Zero-shot optical flow extraction using Local Random Access Sequence (LRAS) Models

Seungwoo (Simon) Kim
Department of Computer Science
Stanford University

Khai Loong Aw
Department of Computer Science
Stanford University

Klemen Kotar
Department of Computer Science
Stanford University

Daniel LK Yamins
Department of Computer Science
Stanford University

1 Introduction

Extracting optical flow from videos remains a core computer vision problem. While specialized models exist [14, 12, 11], they are often trained on synthetic data due to the high cost of labeling, or self-supervised to satisfy training heuristics such as feature/photometric similarity. As a result, both approaches struggle on real-world scenes containing complex dynamics due to an inherent domain gap or mismatch of inductive biases.

A recent paradigm in both language and vision is to extract intermediates from large-scale pretrained models that can implicitly learn features from their extensive training data which are useful for various downstream tasks. Specifically in computer vision, prior work has shown that visual structure such as depth [10] and illumination [4] can be successfully extracted from large-scale video models. These results suggest that generative video models have a strong understanding of scenes, and offer a promising direction of using such models for optical flow. However, unlike in language where a natural inference-time extraction procedure exists in the form of textual prompts, extracting different visual properties from a video model often requires additional task-specific fine-tuning. For the purpose of optical flow, this brings us back to the challenge of data sparsity: obtaining dense flow labels from real-world videos is prohibitively expensive.

Therefore, we are interested in finding a zero-shot procedure for extracting optical flow from generative video models that does not require any additional fine-tuning. Such a procedure would allow us to fully take advantage of the strong scene understanding capabilities of video models, while bypassing any fine-tuning or learned readouts that would likely require training on synthetic data.

We draw inspiration from the Counterfactual World Modeling (CWM) paradigm [1], which proposes a zero-shot, visual prompting procedure for extracting optical flow from a next frame predictor by placing a perturbation on the input image and tracking how that signal propagates to the next frame. In particular, we identify the Local Random Access Sequence (LRAS) [7] model as a large-scale video model with key properties: (1) locality of patch encodings and (2) the ability to decode the next frame in any order, that make it particularly amenable to this CWM-like procedure. Finally, we propose an enhanced KL-divergence-based procedure that directly utilizes the fact that LRAS is capable of predicting *entire distributions* of plausible next-frame patches. Combined, these improvements lead to a generalized, zero-shot flow extraction method that is robust against challenging, in-the-wild scenes, achieving state-of-the-art results compared to specialized baseline models on both TAP-Vid DAVIS, consisting of real-world videos, and TAP-Vid Kubric, which is synthetic.

2 Related Work

2.1 Specialized Models for Optical Flow and Point Tracking

Optical flow quantifies the per-pixel 2D motion between a pair of frames. Most existing models, trained specifically for this task, are supervised on synthetic video datasets with ground-truth dense flow labels, e.g., Sintel [3], FlyingChairs [5], and FlyingThings [8]. However, such a training scheme is vulnerable to a domain gap that prevents generalization to real-world videos. The wide variety of challenging motion and scene appearances that can appear in real-world videos are not fully simulated in synthetic datasets. An alternative approach has been to develop self-supervised flow models [11, 6] that can train on real-world videos. However, such models often rely on task-specific architectures and heuristics such as photometric consistency or smoothness. These hard-coded assumptions about motion can breakdown in more complex (and thereby more realistic) scenes, e.g., in the presence of sudden changes in light intensity or large motion.

2.2 Generative Video Models

On the other hand, large-scale video models are trained on diverse real-world data, without strong assumptions of the underlying scene dynamics. State-of-the-art models such as Cosmos [9] or Stable Video Diffusion [2] are able to produce photorealistic next frame predictions, suggesting a deep understanding of motion and scene transformations that emerges during pretraining. However, it is non-trivial to utilize this property to extract useful visual structure for downstream tasks. In this paper, we identify a zero-shot flow extraction procedure that can be applied to a class of generative video models known as Local Random Access Sequence (LRAS) models [7] to harness the power of large-scale training for state-of-the-art optical flow.

3 Methods

3.1 Test-time inference procedure for extracting optical flow from video world models

Our base procedure of extracting optical flow from a next frame predictor has already been proposed in the Counterfactual World Model (CWM) paper [1].

1. **Step 1. Inject a small perturbation.** Given duplicate initial frames F_1 and $\tilde{F}_1$, we perturb $\tilde{F}_1$ with a small, "white" Gaussian centered at the query location x_q (i.e., amplitude 255 on each RGB channel, and a small fixed standard deviation σ).
2. **Step 2. Clean and Perturbed Predictions.** We pass the clean initial frame F_1 and the perturbed initial frame $\tilde{F}_1$ through the model, each with a sparse mask of the second frame ($F_2^{\texttt{masked}}$). From this, we get two reconstructions of the second frame: $(F_1, F_2^{\texttt{masked}} \to \hat{F}_2)$ and $(\tilde{F}_1, F_2^{\texttt{masked}} \to \hat{\tilde{F}}_2)$. In the ideal case, the model will push the perturbation forward into the next frame in $\hat{\tilde{F}}_2$, while no perturbation trace will exist in the clean prediction $\hat{F}_2$.
3. **Estimate optical flow.** Finally, we can compute $\arg\max \|\hat{\tilde{F}}_2 - \hat{F}_2\|$ in the RGB space to find the location x_t to which the perturbation was carried, as that should be the only difference between the conditioning of the two predictions.

This process can be generalized for any frame predictor that exposes some kind of interface for sparse next-frame conditioning, e.g. certain classes of generative video models. Note that sparse conditioning is necessary – unconditional generation of next frames would yield inconsistent predictions between the two forward passes ($\hat{F}_2$ and $\hat{\tilde{F}}_2$), and the delta will not be constrained to the injected perturbation. Further, if the conditioning is not sparse, i.e. the next frame is entirely revealed, then the model cannot carry over the perturbation, as the ground-truth (in which the perturbation does not exist) is already leaked to the model.

3.2 Applying this procedure to LRAS

The process described above, while in theory applicable to different classes of generative video models, is particularly suited for the recently proposed LRAS architecture [7]. Not only does LRAS directly expose the sparse-conditioning API, its differentiating properties also contain key desiderata for controllable extraction: **(1) Locality.** It uses a local tokenizer that encodes each spatio-temporal patch indepedent of global context. This is advantageous because a small perturbation added to x_p does not propagate globally, with its effects showing up in the reconstructed next frame in unpredictable ways. In contrast, many video models [9] utilize *global* tokenizers. This may hurt consistency between predictions, which is essential for the CWM-like procedure described above. **(2) Random Access.** By incorporating *pointer tokens*, LRAS can perform random-order decoding of the next frame. This provides a stronger conditioning, as the input sequence can consist of random patches in the next frame instead of following some arbitrary structure, e.g. raster order. This can expose more information of the next frame, leading again to more consistent generation and more informative motion conditioning. **(3) Predicting entire distributions.** Finally, the LRAS model has the capability to produce entire patch-wise distributions (which are subsequently sampled to produce next-frame reconstructions). This distribution can be understood as capturing the entire world of plausible outcomes, instead of regressing to a single state. As described in the next section, we make use of this property to improve upon the CWM-like extraction procedure.

3.3 KL-tracing: Improving the zero-shot extraction procedure

Unlike deterministic regression models such as CWM [1], LRAS has the unique advantage of being able to produce entire distributions of next frame patches. This property allows for an improved version of the flow extraction procedure that detects the perturbation signal not in the RGB-space, but rather in the *distribution* space. Intuitively, the model's distribution will shift to account for the added perturbation. However, this effect may be washed out during sampling if the distribution change is subtle, and the delta may contain sampling noise that is irrelevant to the effects of the perturbation. To mitigate this, we can compute the difference directly in the distribution space using KL-divergence.

1. **Step 1. Inject a small perturbation.** This part is equivalent to the CWM method; we inject a white, Gaussian perturbation to the query location x_p, producing clean and perturbed frame 1, F_1 and $\tilde{F}_1$.
2. **Step 2. Clean and Perturbed Predictions.** Similarly, we perform two forward passes. However, instead of inspecting the reconstructed RGBs $\tilde{F}_2$ and $\hat{\tilde{F}}_2$, we directly inspect the logits for the two predictions, yielding two sets of patch-wise logits: $\{z_{ij}\}$ and $\{\tilde{z}_{ij}\}$.
3. **Step 3. Compute KL-divergence map.** For every patch (i, j), we compute the KL-divergence between the clean and perturbed logits: $D_{\mathrm{KL}}(i, j) = \mathrm{KL}[z_{ij} \| \tilde{z}_{ij}]$. Finally, similar to the procedure applied to the RGB difference map, we apply an argmax to the KL-divergence map to find the target location x_t.

In Figure 1, we demonstrate the effect of using a KL-based procedure compared to an RGB-based approach. Intuitively, for non-deterministic video models such as LRAS, there exists samples such that the perturbation is not properly carried or spurious hallucinations are added to the predictions in regions that are not conditioned. RGB difference requires committing to a single sample, which may contain such noise; by computing divergence in the distribution space, we are performing a more holistic analysis that encompasses all plausible samples.

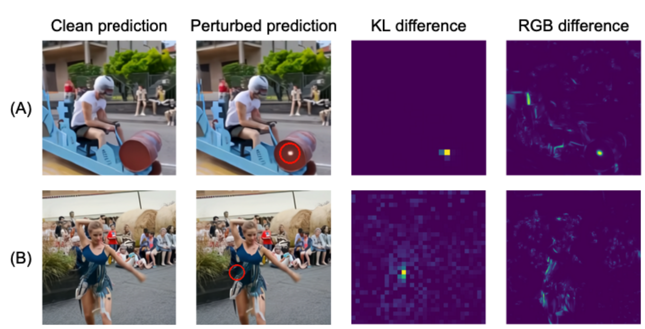

Figure 1: **Using KL divergence of prediction logits is better than using RGB differences.** Using KL-divergence and capturing difference in the distribution space often leads to stronger, and less noisier perturbation signals. This in turn leads to better flow extraction.

4 Results and Discussions

	Method	DAVIS				Kubric			
		AJ ↑	AD ↓	$<\delta^x_{\text{avg}}>$ ↑	OA ↑	AJ ↑	AD ↓	$<\delta^x_{\text{avg}}>$ ↑	OA ↑
TAP-Vid First									
S	RAFT [12]	41.77	25.33	54.37	66.40	71.93	5.60	82.15	88.54
	SEA-RAFT [14]	43.41	20.18	58.69	66.34	75.06	6.54	84.63	89.50
U†	Doduo [6]	23.34	13.41	48.50	47.91	54.98	5.31	72.20	73.56
U	SMURF [11]	30.64	27.28	44.18	59.15	**65.81**	6.81	80.57	**87.91**
	CWM [1, 13]	15.00	23.53	26.30	**76.63**	28.77	11.64	41.63	84.93
	KL-LRAS (ours)	**44.16**	**11.18**	**65.20**	74.58	65.49	**5.06**	**81.66**	87.63

Table 1: **TAP-Vid First: quantitative results on DAVIS and Kubric.** Tracking starts when a point first appears and continues to the video end, so large inter-frame gaps must be bridged. KL-LRAS consistently outperforms both supervised and unsupervised two-frame baselines. For optical flow baselines, occlusion is determined by a cycle consistency threshold. "S" = supervised, "U" = unsupervised; Note that Doduo [6] is weakly supervised.

The results in Table 1 demonstrate the strength of our approach. The zero-shot inference-time flow extraction on LRAS outperforms even specialized flow models that are trained on supervised, albeit synthetic, labels. These results hold across different domains; TAP-Vid DAVIS consists of challening real-world videos with large frame gaps and motion, while TAP-Vid Kubric is a synthetic dataset more in-distribution to the training data of baseline flow models.

While not the main focus of this paper, our approach can naturally extend to compute occlusion. Namely, given the KL-divergence map, we can apply a simple threshold to obtain a binary occlusion signal (e.g. the magnitude of the largest KL divergence value will be different in the case that the perturbation is not properly reconstructed due to it being occluded). For baseline flow models that do not expose an occlusion API, we use cycle-consistency. In particular, occlusion is defined as the event of mismatch between forward and backward flow prediction greater than 6 pixels.

5 Conclusion and Future Work

5.1 Taming LRAS for zero-shot optical flow

We propose a zero-shot procedure to extract optical flow from generative video models. We identify the useful properties of LRAS that make it particularly suitable for this task, and develop a KL-divergence based procedure to apply a CWM-style flow extraction in a more robust fashion. We find that through this procedure, we can effectively harness the deep motion understanding of generative video models that emerge from large-scale pretraining, bridging the domain gap that is present on most specialized flow models trained on synthetic data.

5.2 Future work: Dense Flow via distillation

This paper offers a promising direction for using generative video models for optical flow. The natural next step is to use this method as a labeler to obtain pseudo-labels on real-world videos, and distill into a more efficient architecture for real-time dense flow prediction.

References

[1] Daniel M. Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (Machine) Vision via Counterfactual World Modeling, June 2023. arXiv:2306.01828 [cs].

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets, November 2023. arXiv:2311.15127 [cs].

[3] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, volume 7577, pages 611–625. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. Series Title: Lecture Notes in Computer Science.

[4] Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. Relightvid: Temporal-consistent diffusion model for video relighting. *arXiv preprint arXiv:2501.16330*, 2025.

[5] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning Optical Flow with Convolutional Networks, May 2015. arXiv:1504.06852 [cs].

[6] Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Learning Dense Visual Correspondence from Unsupervised Semantic-Aware Flow, September 2023. arXiv:2309.15110 [cs].

[7] Wanhee Lee, Klemen Kotar, Rahul Mysore Venkatesh, Jared Watrous, Honglin Chen, Khai Loong Aw, and Daniel L. K. Yamins. 3D Scene Understanding Through Local Random Access Sequence Modeling, April 2025. arXiv:2504.03875 [cs].

[8] Nikolaus Mayer, Eddy Ilg, Philip Häusser, 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, June 2016. arXiv:1512.02134 [cs].

[9] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos World Foundation Model Platform for Physical AI, March 2025. arXiv:2501.03575 [cs].

[10] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024.

[11] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-Teaching Multi-Frame Unsupervised RAFT with Full-Image Warping, May 2021. arXiv:2105.07014 [cs].

[12] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow, August 2020. arXiv:2003.12039 [cs].

[13] Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M. Bear, Khaled Jedoui, Klemen Kotar, Felix Binder, Wanhee Lee, Sherry Liu, Kevin A. Smith, Judith E. Fan, and Daniel L. K. Yamins. Counterfactual World Modeling for Physical Dynamics Understanding, December 2023. arXiv:2312.06721 [cs].

[14] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow, May 2024. arXiv:2405.14793 [cs].