Dynamic 3D from Monocular Video: Reality Check

Anonymous Author(s)
Affiliation
Address
email

Figure 1: **Impact of effective multi-view in novel-view synthesis.** Although recent works have shown impressive results on monocular novel-view synthesis of dynamic scenes (left), we find that these evaluation protocols actually operate in a setup that effectively provides a multi-view signal. We propose a method for characterizing the amount of effective multi-view in a video sequence and propose a new dataset which shows that on actual monocular sequences many challenges remain (right).

Abstract

We study the recent progress on inferring dynamic 3D scene representations from a monocular video sequence. We take a closer look and find that although recent approaches demonstrate impressive results, they are evaluated on datasets with a protocol that effectively provide multi-view signals. We propose *effective multi-view factors* (EMF) to quantify the amount of multi-view signal in a sequence based on the magnitude of relative camera and scene motion. We then propose a new dataset of monocular videos with depth data and evaluation protocols based on depth-guided masked PSNR and correspondence. We evaluate representative approaches on this dataset and find that many challenges remain in the monocular setup. We advise future works to report EMF on their input sequences in order to assess the difficulty of the task. We will release our dataset and code including the metric and the dataset generation in hopes that it will be a useful toolbox for future work.

1 Introduction

3D capture of dynamically deforming objects has been a fundamental challenge in computer vision since the scene's motion violates the assumption of a static world in multi-view geometry. However, such dynamic scenes are ubiquitous in our everyday lives – with moving people, purring cats, or perhaps a tree swaying in the wind. Although there have been customized solutions with hardware such as motion capture rigs [1] and lightstages [2, 3], these setups are cumbersome to set up and not easily accessible to ordinary users. Therefore, the ability to 3D capture dynamic sequences in a "casual" setup, such as monocular videos taken by a smartphone in an uncontrolled environment, will be a cornerstone in scaling up 3D content creation, performance capture, and augmented reality.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Recent works have shown promising and impressive results in spatio-temporal novel-view synthesis on monocular sequences with dynamically moving arbitrary objects [4, 5, 6, 7, 8, 9]. However, upon closer inspection, we find that the capture and evaluation setup employed in these works is *not* exactly a monocular capture setup representative of everyday use cases. The inputs to these novel-view synthesis algorithms either depict slow-moving objects with a fast-moving camera or setups where frames at consecutive time steps come from different cameras with a wide baseline on a static or handheld rig. Effectively, both setups result in a training sequence that has a multi-view signal.

In this paper, we analyze this effect and propose an additional set of best practices for the training and evaluation protocols going forward. As our first contribution, we propose "effective multi-view factors" (EMF) to quantify the amount of effective multi-view in a sequence based on the agnitude of the camera motion relative to scene motion present. We analyze that based on EMF, existing methods that train on high-velocity cameras and/or slow-moving objects that are outside the range of a casual capture operate in a regime that are close to a multi-view capture.

Second, we contribute a new dataset of smartphone videos with accompanying depth information with new evaluation metrics in hopes of pushing the boundaries of previous evaluation protocols toward everyday monocular capture. We show that on actual monocular capture setups, many challenges remain for state-of-the-art methods. Our dataset spans a diverse set of dynamic motions, and, in conjunction with our training and evaluation protocol, is closer to the monocular capture setting as quantified by EMF. It also takes advantage of the depth sensors available on recent smartphones, which we show that significantly improves the performance on current algorithms.

One reason for existing evaluation setups is that current methods evaluate performance soley based on the synthesized photorealistic quality (PSNR). However, training on one camera and evaluating with a different camera not seen during training with PSNR can be problematic, since there may be parts of the scene that were never seen in the training view. Existing methods alleviate this issue by incorporating views from the test camera in an alternating manner, which enable PSNR evaluation, at the expense of injecting multi-view signal that confounds the performance of the methods on monocular sequences. Thus, as a third contribution we propose to add the quality of established correspondences over time in the training sequence as a first-class metric for evaluating dynamic view synthesis methods.

Evaluating performance by correspondences has several benefits: first, it enables evaluation within the input video without requiring a held-out view. Second, it evaluates the quality of the inferred underlying geometry and deformation, which cannot be evaluated by appearance alone. It is also robust to dynamic appearance changes due to non-Lambertian surfaces and/or shadows and invariant to changes in imaging settings such as exposure and white balancing. In fact, the combination of good training PSNR and accurate correspondences is a necessary condition of a model that has accurately learned the underlying 3D scene dynamics, which can lead to a good held-out PSNR from a large range of viewpoints. To support this protocol, we provide ground truth correspondence annotations in our dataset and show how to read out correspondences from prior methods that only evaluated with PSNR. We also collect sequences with held-out cameras and show that, when the depth is available, one can still compute held-out PSNR without including the held-out camera views in training, as depth can identify regions that are visible in both cameras.

Finally, we evaluate four representative algorithms on novel-view synthesis of dynamic scenes [5, 6, 7, 8] on the new dataset following the proposed protocol. We find that all methods have a large train *v.s.* test performance gap both in PSNR and correspondence accuracy, indicating a large room for improvement for future works. We end our analysis showing that leveraging the depth information along with a sparsity regularizer yields improved PSNR and correspondence accuracy. We will release the data and the code for computing the metric for capturing new sequences from a smartphone. We believe characterizing the amount of effective multi-view in a sequence is essential for making progress on this challenging problem and we hope our dataset, metric, and evaluation protocols will be a useful toolkit for researchers in the future.

2 Related work

Multi-view Dynamic Capture. Spatio-temporal novel view synthesis of dynamic objects, also known as video view synthesis or free view point video, have a long-standing history in computer

vision and graphics. Major highlights include the pioneering work of Kanade *et al.* [1], which used a 3D dome with 51 cameras to synthesize novel views based on depth maps, Zitnick *et al.* [10], which also relied on depth maps and layers estimated from 8 videos, Einarsson *et al.* [2] who take a light field based rendering approach from a capture in a lightstage with subjects moving on a rotating treadmill, and Collet *et al.* [3] who use 53 RGB and IR camera pair rig to capture high quality geometry and appearance using multi-modal multi-view information. More recently, Broxton *et al.* [11] employ a circular 2D rig to capture a multi-sphere image that they convert to layered mesh, and Li and Slavcheva *et al.* [12] use a time-conditioned neural radiance fields on video recorded by 18 cameras. Both approaches can capture challenging content such as fire. While these approaches have the potential for high-quality dynamic video synthesis, they all require a specialized time-synchronized multi-camera rig, which are not easily accessible for ordinary users.

Monocular nonrigid structure from motion (NR-SfM). Traditional non-rigid structure-from-motion tackles the task of dynamic 3D inference by assuming that the underlying structure going through non-rigid motion is low-rank [13, 14], or by explicitly enforcing this assumption via fitting 3D morphable models when the category of the objects are known [15, 16, 17]. Some approaches learn morphable models from image or video sequences during inference [18, 19, 20, 21, 22]. In contrast, there also exists approaches that are non-parametric in its shape representation that fuse depth scans of generic dynamic scenes in offline [23, 24, 25, 26, 27] or online approaches [28, 29, 30, 31]. All of these approaches aim to recover accurate surface geometry at each time step and their performance is measured with ground truth 3D geometry or 2D correspondences with PCK [32] when such ground truth is not available. In this paper we analyze recent novel-view synthesis methods whose goal is to generate a photo-realistic novel view. Due to its goal, these methods do not allow evaluation using ground truth 3D geometry, but we take inspiration from prior NR-SfM works to evaluate the quality of the inferred 3D dynamic representation based on correspondences. We also draw inspiration from previous NR-SfM work that analyzed camera/object speed and 3D reconstruction quality [33, 34].

Dynamic neural radiance fields (Dynamic NeRFs). Recently, dynamic NeRFs sought to synthesize high-quality novel views of generic dynamic scenes by jointly modeling scene dynamics and view-dependent surface appearance via volume rendering [4, 5, 6, 7, 8, 9]. We note that there has been work to synthesize novel views of moving human bodies [35, 36, 37, 38] and faces [39], but they require a known shape template and 3D pose as initialization. We focus on the aforementioned object-agnostic general scenes. The evaluation protocol in these lines of work inherit from the original static-scene NeRF [40] that quantify the rendering quality of held-out viewpoints using the PSNR and SSIM metrics. However, in dynamic scenes, PSNR from an unseen camera view is not a meaningful metric since the novel view may include parts of the scene that was never seen in the training view[1]. Existing approaches resolve this issue by incorporating views from multiple cameras during training, which result in an effectively multi-view setup. Our work introduces a metric to measure the difficulties of a sequence to validate this, and introduce a monocular dataset with evaluation protocol which shows that existing methods still have a long way to go for monocular capture of dynamic 3D sequences.

3 Existing capture protocol

While tremendous efforts have been spent on developing better methods for novel-view synthesis (NVS) from monocular dynamic sequences [4, 5, 6, 7, 8, 9] , little attention has been paid to the actual protocol on which methods are evaluated. We first introduce the notion of effective multi-view, and review existing monocular capture protocols with this perspective. We then introduce effective multi-view factors (EMF) to quantify this notion and discuss the motivations behind the existing capture designs.

3.1 Existing protocols under the lens of effective multi-view

We first define the notion of "effective multi-view". A monocular dynamic capture setup consists of a single camera observing a moving scene. This is in contrast to a multi-view capture where the scene is captured by multiple cameras simultaneously. We propose to characterize the amount of "effective

[1] unless the method can infer unseen regions using learning based approaches.

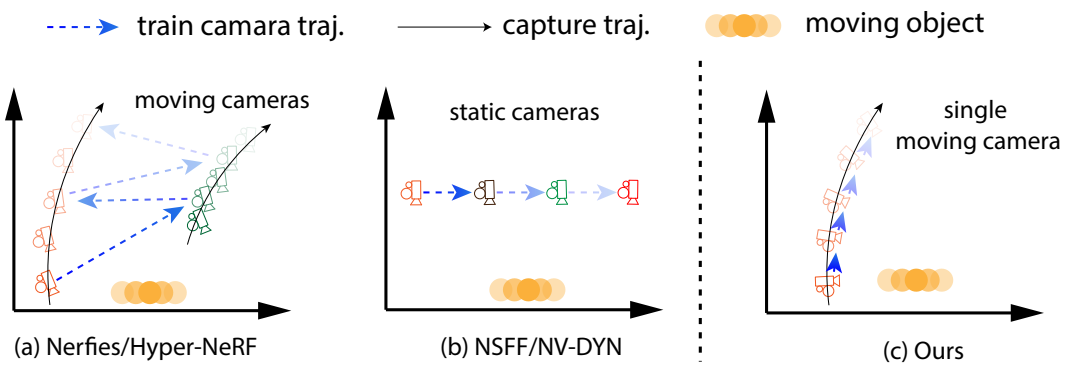

Figure 2: **Illustrations of existing and proposed capture setups in 2D.** (a) Nerfies and HyperNeRF employ two moving cameras to record the moving scene, while (b) NSFF (building upon NV-DYN [41]) employs twelve static cameras. These protocols then *simulate* a monocular train camera trajectory by interleaving frames from each of the cameras, resulting in large baseline camera motion that is not possible to capture with a single handheld camera and is effectively multi-view. (c) On the other hand, we propose a strictly monocular capture setup, using a single camera for capturing a training sequence.

multi-view" using the magnitude of the scene and camera motion, where when the scene motion is significantly smaller than that of the camera, a monocular capture is effectively multi-view. To see this, consider an extreme case when the scene motion has zero speed, *i.e.*, a static scene. In this case, a capture from a single camera is equivalent to a multi-view setup. Similarly, consider an extreme case where the cameras move very fast (without incurring motion blur) relative to the scene motion so much so that it can see the exact same scene from different view points, again reducing the monocular case to the multi-view case. Next we review capture protocols from two existing works under this lens.

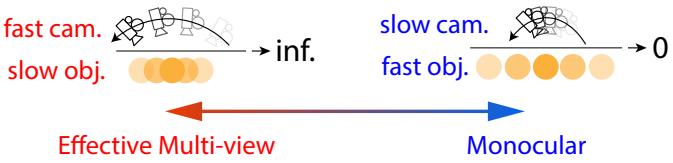

Figure 3: **Spectrum of monocular capture.** A sequence captured by a single camera can result in multi-view capture as the relative camera-object speed approaches infinity.

Nerfies [6] and HyperNeRF [7]. The sequences used for qualitative evaluation in these works are captured by two synchronized smartphones mounted on a handheld rig. Then, a monocular training sequence is simulated by interleaving every other frames from the two cameras. The approach is evaluated on the unused held-out alternative frames from both cameras not used in training. See Figure 2(a) for a 2D illustration. Note that Nerfies [6] focus on "quasi-static" object motion, where the object does not move too much.

NSFF [5]. NSFF builds its quantitative evaluation on the Dynamic Scene Dataset from Yoon *et al.* [41]. The original dataset consists of sequences captured by a rig with 12 static cameras in a 2D grid. For quantitative evaluation NSFF subsamples 24 training frames interleaved over the 12 static cameras over time for its monocular training sequence. The unused frames are used for evaluation. See Figure 2(b) for illustraion.

Due to evaluation considerations using PSNR that we will expand later, these protocols all deploy an interleaving strategy for view coverage. Unfortunately, this setup results in cameras that move instantaneously between frames at a much higher rate than the scene motion, making these setups effectively multi-view, which we quantify next.

3.2 Quantifying effective multi-view

We propose to quantify the amount of effective multi-view in a sequence by the camera and scene motion magnitude, and propose two metrics Ω and ω as effective multi-view factors (EMF). Note that in this discussion we assume an object motion but this analysis applies to any free-form scene

Dataset	Motion	EMF Ω	EMF ω	Train Cam.	Depth	Kpt.	Sequences
Nerfies [6]	Quasi-Static	1.73	$135°/\text{s}$	2			4 MV
HyperNeRF [7]	Quasi-Static	2.77	$208°/\text{s}$	2			4 MV*
NSFF [5]	Dynamic	0.43	$150°/\text{s}$	12	✔		8 MV
iPhone (Ours)	Dynamic	0.08	$13°/\text{s}$	1	✔	✔	6 SV + 4 MV

Table 1: **Benchmark statistics with Effective Multiview Factors (EMF).** We report EMF by relative camera-scene motion Ω and camera angular velocity ω, where high values indicate more effective multi-view. Unlike previous protocols, the proposed iPhone dataset consist of dynamic sequences captured by a *single* camera from daily life with low EMF. It also has accompanying depth maps and labeled keypoints for evaluation. *HyperNeRF provides 13 extra single-view videos not used for novel-view synthesis evaluation.

deformation. Formally, consider a point $\mathbf{x}_t \in \mathbb{S}^2_t$ on the object's surface and a camera parameterized by its origin $\mathbf{o}_t$ at time $t \in \mathcal{T}$. We can compute 3D point displacement $\mathbf{x}_{t+1} - \mathbf{x}_t$ and 3D camera displacement $\mathbf{o}_{t+1} - \mathbf{o}_t$ given their correspondence over time. We define the relative camera-scene motion as the expected ratio,

$$\Omega = \mathop{\mathbb{E}}_{t \in \mathcal{T}} \left[\mathop{\mathbb{E}}_{\mathbf{x}_t \in \mathbb{S}^2_t} \left[\frac{\|\mathbf{o}_{t+1} - \mathbf{o}_t\|}{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|} \right] \right]. \tag{1}$$

We use a ratio over other alternatives as it allows this metric to be scale agnostic. While the numerator is straightforward to compute from the camera trajectory available from SfM systems [42] or an on-device gyroscope, the denominator requires computing the 3D scene flow. We estimate 3D scene flow using dense 2D optical flow and depth maps, which together with the camera parameters can be converted into a 3D scene flow for visible surfaces. For our dataset, we use the depth information available from the iPhone, with its scale aligned with the camera. When metric depth is not available, it is estimated using DPT [43]. 2D optical flow is obtained from RAFT [44] through which occluded regions can be obtained by a forward-backward photometric check. Please see the supplemental for more details for detailed computation. Note that although Ω is theoretically intuitive and incorporates both scene and camera motion, computing the 3D scene flow is non-trivial and thus it can suffer from noisy estimates.

We therefore introduce a second measure ω that empirically works well in practice, which only depends on the camera trajectory. Assuming that the camera is fixated on the object and assuming a constant object motion, the angular velocity of the camera encapsulates how much multi-view signal is in a sequence. Specifically, given a a single fixed "look-at" point $\mathbf{a}$ computed by triangulating the optical-axes of all cameras (as per [6]) and the frame rate N, the camera angular velocity ω is computed as the scaled expectation,

$$\omega = \mathop{\mathbb{E}}_{t \in \mathcal{T}} \left[\arccos \left(\frac{\langle \mathbf{a} - \mathbf{o}_t, \mathbf{a} - \mathbf{o}_{t+1} \rangle}{\|\mathbf{a} - \mathbf{o}_t\| \cdot \|\mathbf{a} - \mathbf{o}_{t+1}\|} \right) \right] \cdot N. \tag{2}$$

In Table 1, we report EMF values Ω and ω on existing datasets. For both metrics, higher values indicate higher the amount of effective multi-view. We find that existing protocols have very large EMF values on both metrics, (as large as $\sim 200°/\text{s}$ for HyperNeRF) indicating that existing protocols are effectively leveraging multi-view clues. To put these numbers in context, a person imaging an object 3m away has to move at 1m/s to get a $\omega = 20°/\text{s}$ (close to the statistics in the proposed dataset). Higher ω like $120°/\text{s}$ implies a speed faster than the Olympic 100m sprint record (without incurring any motion blur).

One reason behind existing protocols leading to high effective multi-view factors is for enabling meaningful photometric quality evaluation via PSNR. As these methods are not designed to synthesize parts of the scene never seen during training, current protocols ensure that both test and train views have been seen during training via interleaving frames from different cameras at the expense of leaking multi-view information in the training sequence. Next we introduce a new dataset and two evaluation metrics, masked PSNR and correspondence accuracy, which enable training with purely monocular sequences.

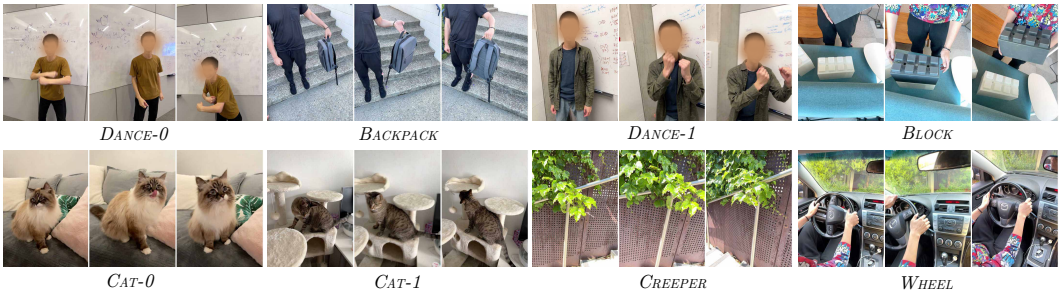

Figure 5: **Sampled sequences in our iPhone dataset.** It consists of diverse moving objects and motion in real-life using a monoculr camera.

4 Proposed monocular capture, dataset, and evaluation protocol

4.1 Capture setup and iPhone dataset

In contrast to existing data, we capture training sequences using a single handheld smartphone in an uncontrolled environment. For evaluating novel views, we additionally set up one or two cameras for some captured sequences as shown in Fig. 4. Note those additional cameras are *not* accessible during training.

Our dataset consists of 6 single-view sequences and 4 multi-view sequences spanning a diverse set of objects and motions, including articulating people and pets, non-rigid deformation and rigid transforms in objects. Fig. 5 shows sampled sequences of our dataset. All captures include per-frame RGB-D data as well as camera poses acquired from iPhone AR toolkit. We find that depth maps are noisy around depth discontinuities due to sensing quality and therefore apply Sobel filter [45] to mask out these regions. For multi-view sequence, cameras are time synchronized using the same technique used in [6]. Each training sequence is annotated with 10 to 20 keypoints in every 10 frames, which enables correspondence evaluation discussed next.

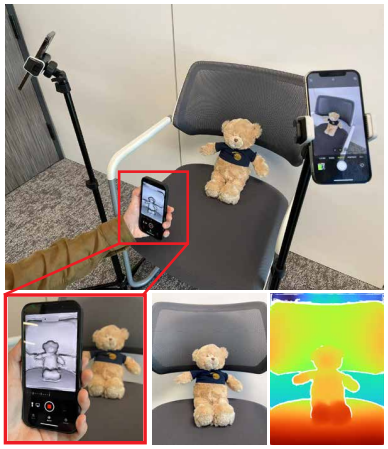

Figure 4: **Our Capture Setup.** We use a handheld iPhone to capture the monocular RGB-D data in daily life. For some data, we also setup two static cameras for the purpose of novel-view evaluation.

4.2 Evaluation protocols

Masked PSNR. Existing works evaluate novel-view synthesis with per-pixel metrics *on the full image* with metrics such as PSNR, SSIM [46] and LPIPS [47], as done in novel-view synthesis evaluation on static scenes. However, in dynamic scenes, particularly for monocular capture setups, there can be parts of the scene which are unconstrained due to motion. To circumvent this issue without resorting to frame interleaving, we leverage depth maps available in our iPhone dataset and use it to identify which parts of the scene should be visible from the training view. We then only evaluate these parts of the scene from a held-out camera, and denote this metric as Masked PSNR_M. The visibility mask is computed by first back-projecting all pixels into 3D space using depth from the training view, followed by projecting them to the target novel view image. Figure 6 shows examples of the visibility mask on our iPhone dataset.

Correspondence. Setting up a multi-view capturing system for novel-view rendering evaluation can be exceedingly cumbersome due to time synchronization and camera registration. Besides, pixel-level metrics like PSNR only evaluate rendering quality and do not directly reflect underlying geometry and its deformation. On the other hand, correspondence has been demonstrated to play key role in non-rigid reconstruction [31]. To this end, we propose to evaluate the 2D correspondence across training frames with PCK-T metric (percentage of correctly transferred key-points) [48], which has at least three benefits. First, it directly evaluates on the deformation modeling of a method. Second, since a good deformation modeling is essential for photorealistic novel-view rendering,

correspondence is also a good indicator on the performance of novel-view rendering as shown in our experiments (see Table. 2). Last, and most importantly, it can be calculated from only the training frames without requiring a multi-view capturing setup. Thus we argue that the correspondence should be a first-class metric for evaluating dynamic view synthesis methods going forward, as long as the key-points annotation are accessible (provided in our iPhone dataset).

5 Approaches, losses, and correspondence computation

A neural radiance field (NeRF) [40] represents a *static* scene as a continuous volumetric field F that transforms a point's position $\mathbf{x}$ and auxiliary variables $\mathbf{w}$ (*e.g.*, view direction, latent appearance vector) to color $\mathbf{c}$ and density σ,

$$F : (\mathbf{x}, \mathbf{w}) \mapsto (\mathbf{c}, \sigma). \tag{3}$$

Here we briefly review representative approaches that extend NeRFs to dynamic scenes.

Nerfies [6]/HyperNeRF [7]. Similarly to traditional non-rigid reconstruction methods that explains non-rigid scenes with a static *canonical* space and a per-frame deformation model [31], Nerfies [6] capture a non-rigid scene with one canonical NeRF F and a per-time step view-to-canonical deformation $W_{t\to c}$ that takes a point $\mathbf{x}$ with a time-conditioned latent vector $\boldsymbol{\varphi}_t$ to a canonical point $\mathbf{x}_c$

$$W_{t\to c} : (\mathbf{x}, \boldsymbol{\varphi}_t) \mapsto \mathbf{x}_c. \tag{4}$$

At each time step the resulting volumetric field is $F_t = F \circ W_{t\to c}$. HyperNeRF [7] addresses topological change on top of Nerfies by outputting a two-dimensional "ambient" coordinate $\mathbf{w}$ encoding the topological change in addition to the canonical point $\mathbf{x}_c$,

$$W_{t\to c} : (\mathbf{x}, \boldsymbol{\varphi}_t) \mapsto (\mathbf{x}_c, \mathbf{w}). \tag{5}$$

These two output variables are passed to the (topologically varying) canonical space mapping F.

Time-conditioned NeRF. Another way to handle non-rigid scenes is to directly map space-time to the output color and density by a time-conditioned latent vector $\boldsymbol{\varphi}_t$, which we refer to as T-NeRF:

$$F_t : (\mathbf{x}, \boldsymbol{\varphi}_t) \mapsto (\mathbf{c}, \sigma). \tag{6}$$

Note that since T-NeRF implicitly handles deformation, it is difficult to compute correspondences over time. Prior work has shown that T-NeRF's implicit deformation handles motion poorly [8, 5, 6, 7]. Therefore, NSFF [5] augments T-NeRF's implicit function F_t to output an explicit scene flow field $W_{t\to t+\delta}$ between adjacent time steps t and $t+\delta$,

$$W_{t\to t+\delta} : \mathbf{x} \mapsto \mathbf{x}', \;\; \delta \in \{+1, -1\}. \tag{7}$$

This explicit flow field is used to regularize motion and, as shown below, can be chained to compute long-range point correspondences across views and times.

All the above methods are originally trained with pure RGB images and camera poses, except for NSFF which additionaly leverages predicted monocular depth [49, 43] and optical flow [44]. To augment their performance, we further equip them with depth supervision, as well as volume sparsity prior [50] which encourages solid surface in the scene.

Depth and sparsity losses We study two regularizations based on depth and sparsity in our experiments, which we find to be necessary for these methods to get a reasonable performance in monocular setups. Recent works [5, 9, 51] seek to additionally supervise NeRF models with depth. Similar to pixel reconstruction loss, expected depth along each ray can be volumetrically rendered and compared against ground-truth. Sparsity regularization is often used to resolve "fuzzy" surface and floater artifacts [52, 53, 54, 55, 56]. Among all the proposed variants, mip-NeRF 360 [50] advocates an interval-based regularization, which we use as a representative implementation.

5.1 Computing correspondences

We are interested in reading out 2D view-to-view correspondence from existing models, which requires different composing methods for different models. To obtain 2D correspondence, we first need a 3D transformation map $W_{t_1\to t_2}$ between two frames. For Nerfies and HyperNeRF, it is chained by composing the learned inverse map $W_{t_1\to c}$ and a forward map $\tilde{W}_{c\to t_2}$ obtained by

	+D	+S	TRAIN PSNR ↑ (MV + SV)	TRAIN PSNR ↑ (MV)		TEST PSNR_M ↑ (MV)	TEST PCK-T (%) ↑ (MV + SV)
Depth Re-proj.			–	–		13.75	–
T-NeRF			28.27	25.36	$\xrightarrow{-11.3}$	14.05	–
T-NeRF	✔		28.10	25.46	$\xrightarrow{-9.6}$	15.82	–
T-NeRF	✔	✔	27.93	25.46	$\xrightarrow{-9.0}$	16.48	–
NSFF	✔		27.81	25.70	$\xrightarrow{-11.7}$	13.96	52.4
HyperNeRF			30.95	28.82	$\xrightarrow{-14.4}$	14.43	59.9
HyperNeRF	✔		30.77	28.52	$\xrightarrow{-12.0}$	16.48	63.3
HyperNeRF	✔	✔	29.75	28.05	$\xrightarrow{-11.4}$	16.70	64.0
Nerfies			30.03	27.72	$\xrightarrow{-13.8}$	13.89	58.2
Nerfies	✔		29.76	27.34	$\xrightarrow{-11.3}$	16.04	66.0
Nerfies	✔	✔	28.22	25.98	$\xrightarrow{-9.0}$	16.93	69.9

Table 2: **Evaluation on our captured iPhone dataset.** Best second-best third-best . All approaches have gap between train and test on monocular sequences. PCK correlates well with held-out PSNR. All approaches benefit from depth and sparsity supervision.

root finding [57]: $W_{t_1 \to t_2} = \tilde{W}_{c \to t_2} \circ W_{t_1 \to c}$. For NSFF, it is chained by composing all local warps spatial-temporally: $W_{t_1 \to t_2} = W_{t_2-1 \to t_2} \circ \cdots \circ W_{t_1+1 \to t_1+2} \circ W_{t_1 \to t_1+1}$. Please see the supplemental material for more details.

6 Experiments

We study the approaches outlined in Section 5 as well as their variations with depth supervision ("+D") and sparsity regularization ("+S") under a strict monocular setting using our proposed dataset and evaluation protocol. Results are reported in terms of both photometric and correspondence accuracy.

Metrics and baselines. We report metrics for frames seen in training as well as views held-out for evaluation. For training metrics, we report the average PSNR over all sequences ("MV + SV") and multi-view sequences only ("MV"). For testing evaluation metrics, we report both photometric quality on held-out novel views using average PSNR_M over all multi-view sequences ("MV") and average PCK-T under 0.05 threshold over all seen views during training ("MV + SV"). We provide a simple baseline by re-projecting depth from training view to testing novel view, noted as "Depth Re-proj.".

Implementation details. We re-implement all methods in JAX [58] except for NSFF, which we follow the public re-implementation [59][2]. The same set of training hyper-parameters are used for all the re-implemented methods. Concretely, following [60], we use a batch size $B = 3072$ for a max iteration $K = 3 \times 10^5$, using ADAM optimizer [61] with an initial learning rate $\eta = 1 \times 10^{-3}$. For experiments with extra supervision or regularization, we use weights $w_{\text{depth}} = 1$ and $w_{\text{sparsity}} = 1 \times 10^{-1}$ for depth and sparsity loss, respectively. We disable view direction and appearance encoding for all methods as we find that models tend to overfit with this extra degree of freedom.

Results. We report our main benchmark results across different methods and sequences in Table 2 and Figure 6. Please see the supplement for a detailed report per sequence and model. PSNR_M of depth re-projection serves as a lower bound for evaluating a model's performance: a model that denoises over input depth should at least be better than this simple baseline. We find that without the depth and sparsity losses, models are only marginally better than this lower bound. These additional losses are also necessary to reduce blur artifacts as shown in Figure 7. We find that improvements in correspondence correlates with better novel rendering view rendering, validating that evaluating correspondence transfer on the seen views during training provides a meaningful signal for novel-view rendering quality. Visualizations for key-point transfer can be found in supplement.

In all, we find that the PSNR for novel view synthesis is significantly worse compared to the fitting accuracy on training frames. This is also reflected in the qualitative results, where even the best-performing methods don't generate photo-realistic images from novel viewpoints *e.g.* boundaries are not sharp, surfaces are incorrectly stretched as seen in Figure 6. Our results thus show that dynamic 3D capture in everyday monocular setups is far from solved and there is much room for improvement.

[2] We find that this code base performs better than the original code release

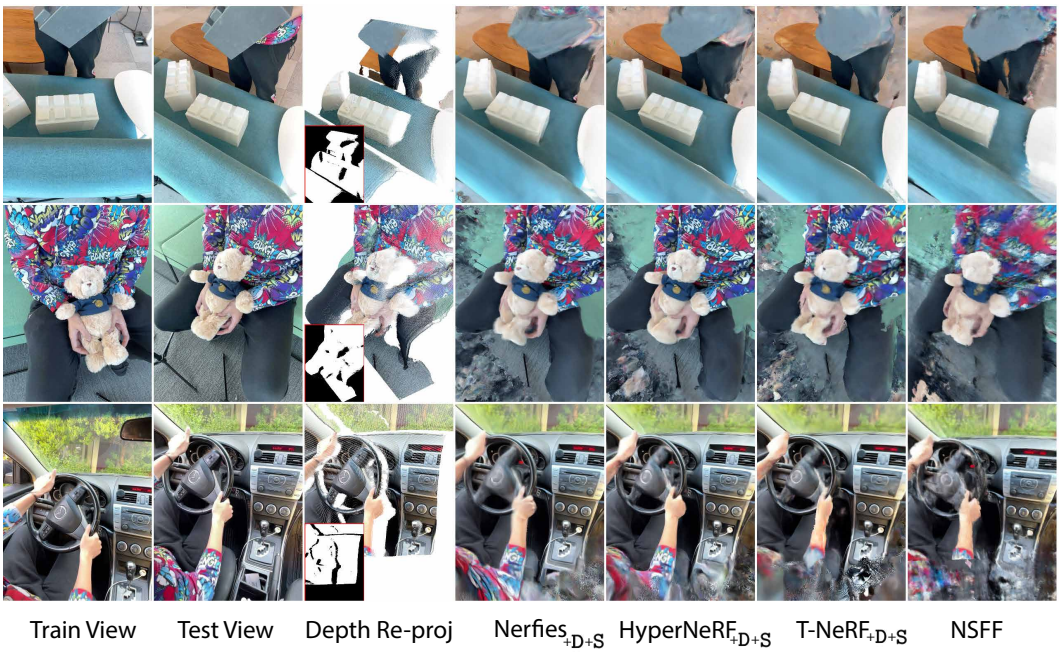

Figure 6: **Novel-view synthesis results on our iPhone dataset.** We show the four approaches we evaluate with a baseline method method which simply re-projects the train view to the test view using depth, denoted as "Depth Re-proj". We visualize the visibility mask used to compute the PSNR_M (shown as an inset) together with the depth re-projection.

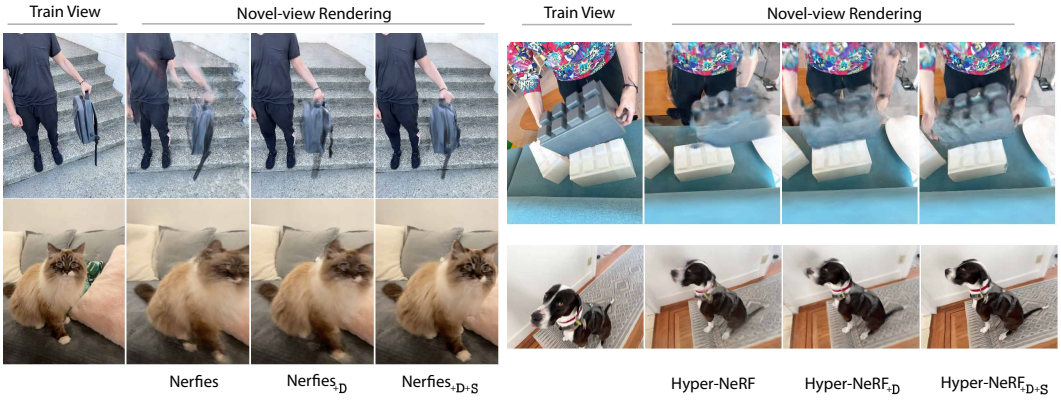

Figure 7: **Augmenting existing works with depth (+D) and sparsity (+S).** We show results on four single-view sequences from our dataset. For each sequence, the novel view point is selected by randomly selecting a view from one of the training camera poses seen at a different time. Please see more in the supplemental.

7 Discussion

In this work we analyze existing monocular dynamic 3D novel view synthesis methods by introducing the notion of effective multi-view and introduce Effective Multiview Factors (EMF) to quantify this based on camera and the scene motion magnitudes. Through this lens we identify that existing evaluation protocols operate in a regime that are far from monocular sequence capture that can be acquired from a single camera. We introduce a new dataset of monocular sequences with depth information and propose two evaluation methods for monocular sequences: depth-guided PSNR and correspondence accuracy measured by PCK-T. Our analysis show that existing approaches suffer without depth guidance and sparsity regularization and have a room to improve under these more challenging scenarios. We encourage future works to report the EMF factors Ω and/or ω in their training sequences to characterize the difficulties of the input sequence.

References

[1] Takeo Kanade, Peter Rander, and PJ Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE multimedia*, 4(1):34–47, 1997.

[2] Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. Relighting human locomotion with flowed reflectance fields. In *ACM SIGGRAPH 2006 Sketches*, pages 76–es. 2006.

[3] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015.

[4] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *CVPR*, 2021.

[5] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021.

[6] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021.

[7] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), Dec 2021.

[8] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2021.

[9] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021.

[10] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 23(3):600–608, 2004.

[11] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020.

[12] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] Lorenzo Torresani, Aaron Hertzmann, and Chris Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):878–892, 2008.

[14] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4447–4455, 2015.

[15] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[16] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European conference on computer vision*, pages 573–586. Springer, 2012.

[17] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[18] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1966–1974, 2015.

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

[20] Nilesh Kulkarni, Abhinav Gupta, David Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021.

[22] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. Banmo: Building animatable 3d neural models from many casual videos. *arXiv preprint arXiv:2112.12761*, 2021.

[23] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[24] Benedict J. Brown and Szymon Rusinkiewicz. Non-rigid range-scan alignment using thin-plate splines. In *Symposium on 3D Data Processing, Visualization, and Transmission*, 2004.

[25] Benedict J. Brown and Szymon Rusinkiewicz. Global non-rigid alignment of 3-d scans. *ACM Trans. Graph.*, 26(3), 2007.

[26] Will Chang and Matthias Zwicker. Range scan registration using reduced deformable models. *Comput. Graph. Forum*, 28(2):447–456, 2009.

[27] Hao Li, Robert W. Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. *Computer Graphics Forum (Proc. SGP'08)*, 27(5), 2008.

[28] Aljaz Bozic, Pablo Palafox, Michael Zollöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural non-rigid tracking. In *NeurIPS*, 2020.

[29] Wei Gao and Russ Tedrake. Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. In *Robotics: Science and System (RSS)*, 2018.

[30] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobald, and Marc Stamminger. VolumeDeform: Real-Time Volumetric Non-rigid Reconstruction. In *ECCV*, 2016.

[31] Richard Newcombe, Dieter Fox, and Steve Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015.

[32] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[33] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d reconstruction of a moving point from a series of 2d projections. In *ECCV*, 2010.

[34] Kaan Yucer, Oliver Wang, Alexander Sorkine-Hornung, and Olga Sorkine-Hornung. Reconstruction of articulated objects from a moving camera. In *ICCV Workshop*, 2015.

[35] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021.

[37] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021.

[38] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021.

[39] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021.

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[41] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020.

[42] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[45] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[47] 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.

[48] Nilesh Kulkarni, Abhinav Gupta, and Shubham Tulsiani. Canonical surface mapping via geometric cycle consistency. In *ICCV*, 2019.

[49] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. 2020.

[50] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

[51] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[52] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

[53] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020.

[54] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *ICCV*, 2021.

[55] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.

[56] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021.

[57] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021.

[58] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[59] https://github.com/kwea123/nsff_pl, 2021.

[60] Boyang Deng, Jonathan T. Barron, and Pratul P. Srinivasan. JaxNeRF: an efficient JAX implementation of NeRF, 2020.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] in Section.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] in Section 6
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] in Section 6
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] in Section 6
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]