Beyond Spatio-Temporal Priors: A Generalizable Approach for Dense Correspondence Matching

Luping Liu[1,2]* **Bingyi Kang**[2] **Yifan Wang**[2,3]* **Dong Xu**[1]†
[1]The University of Hong Kong [2]ByteDance Seed
[3]Zhejiang University
luping.liu@connect.hku.hk bingykang@gmail.com
accwyf@gmail.com dongxu@cs.hku.hk

Abstract

Dense correspondence matching has historically been bounded by simplifying spatio-temporal priors, such as smooth motion and rigid geometry. While effective for classical tasks, these assumptions break down in image editing and reference-guided generation (IEG), where transformations can preserve visual identity while breaking physical continuity. To establish identity-preserving correspondence across such transformations, we introduce FreeMatching, a generalizable framework combining generative and semantic foundation representations with heterogeneous supervision from classical datasets, tracked videos, and synthetic scenes. Teacher-guided iterative refinement further improves correspondence in IEG without dense correspondence annotations. Experimentally, a single FreeMatching model substantially improves correspondence quality on challenging IEG image pairs while retaining competitive performance on classical benchmarks. Furthermore, we demonstrate its utility as a quantitative metric for evaluating identity preservation, with scores that correlate with human judgment. The code is available at https://github.com/luping-liu/FreeMatching.

1 Introduction

Dense correspondence matching has historically been fractured into distinct paradigms: optical flow [1–3] for temporal motion and dense matching [4, 5] for geometric views. While unified frameworks like UniMatch [6] and UFM [7] have bridged these tasks, they remain constrained by simplifying spatio-temporal priors. As illustrated in Figure 1, correspondence exists on a continuous spectrum. While existing methods excel at the left end governed by smooth motion and rigid geometry, they falter at the frontier of image editing and reference-guided generation (IEG) [8–10]. For example, Figure 1 shows a horse in a grassy landscape and a snowy scene, with substantial changes in pose and appearance while its identity remains recognizable. Such transformations can preserve visual identity while breaking physical continuity. Classical spatio-temporal assumptions no longer suffice, and existing methods can produce omissions or distortions, as shown in Figure 4.

To move beyond restrictive spatio-temporal priors, we draw inspiration from human cognition [11]. Human perception can recognize the same instance despite substantial changes in appearance and configuration. Motivated by this principle, we aim to learn identity-preserving [12] correspondence across the diverse tasks shown in Figure 1.

To achieve this goal, we introduce *FreeMatching*, which combines foundation-model representations, heterogeneous correspondence supervision, and teacher-guided iterative refinement. Specifi-

*Work done while Luping Liu and Yifan Wang were interns at ByteDance Seed.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

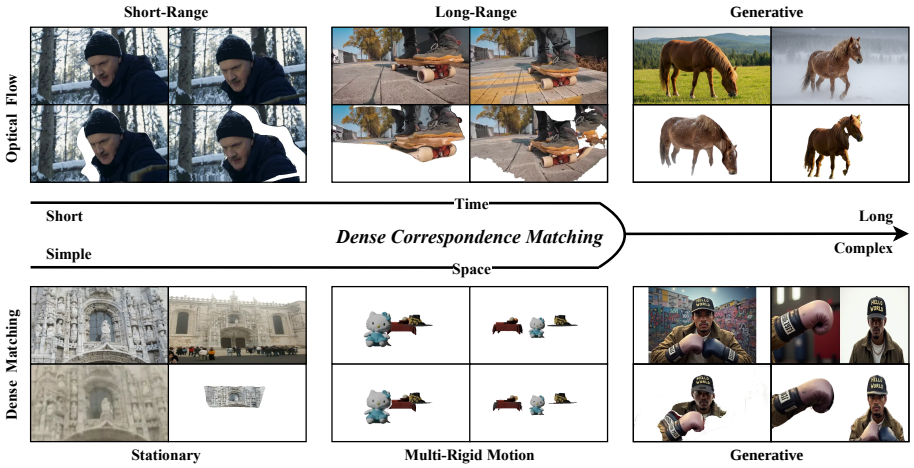

Figure 1: Extending Dense Correspondence to IEG. FreeMatching matches across optical flow, geometric matching, and IEG. Each block shows input pairs above and our bidirectional warps below. The classical examples vary in temporal separation (top) and geometric complexity (bottom).

cally, we utilize FLUX2-4B [13] and DINOv3 [14] to inherit rich generative and semantic knowledge, respectively. We develop this capability through two stages of training: (1) supervised learning on new large-scale tracked video [15, 16] and synthetic datasets, combined with classical datasets [17, 18], where Huber loss [19] is employed to mitigate annotation noise; and (2) teacher-guided iterative refinement on IEG datasets without dense correspondence annotations. This stage refines local correspondences and provides supervision for covisibility. Together, these components extend correspondence learning beyond classical geometric and motion settings.

Our experiments demonstrate the power of FreeMatching's generalized approach. A single model achieves performance competitive with SOTA methods on classical optical flow and dense matching benchmarks, while substantially improving correspondence quality on the challenging IEG task. This success enables an application: using FreeMatching as a perceptual metric for quantifying identity preservation [20], providing objective scores that correlate with human judgment.

Our contributions are threefold:

1. We introduce FreeMatching, a generalizable framework that extends dense correspondence from classical geometric and motion settings to identity-preserving matching in IEG.
2. We combine foundation-model representations, heterogeneous correspondence supervision, and teacher-guided iterative refinement through a two-stage training paradigm, supported by new large-scale video and synthetic datasets.
3. We demonstrate competitive performance on classical tasks and substantial gains on IEG, enabling spatially resolved assessment of identity preservation with scores that correlate with human judgment.

2 Related Work

2.1 Classical and Unified Correspondence

The pursuit of dense correspondence has evolved significantly. Initially divided into optical flow for temporal motion [1] and geometric matching for stationary scenes [4], the field was revolutionized by deep learning. Architectures like RAFT [3] set a new standard with the paradigm of iterative refinement over cost volumes. This paradigm was pushed further by specialized models, such as GMFlow [21] which framed flow as a global matching problem, and RoMa [5], which leverages powerful foundation models like DINOv2 [22] for state-of-the-art wide-baseline matching.

This progress spurred a move towards unifying these tasks. Frameworks like UniMatch [6] and UFM [7] have successfully demonstrated that a single, powerful model can match specialized methods on traditional benchmarks.

2.2 Semantic Correspondence and Foundation Features

Semantic correspondence research addresses geometry and viewpoint sensitivity through geometry-aware matching [23] and viewpoint-guided spherical maps [24]. GECO [25] and diffusion-feature distillation [26] study efficient correspondence representations, while Jamais Vu [27] examines the generalization gap of supervised semantic matching. Geometry-grounded models such as DUSt3R [28] and VGGT [29] provide another source of foundation features, with recent work adapting VGGT priors to dense semantic matching [30]. These advances motivate learning representations that transfer across transformations.

2.3 Learning Paradigms for Correspondence

Training paradigms have co-evolved alongside architectures. The standard of supervised learning on synthetic datasets like FlyingThings [18] has been augmented by two key trends: unsupervised learning through data distillation [31], and the unification of diverse data sources—such as pre-training optical flow models on wide-baseline data—to improve generalization [32].

Building on these advances, FreeMatching extends a single correspondence model from classical geometric and motion settings to IEG. We combine generative and semantic foundation representations with heterogeneous supervision from classical datasets, tracked videos, and synthetic scenes, followed by teacher-guided iterative refinement on IEG image pairs. This combination supports instance-level matching across changes in appearance, pose, and scene composition while retaining classical correspondence capabilities.

3 FreeMatching

We present *FreeMatching*, a unified framework that extends classical correspondence matching to IEG. We first define identity-preserving correspondence, then describe the foundation-model architecture, supervised pre-training, and teacher-guided iterative refinement. Finally, we introduce a reconstruction-based evaluation protocol for matching quality and identity-preserving consistency.

3.1 Task Definition

Classical correspondence tasks use geometric and temporal structure to establish matches. IEG extends this setting: an instance can retain its visual identity even when changes in appearance, pose, and scene composition break physical continuity.

We formulate dense correspondence as identity-preserving visual matching. Rather than requiring correspondences to follow rigid geometry or smooth motion, we seek to match regions depicting the same underlying instance across changes in viewpoint, pose, appearance, and scene composition. This formulation encompasses classical correspondence tasks and extends to IEG, wherever an identifiable visual counterpart remains.

Formally, given a source image I_a and a target image $I_b \in \mathbb{R}^{H\times W\times 3}$, our goal is to predict a dense backward correspondence map $C \in \mathbb{R}^{H\times W\times 2}$ and a covisibility mask $M \in [0,1]^{H\times W}$. Specifically, $C(x,y) = (u,v)$ denotes that pixel (x,y) in I_b corresponds to the absolute coordinate (u,v) in I_a. The mask indicates whether a valid counterpart exists in the source image. Target regions without an identifiable source counterpart, such as newly synthesized content or content revealed after an occluder is removed, have a covisibility target of $M = 0$. Shared regions remain matchable despite changes in appearance or configuration.

3.2 Foundation Models

To achieve this goal, we combine the complementary representations of two foundation models. We initialize our generative transformer with FLUX.2-klein-base-4B [13] (FLUX for brevity) and inject

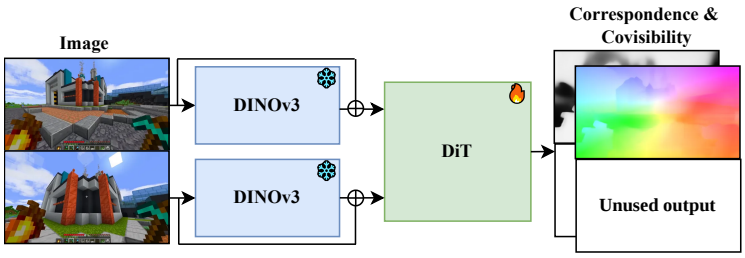

Figure 2: FreeMatching Architecture. Our transformer is initialized from FLUX.2-klein-base-4B. Each image is encoded into latent tokens and augmented with projected DINOv3 features. The two token sequences are concatenated for joint attention, and a task-specific decoder predicts the dense correspondence map and covisibility mask in a single forward pass.

Table 1: **Supervised Training Data.** We combine classical correspondence datasets with our tracked-video and Blender data. Ratios indicate sampling proportions, rounded to multiples of 5%.

Dataset	Ratio	Dataset	Ratio
ScanNet++ [33]	15%	Spring [34]	5%
BlendedMVS [35]	10%	PointOdyssey [36]	5%
FlyingThings [18]	10%	FlyingChairs [37]	5%
DynamicReplica [38]	10%		
Video (Ours)	20%	Blender (Ours)	20%

DINOv3 [14] features to incorporate semantic information. The stage-wise results show improved performance when combining generative initialization with DINOv3 features.

As shown in Figure 2, a frozen FLUX VAE encodes each image into latent tokens. We add linearly projected DINOv3 patch features to the corresponding token embeddings and concatenate the two image sequences. Position encodings retain spatial location and image identity, allowing joint attention to model cross-image relationships. During supervised pre-training, we freeze the foundation encoders and train the transformer, semantic projection, and task-specific output decoder. The model predicts the dense correspondence map C and covisibility mask M in a single forward pass. Architecture details are provided in Appendix A.

3.3 Supervised Pre-training

Dataset Collection To establish a robust generalist foundation across a wide range of transformations, alongside standard optical flow and dense matching datasets, we introduce two new large-scale annotated datasets, capturing both real-world diversity and synthetic precision. Specifically, our new datasets comprise:

1. Real-world Dynamic Scenes: We leverage a SOTA tracking model AllTracker [16] to generate dense pseudo-ground truth for 520k clips from the ViPE [39] dataset. By sampling frame pairs with varying temporal intervals, we capture diverse non-rigid motion patterns that extend beyond the smooth assumptions of classical optical flow.
2. Synthetic Objaverse Scenes: To mitigate inherent annotation noise, we render 360k image pairs with Blender using assets from Objaverse [40]. This provides pixel-perfect ground truth with semantic and structural diversity, serving as a high-fidelity anchor to complement the real-world data.

The supervised data sources and their sampling proportions are summarized in Table 1. Detailed descriptions of all datasets are provided in Appendix B.

Robust Training Loss Since real-world datasets contain inevitable annotation noise, standard ℓ_2 loss can destabilize training. We therefore adopt the robust Huber loss [19] for correspondence C

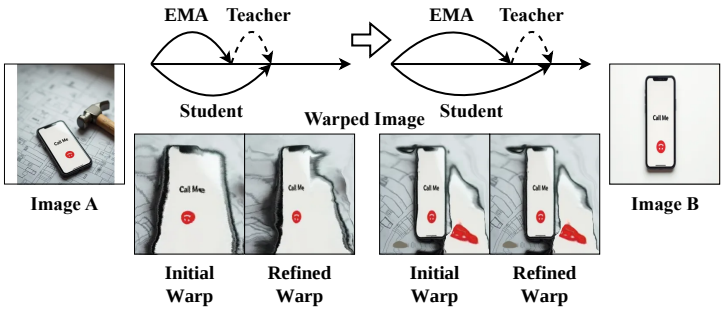

Figure 3: Teacher-Guided Iterative Refinement. An exponential moving average (EMA) copy of the student predicts the correspondence used to produce the Initial Warp. A frozen teacher estimates a local correction; composing it with the initial correspondence produces the Refined Warp and a correspondence target for training the student. The EMA copy evolves with the student while the teacher remains fixed.

regression, while maintaining ℓ_2 for the covisibility masks M. Formally, we minimize:

$$\mathcal{L}_{\text{pre}} = \frac{1}{N} \sum_{x,y} \Big(M^*(x,y) \cdot \mathcal{L}_{\text{Huber}}(C(x,y), C^*(x,y)) + \lambda \| M(x,y) - M^*(x,y) \|_2^2 \Big) \tag{1}$$

where C^* and M^* denote the ground-truth correspondence and covisibility, respectively, and $N = H \times W$ is the number of pixels. The coefficient λ balances mask supervision against correspondence regression; the loss weights are specified in Appendix C. This hybrid objective enables FreeMatching to learn precise matching from clean synthetic data while remaining robust to noisy real-world annotations.

3.4 Teacher-Guided Iterative Refinement

Dataset Collection Supervised pre-training uses classical correspondence datasets, tracked videos, and synthetic scenes. To extend matching to IEG without dense correspondence annotations, we refine the model on image pairs from this domain. The experimental data and refinement configuration are specified in Appendix C.

Refinement Procedure To improve local alignment without dense correspondence annotations, we use a trainable student and a frozen refinement teacher. An exponential moving average (EMA) copy of the student produces the initial correspondence C_0, yielding a warped image $\hat{I}_{\text{inter}} = I_a(C_0)$. The fixed teacher estimates a local correction map S from the target I_b to this intermediate image.

We compose the two maps to obtain the pseudo-target $C_{\text{new}} = C_0 \circ S$. The student learns to match this target in valid regions, while its covisibility prediction is regularized toward the EMA prediction. As the student improves, the EMA copy provides updated initial alignments for the fixed teacher to refine. This separates learning to bridge large changes from correcting the remaining local misalignment.

The refinement teacher can be instantiated with a pretrained correspondence model. Our concrete choice, EMA update, validity criteria, and losses are given in Appendix C.

3.5 Consistency Evaluation

Having established a robust generalist foundation via our two-stage training, we employ a reconstruction-based protocol that serves a dual purpose: evaluating correspondence quality and quantifying data consistency. The core premise is that an ideal correspondence map C should allow the source image I_a to be warped into $\hat{I}_b = I_a(C)$, aligning precisely with the target I_b within the covisible region M. We quantify this alignment fidelity with separately reported reconstruction metrics:

$$\text{Score}_m = \mathcal{D}_m(\hat{I}_b, I_b; M_m), \tag{2}$$

Table 2: Dense Matching and Relative Pose Estimation. Comparison with representative correspondence methods on ETH3D [42] and ScanNet-1500 [17]. A single FreeMatching model achieves competitive performance across outdoor and indoor geometric matching and downstream pose estimation. Best and second-best results are **bold** and underlined, respectively.

	Dense Matching								Pose Estimation (AUC, %)					
	ETH3D				ScanNet-1500				ETH3D			ScanNet-1500		
Method	EPE↓	1px↓	2px↓	5px↓	EPE↓	1px↓	2px↓	5px↓	@5ř↑	@10ř↑	@20ř↑	@5ř↑	@10ř↑	@20ř↑
GMFlow [21]	131.99	93.15	86.95	78.54	53.28	96.95	90.39	73.25	0.17	0.62	2.28	0.42	1.64	5.62
SEA-RAFT [32]	150.79	90.65	86.21	80.93	67.83	95.80	88.03	71.53	0.35	1.25	3.83	1.40	3.94	9.62
UniMatch [6]	112.97	82.20	75.09	67.71	44.64	94.57	84.05	60.83	0.80	2.79	7.26	1.91	5.12	12.02
RoMa [5]	11.26	50.69	37.18	24.53	12.03	91.91	76.48	43.31	53.48	68.32	**78.61**	**31.80**	53.40	**70.90**
UFM [7]	12.98	54.01	35.90	**20.63**	12.03	91.35	74.57	**37.52**	38.77	53.59	66.05	31.60	**54.10**	-
FreeMatching (Ours)	**10.23**	**50.53**	**35.67**	20.71	**8.56**	**91.21**	**74.18**	38.66	**54.22**	**68.76**	78.23	31.51	53.62	70.03

where m indexes the metric and M_m specifies its evaluation region. We report MSE for pixel precision, LPIPS [41] for perceptual similarity, and feature similarities for semantic alignment, without combining them into a single score. Ref-MSE and Ref-DINOv2 use a shared frozen reference mask across methods; Full-MSE and whole-image OpenCLIP use the full image ($M_m \equiv 1$). We assess the consistency of the results across pixel-level, perceptual, and feature-similarity metrics. DINOv3 scores and heatmaps quantify feature alignment, while MSE, LPIPS, DINOv2, SigLIP 2, and OpenCLIP provide complementary views of reconstruction quality. Definitions of all metrics, including SigLIP 2, are provided in Appendix D.

This enables a bi-directional evaluation paradigm:

1. Evaluating Matching Accuracy: On benchmarks where image pairs are known to be consistent, the reconstruction error serves as a proxy for the quality of our predicted correspondence C.
2. Evaluating Identity-Preserving Consistency: Conversely, assuming the matching model is sufficiently robust, this protocol can be inverted to evaluate the data itself. For new image pairs, significant reconstruction error within the covisible region can indicate inconsistencies in the image content, subject to the accuracy of the estimated alignment. Consequently, our model functions as a quantitative metric for identity preservation, enabling the objective assessment of consistency in novel data samples.

4 Experiments

In this section, we empirically validate FreeMatching. We first demonstrate its robustness by achieving competitive performance on classical correspondence benchmarks (optical flow and dense matching). We then showcase its distinct advantage on our proposed IEG-Bench, designed to evaluate correspondence between shared subjects in source or reference images and IEG outputs. Finally, we provide extensive ablation studies, qualitative visualizations, and an analysis of FreeMatching's utility as a perceptual metric.

For refinement, we instantiate the fixed teacher with RoMa [5]. Experimental details are provided in Appendix C, with training and inference costs in Appendix C.4.

4.1 Classical Benchmarks

Tables 2 and 3 compare FreeMatching with representative classical correspondence methods. After 10k IEG refinement, FreeMatching achieves lower EPE than the similarly adapted UFM and RoMa variants on ScanNet and ETH3D. On optical flow, the adapted UFM is stronger on Sintel Clean and KITTI, while FreeMatching is stronger on Sintel Final; UniMatch achieves the lowest EPE among the methods shown. These results demonstrate how FreeMatching extends matching to IEG while maintaining competitive classical performance. The complete 10k adaptation comparison is reported in Table 9.

Table 3: Optical Flow Estimation. Comparison with representative optical-flow and dense-matching methods on Sintel [43] and KITTI-2015 [44], evaluating the cross-task generalization of the same FreeMatching model. Best and second-best results are **bold** and underlined, respectively.

	Sintel (Clean)				Sintel (Final)				KITTI-2015			
Method	EPE↓	1px↓	2px↓	5px↓	EPE↓	1px↓	2px↓	5px↓	EPE↓	1px↓	2px↓	5px↓
GMFlow [21]	1.21	41.3	12.6	1.9	1.56	42.7	14.4	3.2	4.32	60.3	35.5	16.1
SEA-RAFT [32]	0.43	5.9	2.7	0.9	0.81	**8.7**	**4.8**	**2.1**	1.97	29.7	16.8	7.2
UniMatch [6]	**0.40**	**5.4**	**2.4**	**0.8**	**0.75**	8.9	4.9	2.2	**1.52**	**24.1**	**12.5**	**5.1**
RoMa [5]	1.24	25.2	8.1	4.5	2.49	35.6	16.6	8.7	5.91	41.4	27.5	18.4
UFM [7]	0.92	14.2	6.4	2.1	1.48	16.9	9.0	3.8	4.75	39.6	24.5	15.5
FreeMatching (Ours)	1.05	7.1	3.6	1.3	1.36	12.2	6.5	3.1	4.95	32.5	17.6	9.2

4.2 Identity-Preserving Correspondence

The next part, IEG-Bench, probes a model's core ability to establish correspondences while preserving an object's identity. To quantify this, we employ the reconstruction-based evaluation protocol described in Section 3.5 and Appendix D. We compute a perceptual similarity score between the warped source and the target image. A higher score indicates a more accurate correspondence field that better preserves the object's identity.

As detailed in Table 4, our IEG-Bench challenges existing paradigms. While specialized models like SEA-RAFT and UniMatch are ill-equipped for this semantic task, contemporary generalists like RoMa and UFM demonstrate impressive generalization. Thanks to their more relaxed priors and powerful feature backbones, they achieve respectable performance, proving their robustness beyond classical tasks.

RoMa and UFM nevertheless exhibit local distortions or incomplete matches in the IEG examples in Figure 4. The adapted comparison below assesses whether these differences persist after the models receive IEG refinement data.

FreeMatching achieves the strongest reconstruction scores among the compared methods in Table 4. These gains support its ability to establish identity-preserving correspondences across challenging IEG image pairs. In Appendix E, we use human evaluations to assess how our reconstruction-based metric correlates with human judgments of identity preservation.

Table 4: IEG-Bench Evaluation. We report the mean reconstruction similarity scores (higher is better) using SigLIP 2 [45] and DINOv3 [14] features, along with reference-mask MSE (Ref-MSE) and LPIPS [41] (lower is better) between the warped and target images. The identity baseline compares the unwarped source image with the target using an identity correspondence map. Best and second-best results are **bold** and underlined, respectively.

Model	Ref-MSE ↓	LPIPS ↓	SigLIP 2 ↑	DINOv3 ↑
Identity	23.65	81.97	86.67	52.38
SEA-RAFT [32]	18.49	72.45	85.18	58.14
UniMatch [6]	11.07	58.42	87.86	66.35
RoMa [5]	8.08	48.91	87.73	70.15
UFM [7]	10.41	53.05	86.97	66.77
FreeMatching (Ours)	**6.22**	**45.41**	**90.77**	**80.01**

Comparison after weak adaptation. We additionally compare FreeMatching, UFM, and RoMa after 10k adaptation updates using UNO+OmniGen2 pairs. Table 5 reports reference-mask and full-image MSE, together with DINOv2 and OpenCLIP similarities as complementary feature metrics. FreeMatching achieves the best scores across all four metrics. Together with the MSE, LPIPS, SigLIP 2, and DINOv3 results in Table 4, these results show consistent advantages across metrics under both evaluation settings. Its paired Full-MSE advantages over UFM and RoMa are 2.48 (95%

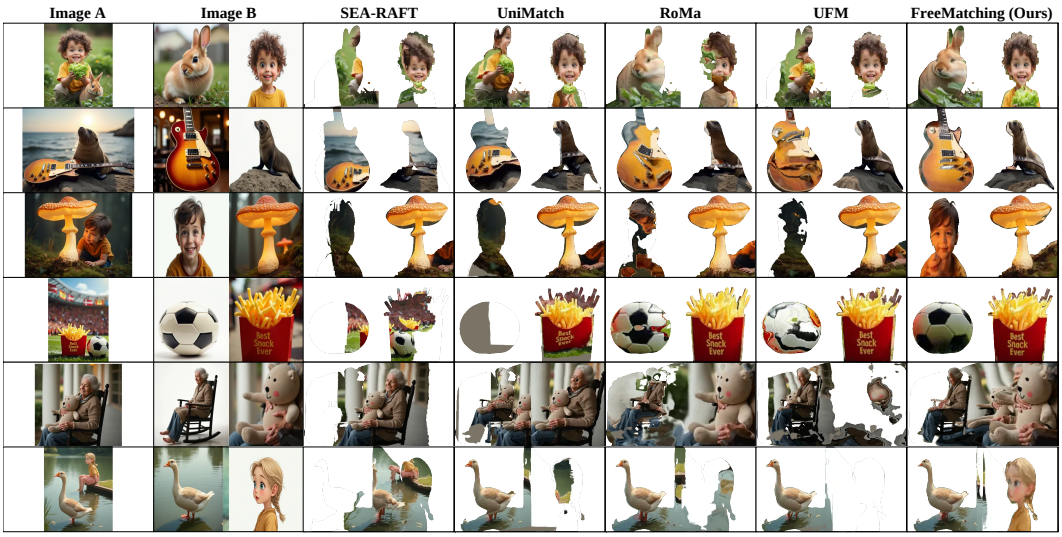

Figure 4: Visual comparisons on IEG-Bench using UNO examples. The references in Image B are composed into Image A ($B \rightarrow A$). Each method warps Image A toward Image B to evaluate correspondence quality.

Table 5: IEG-Bench comparison after 10k adaptation updates. All variants use UNO+OmniGen2 refinement data. Ref-MSE and Ref-DINOv2 use the shared frozen reference mask; Full-MSE and OpenCLIP use the full image. All metrics are scaled by 100. DINOv2 [22] and OpenCLIP [46] complement the DINOv3 metric used in the stage-wise analysis.

Method	Ref-MSE ↓	Full-MSE ↓	Ref-DINOv2 ↑	OpenCLIP ↑
FreeMatching (10k-IEG)	**6.22**	**10.99**	**54.31**	**70.56**
UFM (10k-IEG) [7]	8.55	13.48	39.65	56.48
RoMa (10k-IEG) [5]	9.70	16.57	47.14	61.18

CI: [1.54, 3.43]) and 5.57 ([4.06, 7.13]), respectively. Evaluation definitions and paired statistics are given in Appendix D.5.

4.3 Qualitative Results

Figure 4 illustrates the matching challenges behind the quantitative results. The examples require preserving instance-level correspondences across substantial changes in appearance, pose, and composition. The compared methods exhibit local distortions or incomplete matches, while FreeMatching produces more coherent alignments in these examples.

4.4 Fine-Grained Identity Assessment

Leveraging FreeMatching's precise alignment, we extend its utility to act as a fine-grained metric for identity preservation. Following Section 3.5, by warping the source image to the target and computing local similarity strictly within the covisible region, we generate a spatial heatmap of perceptual divergence.

This approach addresses a critical blind spot of global metrics [20] (e.g., global DINO or CLIP scores). While global metrics effectively measure high-level semantic coherence, they inherently aggregate spatial information, lacking the capability to provide granular feedback on local details. Consequently, they often mask subtle yet critical deviations in texture or structure that FreeMatching reveals.

Figure 5 illustrates this spatially resolved assessment. In the second row, the heatmap highlights the inconsistency in the balloon's text. By showing where local appearance departs from the refer-

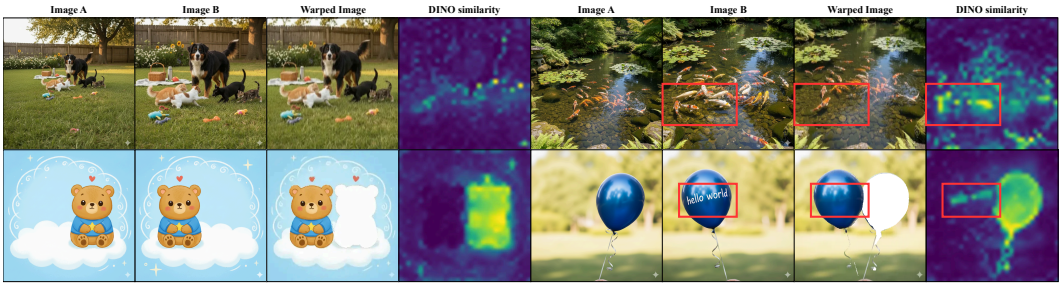

Figure 5: Fine-Grained Identity Assessment. Nano Banana edits Image A into Image B ($A \rightarrow B$). DINOv3 similarity heatmaps compare the warped source with the target within covisible regions, revealing local appearance inconsistencies. Red boxes highlight affected regions, including the altered balloon text.

Table 6: Stage-wise System Construction. We progressively add components and report DINOv3 similarity ($\times 100$) on IEG-Bench. The checkmark (✓) indicates the component is included. FLUX Init: FLUX generative initialization; DINOv3: DINOv3 semantic features; Video & Blender: Video & Blender datasets; Refine: Teacher-guided iterative refinement.

ID	FLUX Init	DINOv3	Video & Blender	Refine	DINOv3 ↑
1					67.31
2	✓				69.28
3	✓	✓			72.76
4	✓	✓	✓		76.52
5	✓	✓	✓	✓	**80.01**

ence after alignment, FreeMatching provides spatial feedback for assessing identity preservation in generative editing.

4.5 Stage-wise System Construction

Table 6 reports the cumulative construction of FreeMatching. ID1 is a randomly initialized base DiT trained on classical supervised datasets with Huber correspondence loss and an ℓ_2 mask loss; it uses neither foundation-model initialization nor DINOv3 inputs, Video/Blender data, or refinement. Subsequent rows add FLUX initialization, DINOv3 features, Video/Blender supervision, and teacher-guided iterative refinement in sequence. DINOv3 reconstruction similarity increases across these stages, measuring the cumulative improvement in feature alignment. The results show that:

1. FLUX initialization and DINOv3 features each improve matching in the cumulative construction, supporting the use of generative and semantic foundation representations.
2. Adding our video and Blender supervision further improves matching on IEG-Bench, extending the capabilities learned from classical correspondence datasets.
3. Teacher-guided iterative refinement further improves matching on IEG, complementing the gains from foundation-model representations and heterogeneous supervision.

Refinement across domains. Table 10 compares Stage 1 with refinement on unlabeled flow-domain pairs (10k-Flow) and IEG pairs (10k-IEG). Both improve IEG-Bench Ref-MSE over Stage 1. The gains depend on the refinement domain: 10k-Flow performs best on Sintel Clean, while 10k-IEG performs best on the remaining reported benchmarks. Compared with Stage 1, 10k-IEG also lowers EPE on all five classical benchmarks, showing that the extension to IEG retains classical matching capabilities.

5 Discussion

In this work, we introduced FreeMatching, a generalizable framework that extends dense correspondence matching beyond classical geometric and motion settings to IEG. By combining foundation-

model representations, heterogeneous correspondence supervision, and teacher-guided iterative refinement, a single model establishes identity-preserving correspondences across diverse visual transformations while retaining competitive performance on classical benchmarks. This broader capability also enables spatially resolved assessment of identity preservation in IEG. We envision a mutually beneficial future where generative models provide diverse data for correspondence learning, while robust correspondence serves as a metric for guiding and evaluating image editing and reference-guided generation.

Generative Ambiguity Generative tasks can also introduce ambiguous correspondences, for example through object replication. In Appendix F, we show how our framework can be extended to a probabilistic diffusion formulation for modeling such ambiguities. This formulation provides a route to representing multiple plausible matches, while the experiments in this paper use the efficient single-pass regression model.

Semantic Matching Semantic correspondence methods such as SD-DINO [47] align semantically related image regions, including across different instances of a category. Our objective instead emphasizes preserving the identity and local details of the same instance across IEG image pairs. The qualitative comparison in Appendix G illustrates how FreeMatching preserves object shape and appearance more consistently in the displayed multi-object example.

Limitations Despite these advances, limitations remain. First, the model can be challenged by extreme cases, such as severe occlusions, topological changes, or drastic non-rigid deformations. Second, its performance ceiling is intrinsically tied to the capabilities of the underlying pre-trained foundation models. Finally, our IEG-Bench, while a valuable starting point, would benefit from expansion to cover more diverse categories. Addressing these challenges presents exciting avenues for future research.

References

[1] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17 (1-3):185–203, 1981.

[2] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[3] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3): 24, 2009.

[5] Johan Edstedt, Qiyu Sun, Georg Bokman, Maarten Wadenback, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024.

[6] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. URL `https://haofeixu.github.io/unimatch/`.

[7] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. UFM: A simple path towards unified dense correspondence with flow. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL `https://proceedings.neurips.cc/paper_files/paper/2025/hash/9d89448b63ce1e2e8dc7af72c984c196-Abstract-Conference.html`.

[8] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025.

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

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

[11] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992.

[12] 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.

[13] Black Forest Labs. FLUX.2 [klein] 4B Base. Model card, 2026. URL `https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B`.

[14] 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.

[15] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024.

[16] Adam W Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, et al. Alltracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5253–5262, 2025.

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

[18] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

[19] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[21] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[22] 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.

[23] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3076–3085, 2024. URL `https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Telling_Left_from_Right_Identifying_Geometry-Aware_Semantic_Correspondence_CVPR_2024_paper.html`.

[24] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19521–19530, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mariotti_Improving_Semantic_Correspondence_with_Viewpoint-Guided_Spherical_Maps_CVPR_2024_paper.html.

[25] Regine Hartwig, Dominik Muhle, Riccardo Marin, and Daniel Cremers. GECO: Geometrically consistent embedding with lightspeed inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Hartwig_GECO_Geometrically_Consistent_Embedding_with_Lightspeed_Inference_ICCV_2025_paper.pdf.

[26] Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of diffusion features for semantic correspondence. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/WACV2025/papers/Fundel_Distillation_of_Diffusion_Features_for_Semantic_Correspondence_WACV_2025_paper.pdf.

[27] Octave Mariotti, Zhipeng Du, Yash Bhalgat, Oisin Mac Aodha, and Hakan Bilen. Jamais vu: Exposing the generalization gap in supervised semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/17826a22eb8b58494dfdfca61e772c39-Abstract-Conference.html.

[28] 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.

[29] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

[30] Songlin Yang, Tianyi Wei, Yushi Lan, Zeqi Xiao, Anyi Rao, and Xingang Pan. Towards geometry-grounded dense semantic matching with VGGT priors. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2509.21263.

[31] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8770–8777, 2019.

[32] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024.

[33] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Yeshwanth_ScanNet_A_High-Fidelity_Dataset_of_3D_Indoor_Scenes_ICCV_2023_paper.pdf.

[34] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mehl_Spring_A_High-Resolution_High-Detail_Dataset_and_Benchmark_for_Scene_Flow_CVPR_2023_paper.html.

[35] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.10127.

[36] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. URL https://arxiv.org/abs/2307.15055.

[37] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015. URL https://arxiv.org/abs/1504.06852.

[38] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13229–13239, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Karaev_DynamicStereo_Consistent_Dynamic_Depth_From_Stereo_Videos_CVPR_2023_paper.html.

[39] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[40] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[42] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[43] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[44] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Menze_Object_Scene_Flow_2015_CVPR_paper.html.

[45] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[46] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.pdf.

[47] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36: 45533–45547, 2023.

[48] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

[49] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Appendix

A Architecture

In this section, we provide the detailed architectural specifications. Our framework leverages the generative prior of a Diffusion Transformer (DiT) and the semantic robustness of DINOv3 to predict dense correspondence and covisibility masks.

A.1 Architecture Overview

Backbone. We initialize the generative transformer from `FLUX.2-klein-base-4B` [13], an undistilled rectified-flow model. We retain its latent-token input projection and add semantic feature injection and a task-specific output decoder for correspondence prediction.

Image Compression (VAE). The frozen FLUX VAE deterministically encodes each image $I \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $z \in \mathbb{R}^{h\times w\times 32}$, where $h = H/8$ and $w = W/8$. We pack each 2×2 latent neighborhood into a 128-dimensional token and normalize it using the VAE's stored statistics. At input resolution $H \times W = 288 \times 512$, each image yields a 36×64 latent grid and $18 \times 32 = 576$ packed tokens.

Semantic Feature Extraction. A frozen DINOv3 ViT-L/16 encoder extracts 1024-dimensional patch features at the same $H/16 \times W/16$ spatial grid. We discard the class and register tokens, then use a zero-initialized linear layer to project the patch features to the transformer's hidden dimension. These features are added to the corresponding projected latent tokens.

A.2 Input Representation and Conditioning

The source and target token sequences are concatenated along the sequence dimension, yielding 1152 image tokens at our input resolution. FLUX position encodings retain both spatial coordinates and image identity, enabling attention across the two views. We use cached empty-prompt embeddings for the text-conditioning path; correspondence prediction requires only the image pair.

A.3 Prediction and Output Decoding

The transformer predicts output tokens in a single forward pass. We select the token sequence associated with the correspondence query image, reverse the latent normalization, and unpack the tokens into a 32-channel latent grid. A task-specific convolutional decoder with progressive upsampling produces a full-resolution three-channel output: two channels for correspondence coordinates and one for covisibility. The predicted coordinates are converted to pixel coordinates to obtain $C \in \mathbb{R}^{H\times W\times 2}$, with the remaining channel providing $M \in \mathbb{R}^{H\times W\times 1}$. The parameter-update scope for both training stages is specified in Appendix C.4.

Table 8 reports measured inference latency and peak allocated memory under a common external input resolution on an H20 GPU. Improving inference efficiency remains a direction for future work.

B Datasets

We present a comprehensive overview of the data strategy employed for pre-training and evaluation, combining established benchmarks with novel large-scale datasets to ensure both robustness and generalization. Some visual examples are provided in Figure 6.

B.1 Standard Datasets

Our supervised mixture includes BlendedMVS [35] and ScanNet++ [33] for geometric correspondence, FlyingThings [18], Spring [34], and FlyingChairs [37] for optical flow, and PointOdyssey [36] and DynamicReplica [38] for synthetic video correspondence. We follow the relevant preprocessing protocols in UFM [7]. These sources complement our tracked-video and Blender data. Following Table 1, the approximate sampling proportions are 10% BlendedMVS,



Figure 6: Overview of the constructed datasets. We utilize Real-World Video data for learning natural deformations, Synthetic Blender data for precise geometric supervision, and the IEG benchmark for evaluating identity-preserving correspondence between source or reference images and generated outputs.

15% ScanNet++, 10% FlyingThings, 5% Spring, 5% PointOdyssey, 10% DynamicReplica, 5% FlyingChairs, 20% Video, and 20% Blender. These proportions describe sampling frequency rather than dataset size and are rounded to 5-percentage-point increments.

B.2 Real-World Video Dataset

To capture real-world non-rigid deformations, we annotate 520k clips from the ViPE [39] dataset. We generate dense pseudo-ground truth annotations using the AllTracker [16] model. By randomly sampling frame pairs from video clips, our training data encompasses diverse motion patterns and natural occlusions that static datasets typically lack. Due to the data volume and copyright restrictions on the source videos, we do not distribute these videos or their precomputed trajectory annotations. We provide data-loading and trajectory-format conversion code, together with preparation instructions for videos that users obtain and are authorized to use.

Specifically, we utilize the official AllTracker checkpoints from `https://github.com/aharley/alltracker`. To handle long-range dependencies, we employ a sliding-window inference mode with a window length of $L = 16$ frames. The model performs $N_{\text{iters}} = 4$ iterative updates per step to refine flow estimates, ensuring precise motion propagation. It outputs dense 2D trajectories $\mathbf{P} \in \mathbb{R}^{T\times H\times W\times 2}$ and visibility confidence maps $\mathbf{V} \in \mathbb{R}^{T\times H\times W}$. We utilize these visibility masks to handle occlusions, ensuring that the loss is computed only on valid pixel correspondences.

B.3 Synthetic Blender Dataset

To mitigate the noise inherent in pseudo-labeled real-world data, we construct a synthetic dataset of 360k image pairs using 3D assets from Objaverse [40]. Each scene is composed of randomly selected objects (2–3 per scene) subject to random rigid transformations. By rendering these scenes from varied viewpoints, we obtain precise pixel-level correspondence and covisibility masks. This provides a clean signal for learning occlusion handling without ambiguity.

Data Curation and Preprocessing. To ensure high-quality geometric supervision, we implement a rigorous automated filtering pipeline for the raw Objaverse assets. First, we normalize the scale of all objects such that their bounding box diagonal equals a fixed unit length ($d = 2.0$). To filter out assets with poor geometry (e.g., flat planes or needle-like structures), we enforce an aspect ratio threshold of 3.0 on the bounding box dimensions; objects exceeding this ratio are either discarded or anisotropically scaled to meet the criteria. Furthermore, we compute the ratio between the convex hull area of the object's projected silhouette and its 2D bounding box area. Assets with a ratio below 0.2 are rejected to ensure substantial pixel occupancy and avoid sparse wireframe-like structures.

Dynamic Scene Composition. We simulate challenging dynamic scenarios involving independent rigid body motions. Each scene is initialized with N objects ($N \in \{2, 3\}$) arranged in a randomized layout (linear for pairs, triangular for triplets). The motion generation involves two temporal frames:

1. Initialization (Frame 1): Objects are placed with random spatial jitter and subjected to random 3D rotations (max tilt 30°) and scaling factors $\sigma \in [0.8, 1.2]$.
2. Motion Modeling (Frame 2): To simulate large displacements, we apply independent rigid transformations to each object relative to Frame 1. This includes a rotation delta magnitude sampled uniformly from $[30^\circ, 50^\circ]$ and further scale variation ($0.8\times$ to $1.2\times$) to simulate depth changes.

Adaptive Rendering Pipeline. Unlike static camera setups, we employ an adaptive camera control system. For each scene, the camera tracks the centroid of the collective bounding box of all objects, dynamically adjusting its distance to ensure all targets remain within the viewing frustum. Rendering is performed using the Blender Cycles engine with 128 samples per pixel at a resolution of 640×360. We leverage Blender's compositor nodes to export ground-truth data: optical flow is extracted directly from the motion vector pass and saved as high-precision 32-bit OpenEXR files, while occlusion masks are generated from the alpha channel, providing ambiguity-free supervision for motion learning.

B.4 IEG Dataset

We curate the image editing and reference-guided generation (IEG) dataset based on UNO [8] to further extend the scope of dense correspondence matching.

For UNO data preparation, we annotate shared object regions through semantic concept extraction and mask generation. These object-region annotations are distinct from the correspondence-validity masks used by the EMA-based refinement objective described in Appendix C.

0. Multi-Object Composition. To enhance semantic complexity and prevent overfitting to center-bias, we modify the standard UNO training format. Specifically, we concatenate two distinct reference–generated image pairs into a single composite training instance. This forces the model to learn correspondence across multiple objects and cluttered semantics rather than simple single-object alignment.

1. Shared Concept Extraction. Given a reference–generated image pair (I_a, I_b), we first identify the primary object categories shared between the two images. We employ Qwen3-VL-8B [48] to perform Open-ended Visual Question Answering. We prompt the model with both images and the instruction: "Identify the 1–2 most prominent common objects present in both images and output their class names." This yields a set of text labels $\mathcal{L} = \{l_1, \ldots, l_k\}$ (where $k \in \{1, 2\}$).

2. Union Mask Generation. Using the extracted labels $\mathcal{L}$ as text prompts, we leverage SAM 3 [49] to segment the corresponding regions in both I_a and I_b. Since a single label (e.g., "car") may correspond to multiple instance masks in the scene, we perform a pixel-wise binary union to aggregate all instances associated with $\mathcal{L}$. This results in unified binary masks M_a and M_b, where $M(\mathbf{p}) = 1$ indicates the presence of the target objects and 0 denotes the background.

C Experimental Setup

C.1 Benchmarks and Baselines

Benchmarks Our evaluation spans two major domains of correspondence matching.

- **Classical Correspondence:** We evaluate on a comprehensive set of datasets to ensure generalization. For dense matching, we use the outdoor ETH3D [42] and indoor ScanNet-1500 [17]. For optical flow, we use the synthetic Sintel [43] and the real-world KITTI-2015 [44]. These benchmarks are governed by strong geometric and motion priors.
  For most benchmarks, we report the End-Point Error (EPE) and the percentage of pixels with errors greater than certain thresholds (1px, 2px, 5px). We compute these metrics only on covisible regions and recompute the results ourselves. For relative pose estimation, we follow RoMa's pipeline and report pose-error AUC at 5°, 10°, and 20°. Pose error is the maximum of the angular errors in rotation and translation direction. AUC is the area under the cumulative pose-error curve up to each threshold, normalized by that threshold and multiplied by 100; higher is better. Unreported baseline entries are left blank.
- **Identity-Preserving Correspondence:** For this domain, we leverage the UNO [8] dataset, which features two disjoint subsets, ensuring a strict separation between training and evaluation. We utilize the single-reference subset exclusively for our teacher-guided iterative refinement stage. Conversely, our test set, the **IEG Correspondence Benchmark (IEG-Bench)**, is constructed from approximately 100 challenging image pairs drawn from the held-out multi-reference subset. To evaluate performance, we follow the approach in Section 3.5, using reconstruction consistency to assess correspondence quality.

Baselines We compare against baselines across the correspondence spectrum.

- For classical correspondence, we benchmark against top performers like GMFlow [21], SEA-RAFT [32], UniMatch [6], RoMa, and UFM [7].
- For IEG correspondence, we compare SEA-RAFT, UniMatch, RoMa, and UFM on IEG-Bench. We additionally report a 10k IEG adaptation comparison for FreeMatching, UFM, and RoMa in Table 5.

Our training pipeline consists of supervised pre-training on classical datasets, tracked videos, and synthetic scenes, followed by teacher-guided iterative refinement on IEG pairs.

C.2 Supervised Pre-training

In the first stage, we train the model using a composite loss function that enforces geometric consistency and mask accuracy on the heterogeneous correspondence datasets. Since real-world applications often involve noise, standard ℓ_2 loss can destabilize training. We therefore adopt a robust formulation. The total supervised loss $\mathcal{L}_{\text{sup}}$ is defined as:

$$\mathcal{L}_{\text{sup}} = \lambda_{\text{huber}}\mathcal{L}_{\text{huber}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} \tag{3}$$

where we set $\lambda_{\text{huber}} = 10$ and $\lambda_{\text{mask}} = 0.2$. The implementation uses the coordinate encoding and per-image normalization detailed below.

Robust Correspondence Loss ($\mathcal{L}_{\text{huber}}$) For training, the absolute source coordinates $C = (u, v)$ are encoded as $\widetilde{C} = (2u/(W-1) - 1,\ 2v/(H-1) - 1)$. Let V denote the ground-truth-derived mask used for correspondence supervision. For each image pair, the supervised pre-training loss sums the two coordinate components and normalizes by the number of valid pixels:

$$\mathcal{L}_{\text{huber}} = \frac{\sum_{x,y} V(x,y) \sum_{d=1}^{2} \mathcal{L}_{\text{H}}\big(\widetilde{C}_{\text{pred},d}(x,y), \widetilde{C}_{\text{gt},d}(x,y); \delta\big)}{\sum_{x,y} V(x,y) + \epsilon}. \tag{4}$$

The batch loss averages these per-pair values. During Stage 2, correspondence losses additionally average the two coordinate components, giving a denominator of $2\sum_{x,y} V(x,y) + \epsilon$ for the component-wise sum.

We adapt the Huber threshold using an exponential moving average e of the masked L1 error, with the same pixel and coordinate normalization as the correspondence loss. We initialize $e_0 = 0.05$ and update $e_t = 0.998e_{t-1} + 0.002\,\mathrm{MAE}_t$, then set $\delta_t = 2e_t$. Stage 2 maintains separate error averages for supervised and weakly supervised updates.

Mask Loss ($\mathcal{L}_{\text{mask}}$) The task-level covisibility mask is defined in $[0, 1]$, while its training target is encoded as $\widetilde{M}^* = 2M^* - 1$. With raw mask prediction $\widetilde{M}_{\text{pred}}$, the implementation computes

$$\mathcal{L}_{\text{mask}} = \frac{1}{HW} \sum_{x,y} A(x, y) \big\| \widetilde{M}_{\text{pred}}(x, y) - \widetilde{M}^*(x, y) \big\|_2^2, \tag{5}$$

where A selects available mask supervision. Covisibility thresholds are applied after converting the raw prediction back to the $[0, 1]$ convention via $(\widetilde{M}_{\text{pred}} + 1)/2$.

C.3 Teacher-Guided Iterative Refinement

We initialize the student from the supervised checkpoint and refine it using UNO+OmniGen2 image pairs without dense correspondence annotations.

EMA initialization and fixed refinement teacher. An EMA copy of FreeMatching predicts the initial correspondence C_0 and covisibility $\bar{M}$. After each student update, its parameters are updated as $\bar{\theta} \leftarrow \beta\bar{\theta} + (1 - \beta)\theta$, with $\beta = 0.95$. We instantiate the frozen refinement teacher with RoMa [5]. RoMa estimates a correction map S between the EMA-warped source and the target, producing $C_{\text{new}} = C_0 \circ S$. The teacher weights remain fixed, and gradients do not propagate through pseudo-target generation.

Refinement objective. For weakly supervised image pairs, we minimize

$$\mathcal{L}_{\text{weak}} = 10\mathcal{L}_{\text{self}} + 0.2\mathcal{L}_{\text{mask-EMA}}. \tag{6}$$

The validity mask V is the intersection of the RoMa confidence mask (threshold 0.2) and the EMA covisibility mask (threshold 0.7). The correspondence loss is

$$\mathcal{L}_{\text{self}} = \frac{\sum_{x,y} V(x, y)\, \mathcal{L}_{\text{Huber}}\big(\widetilde{C}_\theta(x, y), \text{sg}[\widetilde{C}_{\text{new}}(x, y)]\big)}{\sum_{x,y} V(x, y) + \epsilon}, \tag{7}$$

where $\widetilde{C}$ denotes coordinates normalized to $[-1, 1]$, the Huber loss averages over the two coordinate components, and sg denotes stop-gradient. The mask-consistency term is

$$\mathcal{L}_{\text{mask-EMA}} = \frac{1}{HW} \sum_{x,y} \Big\| \widetilde{M}_\theta(x, y) - \text{sg}[\widetilde{\bar{M}}(x, y)] \Big\|_2^2, \tag{8}$$

with $\widetilde{\bar{M}} = 2\bar{M} - 1$. This term regularizes the student's covisibility prediction toward the EMA output. A batch with no valid pseudo-correspondences contributes no refinement loss. The reported refinement objective uses these correspondence and mask targets, without an additional object-containment loss.

C.4 Implementation Details

Parameter updates. During the selected FreeMatching refinement run, the transformer and semantic projection are updated, while the output decoder is frozen. The foundation feature extractors and RoMa refinement teacher also remain frozen; the EMA copy is updated only by exponential averaging.

The selected FreeMatching checkpoint is the EMA model after 10k total Stage-2 updates, initialized from the supervised checkpoint at step 275k. Refinement combines supervised training with weakly supervised updates on UNO+OmniGen2 pairs. This run uses BF16 mixed precision and a global batch size of 32, with the following settings:

- **Optimizer:** AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$, and weight decay $1e^{-2}$.

- **Learning Rate:** 3×10^{-5}, with 3000 warmup steps followed by a constant learning rate.
- **Timestep Conditioning:** We use a fixed timestep $t = 620$ for single-pass correspondence prediction.
- **Memory:** The training supports gradient checkpointing to reduce GPU memory usage.

Original training cost. The original training configuration reported in the rebuttal uses 150k Stage-1 updates followed by 10k Stage-2 updates on 16 A100-40GB GPUs. Stage 1 takes approximately 2 days (768 A100 GPU-hours), and Stage 2 takes approximately 5 hours (80 A100 GPU-hours). These costs describe that training configuration, rather than measurements of the selected 275k-initialized refinement run.

Baseline adaptation cost. The 10k IEG adaptation experiments use 16 H20 GPUs and a global batch size of 32, with external image resolution 288×512. Table 7 reports their approximate training costs. These runs correspond to the additional adaptation comparison in Table 5 and are distinct from the original A100 training above.

Table 7: Training cost of 10k IEG adaptation on 16 H20 GPUs with global batch size 32. GPU-hours are approximate totals across the training GPUs.

Method	Training cost (H20 GPU-hours)
FreeMatching	42.0
UFM [7]	16.8
RoMa [5]	30.8

Inference profiling. Table 8 reports a separate inference profile on a single H20 GPU at external input resolution 288×512 and batch size 1. We use 10 warm-up calls followed by 50 timed calls to each model's inference wrapper, with CUDA synchronization around the timed loop. We report mean latency per image pair and peak allocated GPU memory.

Table 8: Inference cost on one H20 GPU at external resolution 288×512 and batch size 1. Latency is averaged over 50 calls after warm-up; memory is peak allocated GPU memory.

Method	Latency (ms/pair)	Peak memory (GiB)
FreeMatching	121.9	15.18
UFM [7]	51.3	2.35
RoMa [5]	102.5	2.58

D Objective Evaluation

In this section, we provide a detailed formulation of the reconstruction-based evaluation protocol used in our experiments. As outlined in the main text, our goal is to quantify the fidelity of the warped source image $\hat{I}_b = \mathcal{W}(I_a, C)$ against the ground-truth target image I_b, restricted to the covisible region defined by the binary mask M.

To ensure a holistic assessment covering pixel-level accuracy, perceptual quality, and semantic identity preservation, we report four complementary metrics separately, supplemented by the reference-mask, full-image, and alternative-feature evaluations in Appendix D.5.

D.1 Pixel-Level Fidelity: MSE

To measure low-level structural alignment, we compute the Mean Squared Error (MSE) within the masked region. Let $\Omega = \{p \mid M(p) = 1\}$ denote the set of valid pixels in the covisible area. The MSE is defined as:

$$\text{MSE}(\hat{I}_b, I_b, M) = \frac{1}{3|\Omega|} \sum_{p \in \Omega} \|\hat{I}_b(p) - I_b(p)\|_2^2, \tag{9}$$

where pixel values are normalized to the range $[0, 1]$ and the factor of 3 averages over RGB channels. In our reported tables, we scale this value by 100 for readability. Lower MSE indicates tighter spatial alignment of edges and textures.

D.2 Perceptual Similarity: LPIPS

Pixel-wise metrics like MSE are often overly sensitive to slight misalignments or high-frequency noise that do not affect human perception. To capture perceptual similarity, we utilize the Learned Perceptual Image Patch Similarity (LPIPS) metric [41].

We employ the AlexNet-based backbone. Unlike the standard global LPIPS, we compute the **spatial** LPIPS map, $\mathcal{L}_{\text{map}} \in \mathbb{R}^{H' \times W'}$, to strictly evaluate the covisible region:

$$\text{LPIPS}_{\text{masked}} = \frac{1}{|\Omega'|} \sum_{p' \in \Omega'} \mathcal{L}_{\text{map}}(p'), \tag{10}$$

where Ω' represents the mask M downsampled to match the feature map resolution of the LPIPS network. We report LPIPS $\times 100$. Lower scores indicate better perceptual reconstruction.

D.3 Semantic and Identity Alignment: DINOv3 & SigLIP 2

To evaluate whether the correspondence preserves the semantic identity and high-level details of the subject (crucial for our "Identity-Preserving Consistency" protocol), we employ deep feature similarity metrics using state-of-the-art vision foundation models.

DINOv3 Score. We utilize the DINOv3 model [14]. We extract the patch-level features from the last hidden state. To avoid artifacts from register tokens, we discard the first 5 tokens (CLS + registers) and reshape the remaining sequence into a spatial grid. The score is calculated as the cosine similarity between the warped image features $F_{\hat{I}_b}$ and target features F_{I_b} averaged over the valid mask patches:

$$\text{Sim}_{\text{DINO}} = \frac{1}{|P|} \sum_{k \in P} \text{Cos}(F^{(k)}_{\hat{I}_b}, F^{(k)}_{I_b}), \tag{11}$$

where P is the set of patches overlapping with the mask M.

SigLIP 2 Score. Complementary to DINO, we use SigLIP 2 Large [45] to capture multimodal semantic alignment. Similar to the DINO protocol, we extract the dense feature map from the vision tower's last hidden state, normalize the feature vectors, and compute the average cosine similarity over the masked region.

Higher DINOv3 and SigLIP 2 similarities ($\times 100$) indicate closer feature alignment between the warped source and target. We interpret these scores together with the pixel-level and perceptual metrics, assessing whether the matching improvements are consistent across metrics.

D.4 Implementation Details

All metrics are implemented in PyTorch. For feature extraction (DINOv3, SigLIP 2), input images are resized to the model's native resolution (e.g., derived from patch size constraints) using bilinear interpolation, and masks are resized using nearest-neighbor interpolation.

- **Masking Strategy:** A pixel/patch is considered valid if the mask value > 0.5 (or > 128 in 8-bit depth).
- **Normalization:** LPIPS inputs are normalized to $[-1, 1]$. DINOv3 and SigLIP 2 inputs follow the specific mean/std normalization required by their respective processors.
- **Evaluation Subset (IEG-Bench):** The quantitative results in Table 4 are computed on a curated subset of approximately 100 image pairs selected for their geometric complexity and diverse content, providing a stress test for matching algorithms.

D.5 Complementary Metrics and Paired Statistics

Reference-mask provenance. The reference masks are AI-assisted human annotations. FreeMatching predictions provide initial candidates, which annotators manually screen, annotate, and verify before evaluation. The finalized masks are then frozen and applied identically to all methods, rather than recomputed from each evaluated model's predictions. Full-MSE and whole-image OpenCLIP are evaluated without these masks.

For the 10k adaptation comparison in Table 5, each method is evaluated on the same image pairs. Ref-MSE uses a shared frozen reference mask in the RGB MSE formula above; Full-MSE replaces that mask with the full image. Ref-DINOv2 uses DINOv2 ViT-L/14 [22] patch features at 280×504 resolution. We average patch cosine similarities with weights given by reference-mask coverage, retaining patches with coverage at least 0.25. OpenCLIP [46] uses the LAION-2B-pretrained ViT-H/14 image encoder on 224×224 images and computes the cosine similarity between global image embeddings without masking. Both similarities are multiplied by 100. These metrics evaluate alignment with feature encoders distinct from the DINOv3 features used by FreeMatching.

We compute paired Full-MSE differences on the common evaluation pairs from IEG-Bench, sampling pairs with replacement for 10,000 bootstrap replicates (seed 20260728). The 2.5th and 97.5th percentiles give the reported 95% confidence intervals. Positive advantages denote lower error for FreeMatching. Its image-level win rates on Full-MSE, Ref-DINOv2, and OpenCLIP are 68%, 95%, and 91% against UFM, and 73%, 78%, and 87% against RoMa.

D.6 Weak Adaptation and Refinement-Domain Controls

Table 9 reports classical-task EPE for the three 10k-IEG-adapted methods. Table 10 compares FreeMatching-S1 with two refinement data sources: previously unseen, unlabeled Spring/FlyingThings pairs for 10k-Flow and IEG pairs for 10k-IEG. The flow-domain control uses no ground-truth flow during refinement. Each row reports one fixed checkpoint across the benchmarks.

Table 9: Classical-task EPE after 10k IEG adaptation. Lower is better; best results within this comparison are bold.

Method	Sintel Clean	Sintel Final	KITTI	ScanNet	ETH3D
FreeMatching (10k-IEG)	1.05	**1.36**	4.95	**8.56**	**10.23**
UFM (10k-IEG) [7]	**0.92**	1.48	**4.75**	12.03	12.98
RoMa (10k-IEG) [5]	1.24	2.49	5.91	12.03	11.26

Table 10: Effect of the refinement data source. We report IEG-Bench Ref-MSE ($\times 100$) and EPE on classical benchmarks. Lower is better; best results within this comparison are bold.

Checkpoint	Ref-MSE	Sintel Clean	Sintel Final	KITTI	ScanNet	ETH3D
FreeMatching-S1	8.08	1.11	1.73	6.26	8.85	11.68
FreeMatching (10k-Flow)	7.26	**0.98**	1.59	6.46	8.91	11.64
FreeMatching (10k-IEG)	**6.22**	1.05	**1.36**	**4.95**	**8.56**	**10.23**

E Human Evaluation

To rigorously assess the perceptual quality of long-range correspondence and the effectiveness of our reconstruction consistency metric, we conducted a user study on the IEG-Bench dataset. While objective metrics provide scalable quantitative insights, human judgment remains the gold standard for evaluating semantic alignment and visual consistency. We compared FreeMatching with three representative baselines and recruited ten independent evaluators to rate the results blindly.

Protocol and Criteria. The evaluation followed a blind protocol where method identities were anonymized and display orders were randomized. Evaluators were instructed to rate the alignment quality based on semantic consistency and texture preservation using a discrete 4-point scale:

Table 11: **Quantitative comparison on IEG-Bench.** We report the Human Mean Opinion Score (MOS) and objective metrics. **Human (MOS)** is rated on a scale of 0-3 (higher is better). LPIPS [41] and DINOv3 [14] are scaled by 100. Our method achieves the highest ratings across all metrics. Best results are highlighted in **bold**.

Model	Human ↑	LPIPS ↓	DINOv3 ↑
SEA-RAFT [32]	0.49	72.45	58.14
RoMa [5]	1.39	48.91	70.15
UFM [7]	1.13	53.05	66.77
FreeMatching (Ours)	**2.25**	**45.41**	**80.01**

- **0 (Failure)** for significant misalignment or severe distortion;
- **1 (Poor)** for visible errors affecting the main subject;
- **2 (Good)** for generally correct alignment with minor artifacts;
- **3 (Perfect)** for high-fidelity reconstruction indistinguishable from the target context.

Results and Analysis. As reported in Tab. 11, FreeMatching achieves the highest MOS of **2.25**, significantly outperforming the strongest baseline, RoMa (1.39), and demonstrating substantial improvement over dense optical flow methods like SEA-RAFT (0.49).

Furthermore, to validate the reliability of our automated evaluation pipeline, we analyzed the correlation between human ratings and objective DINOv3 scores. We observe a strong positive Pearson correlation coefficient ($r = 0.738$). This high correlation indicates that the DINOv3 metric aligns closely with human perception of correspondence quality, suggesting that our objective metrics serve as a reliable and efficient proxy for labor-intensive human evaluation in long-range alignment tasks.

F Beyond One-to-One Correspondence

In IEG tasks involving object replication or synthesis, a query may admit multiple plausible correspondences. We outline a probabilistic extension of our FLUX-based framework to model the conditional distribution $p(C|I_a, I_b)$, providing a formulation for representing this ambiguity.

Training The extension replaces single-step regression with an iterative denoising objective. Given correspondence targets C_{gt}, we define noisy states C_t according to a noise schedule. The model f_θ, conditioned on image features and time t, can then be trained to recover clean coordinates directly (x-prediction):

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\mathcal{L}_{\text{Huber}}(f_\theta(I_a, I_b, C_t, t), C_{gt}) \tag{12}$$

Conditioning on the noisy correspondence state provides a mechanism for representing different matching hypotheses for the same image pair.

Sampling The corresponding sampling procedure starts from $C_T \sim \mathcal{N}(0, I)$ and iteratively refines the predicted clean correspondence. Different initial noise samples would allow the model to represent alternative matching hypotheses for the same image pair. Evaluating the validity and diversity of these hypotheses requires a dedicated experimental study.

This appendix describes a modeling extension; all reported experiments use the efficient single-pass regression model.

G Semantic Matching

Semantic correspondence commonly studies alignment across different instances of a category, whereas our task emphasizes the identity and local details of the same instance across IEG image pairs. SD-DINO [47] combines Stable Diffusion and DINOv2 features for zero-shot semantic correspondence.



Figure 7: Comparison with Semantic Matching. In this multi-object example, the displayed SD-DINO [47] reconstructions, obtained with the prompts “book” and “vase”, exhibit incomplete shapes and mixed object appearance. FreeMatching preserves more complete object shapes and local details without text prompts.

Figure 7 compares the methods on a multi-object IEG example. The displayed SD-DINO results use the prompts “book” and “vase” in our evaluation configuration. Its reconstructions omit parts of the objects and mix their appearance, while FreeMatching preserves more complete object shapes and local details using only the image pair.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: Yes. The abstract and introduction summarize the FreeMatching framework, its two-stage training strategy, and the intended evaluation scope. The claims are supported by quantitative results and ablations in Sections 4, 4.2 and 4.5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The discussion includes a limitations paragraph describing challenging cases such as severe occlusion, topological changes, drastic non-rigid deformation, dependence on foundation models, and the limited category coverage of IEG-Bench.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper is empirical and does not present formal theoretical results requiring assumptions and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Yes. The paper provides the model architecture, datasets, training objectives, evaluation benchmarks, and metric definitions, with implementation details in Appendices A to D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper links to a public project repository and describes the implementation and data construction pipeline. The repository currently contains introductory material; runnable code and exact reproduction commands have not yet been released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Yes. The paper specifies benchmark splits, baselines, evaluation metrics, training losses, optimizer, learning rate, warmup, mixed precision, fixed timestep conditioning, and hardware in Appendices B to D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Yes. The paper reports paired bootstrap 95% confidence intervals for Full-MSE differences, with the resampling unit and procedure specified in Appendix D.5. Human-evaluation correlations are reported in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Yes. The paper distinguishes the original training on 16 A100 GPUs from adaptation on 16 H20 GPUs, and reports training costs, inference latency, and peak GPU memory in Appendix C.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses established datasets and foundation models with citation, evaluates correspondence rather than generating harmful content, and reports a low-risk blind rating study without collecting sensitive personal information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No].

    Justification: The paper discusses positive uses for evaluating identity preservation in IEG, but it does not currently include a balanced broader-impact discussion covering possible negative uses or mitigations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a high-risk generative model or language model; the proposed system is a dense correspondence model and evaluation metric. Any future release should still document data filtering and usage restrictions if real-world images are included.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: Yes. The paper credits the datasets, models, and codebases used, including AllTracker, UNO, Objaverse, FLUX.2-klein-base-4B, DINOv3, SAM, and Qwen, and follows the corresponding usage terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: Yes. The real-world video pseudo-label data, synthetic Blender data, IEG processing, and IEG-Bench evaluation subset are documented in Appendix B, including data sources, filtering, rendering, mask generation, and subset construction.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No].

Justification: No. Appendix E reports the blind rating protocol, rating scale, and evaluation criteria, but the paper does not include interface screenshots or compensation details.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No].

    Justification: The paper includes a human evaluation, but it does not currently describe participant risks, risk disclosure, or IRB/equivalent review status.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: Yes. The IEG data pipeline in Appendix B describes the use of Qwen3-VL-8B for open-ended visual question answering before SAM-based mask generation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.