occluded points) since they lack explicit occlusion reasoning.

Metrics. We report Mean Per Joint Position Error (MPJPE) in millimeters after Procrustes alignment. Results averaged over 3 seeds; standard deviations reported if > 1mm.

Implementation. We use a transformer with 6-12 layers, 8-16 heads, and embedding dimension $D = 256$-1024 scaled by dataset size. Training uses AdamW, 10^{-4} learning rate, batch size 64, and early stopping based on validation MPJPE. Procrustean alignment solves for scale s and rotation R via SVD [18]. Experiments use 1 NVIDIA A100 GPU.

4.1. Can Transformers learn from 2D Only?

Table 1 highlights our key finding: preserving correspondence enables transformers to perform 3D lifting from 2D supervision with multi-category scalability. Unlike MLPs (C3DPO, PAUL) needing per-category networks or transformers (3D-LFM) requiring 3D supervision, our 2DLFM achieves strong results (11.2mm/9.3mm Pascal3D+, 38.1mm/35.8mm Human3.6M) without 3D supervision.

The large gap between input-only ($>$100mm) and per-layer injection ($\sim$10mm) confirms what classical SfM proves: correspondence is not merely helpful but necessary. Without it, the problem admits infinitely many solutions. Our per-layer injection continuously reinforces spatial identity throughout the network, maintaining the necessary condition that enables stable geometric reasoning. Crucially, we achieve this while maintaining the scalability benefits of transformers. A single unified model handles all categories without architectural changes or per-class bottlenecks.

Figure 1 already compares VGGT [19] (scene-level depth estimation) to 2DLFM (semantic object lifting). Scene-centric methods yield dense depth maps, but back-projecting 2D landmarks causes warped geometry. Our semantic approach ensures plausible objects by explicitly modeling correspondence as evident in Table 1

Table 1. **Breaking the Supervision-Scalability Trade-off.** Previous work forced a choice: MLPs enable 2D-only learning but lack scalability, while transformers scale but need 3D labels. Our PE injection restores correspondence, enabling 2D-only learning and multi-category scalability. Notably, standard ViT-style transformers with input-only PE fail ($>$150mm), highlighting the importance of correspondence preservation.

Method	2D-only	Multi-cat	Pascal3D+ ↓	Human3.6M ↓
MLP-based (2D supervision)				
C3DPO [12]	✓	×	15.0	95.6
PAUL [17]	✓	×	9.4	88.3
Transformer-based (requires 3D)				
3D-LFM [3]	×	✓	5.2	46.3
SimpleBaseline3D [11]	×	✓	8.7	52.1
Scene-level reconstruction				
VGGT [19]	✓	✓	89.4*	107.8*
Naive transformers (2D-only)				
+ PE (input only, ViT-style)	✓	✓	92.3	52.4
+ Fourier (input only)	✓	✓	40	35.4
+ Laplacian (input only)	✓	✓	40	34.7
2DLFM (Ours, 2D-only)				
+ Per-layer Fourier	✓	✓	**8.1**	**30.9**

**Back-projected from scene depth (5×5 median); see Fig. 1*

4.2. Foundation Model Capabilities

Table 2 shows our approach achieves true foundation model behavior: a single unified model improves performance across all categories via cross-category knowledge transfer, with particular gains in low-data regimes. We compare *per-category* training (separate models) with *combined* training (a single model). **Low-data categories benefit most.** For *bottle* (1601 samples), error drops from 100mm to 7.2mm (96.5%). *Chair* (949 samples) improves 85.1%, and *sofa* (669 samples) gains 56.4%. **Even high-data categories benefit:** *aeroplane* improves 34.8%. Our per-layer PE design enables parameter sharing across categories while preserving spatial correspondence *within* each class. Transformer weights remain constant, with only the PE varying per category. This, along with shared attention parameters that learn geometric relationships, allows low-data categories to "borrow" structural priors from high-data ones.

Scaling trends. Performance continues improving as we add categories, suggesting our approach scales beyond current experiments. At 45+ categories, overall performance remains stable: while a small number of classes show mild degradation, most categories: including many rare ones benefit from the shared corpus of geometric knowledge.

4.3. Ablation Studies

We conduct a systematic analysis to validate our central hypothesis: that **continuous correspondence signaling** is the critical factor enabling 2D-only lifting, and our **per-layer injection** is the key mechanism. We isolate the components of our approach (Fig. 4) to investigate: **(1)** The impact of the PE *injection strategy* (e.g., every-layer vs. input-only); **(2)** The importance of the PE *type* (e.g., graph-based vs.

Table 2. **Multi-Category Training Benefits.** Combined training with a single unified model dramatically improves performance compared to per-category models, with particularly strong gains for low-data categories. The *bottle* category (only 1601 samples) improves by 96.5% by leveraging knowledge from related shapes. Averaged across the following categories, combined training improves reconstruction accuracy by 59.1%, showing emergent foundation model capabilities enabled by preserved correspondence.

Category	Samples	Per-cat ↓	Combined ↓	Improvement
aeroplane	1,953	15.2	9.9	34.8%
bottle	1,601	100	7.2	92.8%
chair	949	38.2	5.7	85.1%
sofa	669	48.4	21.1	56.4%
bicycle	904	12.3	8.9	27.6%
car	5,531	8.9	6.7	24.7%
drosophilia	80	23.4	1.8	92.3%

conventional); and **(3)** The *scalability* of our architecture.

4.3.1. Where to inject PE?

As shown in Figure 4(a), continuous correspondence signaling is essential. Input-only injection - standard practice in Vision Transformers [4] yields 100.3mm on Pascal3D+ and 63.4mm on Human3.6M, showing complete failure for geometric tasks despite success in ViT. Injecting at first and last layers improves slightly in low data regimes to 92.1mm and 71.2mm, showing boundary conditions help somewhat but remain insufficient. Even skipping alternate layers (*every 2nd layer*: 26.1mm, 34.5mm) degrades performance substantially compared to our *every-layer* approach (8.1mm, 33.1mm).

This phenomenon occurs as attention operations in standard transformers mix token representations without positional awareness: each layer's output becomes increasingly permutation-invariant. By subsequent layers, spatial identity is effectively lost unless actively reinforced. Our per-layer injection maintains correspondence throughout the computational graph, preserving the necessary condition for lifting across deep networks.

4.3.2. What type of PE to inject?

This ablation focuses on exploring the design space of positional encodings, all evaluated with our per-layer injection strategy to provide fair comparison. We entertained two paradigms: (i) *graph-based PE* (Laplacian eigenvectors), motivated by the desire to inject skeletal topology directly, and (ii) *conventional PE* (RFF), which assumes simpler spatial structure, as shown in Fig. 4(b).

Surprising finding : Weighted conventional PE nearly matches graph-based methods: RFF achieves 11.2mm (Pascal3D+) and 38.1mm (Human3.6M) without graph knowledge, closely matching Graph Laplacian (9.3mm, 35.8mm) using true skeletal structure.

Worth noting the **critical distinction:** Standard ViT-style positional encodings (injected only at input) fail (92.3mm), showing that *how* we inject matters more than *what* we inject. Our appropriately weighted sinusoidal features (8.1mm), continuously reinforced per-layer, provide sufficient correspondence signal without requiring topological priors. This validates our hypothesis: the correspondence preservation mechanism (per-layer injection) is more fundamental than the encoding choice (graph vs. conventional). The convergence of all PE types under proper injection ($\sim 10-13$mm) further supports our core claim: once the necessary condition (correspondence) is satisfied, the specific encoding becomes secondary.

Practical implications. For practitioners, RFF offers the best trade-off: zero parameters, no graph knowledge required, and performance within 2mm of graph-based methods. For absolute best performance, Graph Laplacian provides modest improvement at the cost of requiring skeletal topology.

4.3.3. Architecture scaling.

Figure 4(c) validates a key implication of our method: it unlocks the ability to scale network depth for **improved geometric reasoning**. Performance consistently improves from 15.3mm (4 layers) to 9.3mm (12 layers), and further to our best single-model result of 8.1mm (24 layers). This scaling is non-trivial and highlights a fundamental limitation of standard ViTs for geometric tasks, where correspondence **fades** with depth, rendering deep refinement impossible. Our per-layer injection, in contrast, acts as a **continuous spatial anchor**, maintaining stable grounding throughout the entire computational graph. This allows the network to successfully leverage **deep hierarchical refinement**, progressing from coarse spatial structure in early layers to resolving **complex articulation constraints and fine-grained details** (like finger poses or subtle rotations) in deep layers - proving that continuous correspondence is the necessary condition for scaling transformers on complex 3D tasks.

4.4. Qualitative Results and Failure Analysis

Figure1 shows reconstructions for 2D-LFM and scene-centric depth models (VGGT). From 2D keypoints, 2DLFM recovers plausible 3D structures: proper joint geometry, symmetrical vehicles, and proportional objects. In contrast, depth-based back-projection distorts geometry, as shown in Table 1. Dense geometry fails on object-specific shapes; correspondence-aware lifting better preserves landmark-level structure.

Robustness and real-world applicability. We evaluate robustness to noisy 2D inputs and to off-the-shelf ViT-

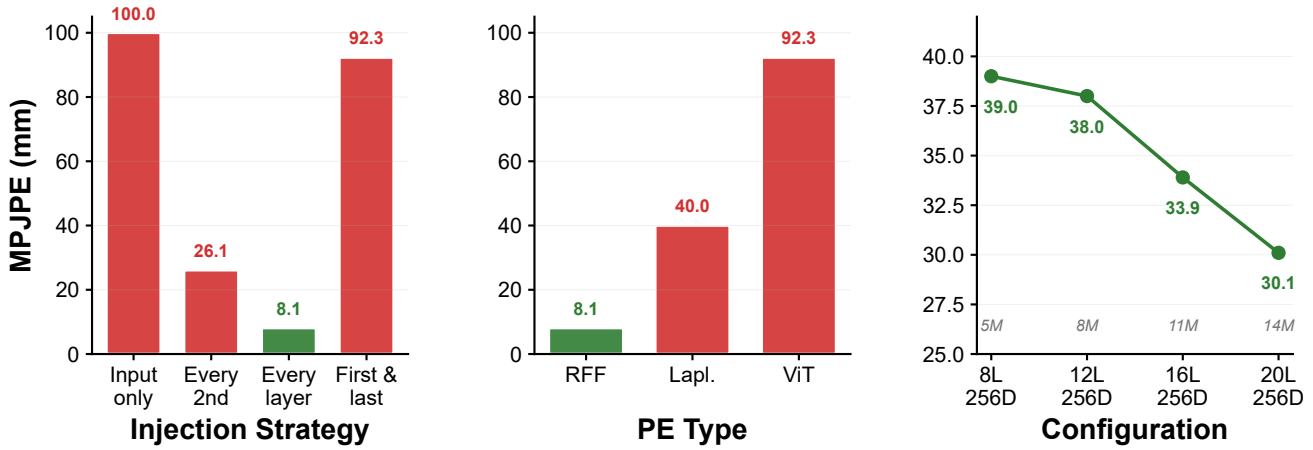

Figure 4. **Ablation Studies.** We systematically evaluate the impact of PE injection strategy, PE type, and architecture scaling, all with per-layer PE on a 16 layer transformer unless otherwise noted (Pascal3D+ scores shown). **(a) PE Injection Strategy.** Continuous correspondence signaling throughout the network is essential. Standard ViT practice (input-only PE) catastrophically fails (100mm), while our *every-layer* approach maintains spatial identity, achieving 8.1mm. Even sparse injection (every 2nd layer) degrades performance substantially (26.1mm). **(b) PE Type Comparison.** With per-layer injection, the specific PE type also matters. Graph-based Laplacian PE achieves 40mm, while our fourier PE (requiring no topological knowledge) far surpasses the others at 8.1mm. Standard ViT-style PE (input-only, 92.3mm) is shown for comparison, emphasizing the importance of injection strategy. **(c) Architecture Scaling with Per-Layer PE.** Performance consistently improves with network depth, from 39.0mm (4 layers) to 30.1mm (24 layers) on human3.6m. This validates that continuous correspondence signaling enables stable scaling of deep transformers for geometric tasks, unlike input-only PE methods that degrade beyond shallow depths.

Pose [23] detections on Human3.6M. With ViTPose keypoints, 2D-LFM achieves 34.4 mm MPJPE (vs. 30.9 mm with ground-truth), and Gaussian noise analysis confirms graceful degradation with no catastrophic failure (see supplementary Sec. 4.3 for full details).

Failure cases: monocular depth ambiguity. Because our model observes only a single frame, monocular ambiguities remain: limbs may flip in depth, extreme foreshortening can shorten limbs, and occluded landmarks may collapse. These single-view limits are addressable with multi-view, motion, or physical priors in future work.

Our experiments validate three key claims: (1) per-layer PE injection restores correspondence, enabling transformers to achieve 8.1mm MPJPE vs. $>$ 90mm for standard ViT approaches; (2) it scales to 45+ categories, with up to 92% gains for low-data classes via cross-category transfer; (3) correspondence signaling is crucial, while PE type matters less.

5. Conclusion & Discussion

Correspondence is Necessary. 2D$\rightarrow$3D lifting is inherently correspondence-dependent: without fixed token identity, the problem becomes permutation-degenerate. Proposition 1 shows this formally, and our experiments confirm it - standard transformers collapse ($>$100 mm), while maintaining correspondence enables accurate lifting (8.1 mm). Thus correspondence is not an architectural convenience but a required ingredient for single-frame non-rigid reconstruction.

Cross-Category Benefits. Preserving correspondence allows geometric structure learned from large categories to transfer to smaller ones, improving reconstruction without 3D supervision. Our results indicate that correspondence-aware models provide a practical path toward scalable 2D-only lifting across many object types.

Implications. The fundamental barrier in 2D-only lifting is not expressiveness but the loss of token identity. Once correspondence is enforced, transformers can recover meaningful non-rigid 3D structure from a single frame. The sparse 3D landmarks from 2D-LFM are useful as structural priors for dense reconstruction, pseudo-labels for depth foundation models, and action signals in robotics and AR/VR.

References

[1] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 2

[2] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, pages 690–696. IEEE, 2000. 3

[3] Mosam Dabhi, László A Jeni, and Simon Lucey. 3d-lfm: Lifting foundation model. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 10466–10475, 2024. 2, 3, 5, 6

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 7

[5] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. *arXiv preprint arXiv:2110.07875*, 2021. 3

[6] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 6

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6

[8] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3314–3323, 2024. 3

[9] Haorui Ji, Hui Deng, Yuchao Dai, and Hongdong Li. Unsupervised 3d pose estimation with non-rigid structure-from-motion modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3314–3323, 2024. 2

[10] Shalini Maiti, Lourdes Agapito, and Benjamin Graham. Unsupervised 2d-3d lifting of non-rigid objects using local constraints. *arXiv preprint arXiv:2504.19227*, 2025. 3

[11] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 3, 6

[12] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7688–7697, 2019. 2, 3, 6

[13] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 3, 5

[14] Ladislav Rampášek, Michael Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems*, 35:14501–14515, 2022. 5

[15] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International journal of computer vision*, 9(2):137–154, 1992. 2, 3

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3, 4

[17] Chaoyang Wang and Simon Lucey. Paul: Procrustean autoencoder for unsupervised lifting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 434–443, 2021. 2, 3, 6

[18] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards unsupervised 2d-3d lifting in the wild. In *2020 International Conference on 3D Vision (3DV)*, pages 12–22. IEEE, 2020. 3, 6

[19] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 2, 5, 6

[20] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 5, 6

[21] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 6

[22] Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, et al. Animal3d: A comprehensive dataset of 3d animal pose and shape. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9099–9109, 2023. 6

[23] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023. 8

[24] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 2

[25] Jianqiao Zheng, Sameera Ramasinghe, and Simon Lucey. Rethinking positional encoding, 2021. 3