# R5DGS: Semantic-Aware 4D Gaussian Splatting with Rigid Body Constraints for Efficient Dynamic Scene Reconstruction

Denis Gridusov[1], Maxim Popov[1] and Sergey Kolyubin[1]

***Abstract*— Reconstructing and predicting dynamic 3D scenes from multi-view videos is a foundational task for robotics, AR/VR, and digital twins. Recent physics-informed Gaussian Splatting methods achieve impressive future frame extrapolation but lack semantic awareness and suffer from large computational overhead. We introduce R5DGS, a framework that augments a physics-driven 4D Gaussian representation with compact Identity Encoding vectors, enabling precise Gaussian-to-object association. By constructing an offline CLIP-based object lookup table, we support open-vocabulary text prompting to retrieve and render object-specific Gaussians across arbitrary timestamps and viewpoints. Furthermore, we propose a rigid-body inference constraint that predicts and integrates physical dynamics exclusively for object centroids, propagating motion to associated Gaussians via relative transformations. This optimization yields a 11 FPS speedup during extrapolation without compromising trajectories plausibility.**

## I. Introduction

Reconstructing and predicting dynamic 3D scenes from multi-view video is a foundational prerequisite for a wide range of embodied applications. In robotics and autonomous navigation, accurate spatiotemporal modeling enables safe trajectory planning and proactive interaction with moving agents. In augmented/virtual reality and digital twins, it facilitates immersive scene editing and realistic predictive simulation. Beyond static view synthesis, the ability to extrapolate future states from limited observations is essential for closing the perception-action loop in unstructured environments.

Traditional 3D reconstruction has evolved from explicit discretized representations (voxels, point clouds, meshes) to implicit continuous fields such as signed distance functions [1], [2] and Neural Radiance Fields (NeRF) [3]. While NeRFs achieve high-fidelity novel view synthesis, their volumetric rendering remains computationally intensive. Recently, 3D Gaussian Splatting (3DGS) [4] has emerged as a dominant explicit alternative, leveraging particle-based Gaussian kernels for real-time, high-quality rendering. To capture temporal dynamics, early approaches extended NeRF by learning time-dependent deformation fields [5], [6], [7], [8] or scene flows [9]. These were rapidly adapted to the 3DGS paradigm, with methods like 4D Gaussian Splatting [10] and Deformable 3DGS [11] achieving state-of-the-art interpolation within observed video durations. However, these deformation-based approaches primarily optimize pixel-level correlations without encoding physical laws, leading to motion drift and complete failure when extrapolating beyond the training horizon.

[1] Biomechatronics and Energy-Efficient Robotics (BE2R) Lab, ITMO University, Saint Petersburg, Russia

To enable predictive modeling, recent works have integrated physical priors into 3D reconstruction. Physics-Informed Neural Networks (PINNs) [12] embed partial differential equations as soft regularization terms to model phenomena like fluid dynamics [13], [14] or continuum mechanics [15]. Despite theoretical guarantees, PINNs suffer from high training costs, require explicit boundary conditions or object masks, and often struggle with complex, multi-part dynamics. Alternatively, explicit physics simulators leverage spring-mass systems [16], [17] or graph networks [18] to model specific material behaviors. While effective for targeted domains, these methods lack generality and depend heavily on predefined object categories or material properties. More recently, TRACE [19] proposes a generalized translation-rotation dynamics system per Gaussian particle, enabling label-free future extrapolation through classical mechanics and Runge-Kutta integration. Nevertheless, this per-particle formulation leads to significant computational overhead during inference and inherently lacks semantic coherence, as independently evolving Gaussians cannot be easily grouped into meaningful objects for downstream perception or control tasks.

In this work, we bridge the gap between physical dynamics, semantic controllability, and inference efficiency. We introduce **R5DGS**, a unified framework that augments a physics-driven 4D Gaussian representation with instance-level semantic grouping and a rigid-body optimization scheme for extrapolation. Our primary contributions are as follows:

Our primary contributions are follows:

- A semantics-augmented physics-informed 4D Gaussian framework that enables precise Gaussian-to-object grouping via compact Identity Encodings.
- A centroid-driven rigid-body inference strategy that accelerates future prediction by 11 FPS while preserving trajectories plausibility, bridging the gap between high-fidelity physical extrapolation and real-time interactive deployment.
- An open-vocabulary querying mechanism powered by an offline CLIP-based lookup table, allowing text-driven retrieval and rendering of dynamic objects across time and viewpoints.

## II. Methodology

We build upon TRACE [19], a physics-informed 4D Gaussian framework for dynamic scene extrapolation. Given multi-view RGB videos, TRACE represents the scene at a canonical timestamp $t = 0$ as a set of 3D Gaussians

$\mathcal{G}^0 = \{(\mathbf{x}_i, \mathbf{r}_i, \mathbf{s}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^3$ is the center, $\mathbf{r}_i \in \mathbb{R}^4$ the rotation quaternion, $\mathbf{s}_i \in \mathbb{R}^3$ the scaling, $\alpha_i$ the opacity, and $\mathbf{c}_i$ the SH-encoded color.

Temporal evolution is modeled via two parallel modules: (1) an auxiliary deformation field $f_{\text{def}}$ that predicts per-Gaussian displacements $(\Delta\mathbf{x}, \Delta\mathbf{r}, \Delta\mathbf{s})$ for interpolation within observed time, and (2) a Translation-Rotation Dynamics (TRD) module that learns physical parameters: equivalent center velocity $\bar{\mathbf{v}}_c$, acceleration $\bar{\mathbf{a}}_c$, angular velocity $\boldsymbol{\omega}_p$, and angular acceleration $\boldsymbol{\varepsilon}_p$ to enable future frame extrapolation via 2nd-order Runge-Kutta integration. While TRACE achieves strong extrapolation quality, it treats each Gaussian independently, resulting in high inference cost and lacking semantic object awareness.

### A. Identity-Augmented Representation

To enable object-level reasoning, we augment each 3D Gaussian with a compact, learnable 16-dimensional identity vector $\mathbf{e}_i \in \mathbb{R}^{16}$, inspired by Gaussian Grouping [20]. During differentiable rendering, these features are $\alpha$-blended to the image plane:

$$\mathbf{E}_{\text{id}} = \sum_{i \in \mathcal{N}} \mathbf{e}_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \tag{1}$$

where $\alpha'_i$ denotes the projected opacity. The rendered identity map is supervised by:

- **2D Identity Loss.** Classifier $f$ maps $\mathbf{E}_{\text{id}}$ to $\mathbf{K}$ object classes, optimized via cross-entropy against coherent multi-view masks.
- **3D Spatial Regularization.** A KL-divergence penalty enforces feature consistency among $k$-nearest spatial neighbors in canonical space, mitigating supervision gaps from occlusions.

These semantic priors are jointly optimized with standard photometric reconstruction, yielding the total objective:

$$\mathcal{L}_{\text{5DGS}} = \mathcal{L}_{\text{render}} + \lambda_{\text{obj}} \mathcal{L}_{\text{obj}} + \lambda_{\text{3d}} \mathcal{L}_{\text{3d}}, \tag{2}$$

where $\mathcal{L}_{\text{render}} = (1 - \lambda_{\text{dssim}})\mathcal{L}_1 + \lambda_{\text{dssim}}(1 - \text{SSIM})$ following the 3DGS formulation [4], $\mathcal{L}_{\text{obj}}$ is the 2D Identity Loss, $\mathcal{L}_{\text{3d}}$ is the 3D Spatial Regularization and $\lambda_{\text{obj}}, \lambda_{\text{3d}}$ are scalars.

This gives a discrete, by-design grouping of Gaussians into object instances while preserving reconstruction quality.

### B. Rigid-Body Constrained Extrapolation

The TRD module in TRACE predicts dynamics and integrates position for all scene Gaussians, which is computationally expensive during extrapolation ($t > t_{\max}$). From classical mechanics, we know that the motion of particles within a rigid body is strictly governed by constraints that maintain constant relative distances. Guided by this principle, we introduce a group-level rigid-body constraint during inference, substituting per-Gaussian dynamics with physically consistent object-level propagation.

Let $\mathcal{G}_k$ denote Gaussians assigned to object $k$. We select a representative Gaussian $r_k \in \mathcal{G}_k$ as the one closest to the geometric centroid and precompute canonical offsets:

$$\mathbf{o}_i = \mathbf{x}_i - \mathbf{x}_{r_k}, \quad \forall i \in \mathcal{G}_k. \tag{3}$$

During extrapolation, we first propagate all Gaussians to $t_{\max}$ using $f_{\text{def}}$. We then integrate TRD dynamics only for the $K$ representatives, obtaining future position $\mathbf{x}_{r_k}^{\text{vel}}$ and orientation quaternion $\mathbf{q}_{r_k}^{\text{vel}}$. Motion is rigidly propagated to remaining Gaussians via:

$$\Delta\mathbf{x}_i = \big(\mathbf{x}_{r_k}^{\text{vel}} - \mathbf{x}_{r_k}^{\text{def}}\big) + \big(\mathbf{R}(\mathbf{q}_{r_k}^{\text{vel}})\mathbf{o}_i - \mathbf{o}_i\big), \tag{4}$$

$$\mathbf{q}_i^{\text{out}} = \mathbf{q}_{r_k}^{\text{vel}} \otimes \Delta\mathbf{q}_i^{\text{def}}, \tag{5}$$

where $\mathbf{R}(\cdot)$ converts quaternions to rotation matrices, $\otimes$ denotes Hamilton product, and superscript $def$ indicates deformation network outputs. This formulation preserves inter-point distances and relative orientations by construction while reducing MLP and integrator queries from $\mathcal{O}(N)$ to $\mathcal{O}(K)$, where $K \ll N$.

### C. Additional Loss Components

To increase reconstruction quality we introduce several loss components.

*a) Rigid Distance Preservation:* To encourage rigid motion during training, we introduce:

$$\mathcal{L}_{\text{rigid}} = \frac{1}{N} \sum_{i=1}^{N} \Big( \big\| \mathbf{x}_i^{\text{after}} - \mathbf{x}_{r_{g(i)}}^{\text{after}} \big\| - \big\| \mathbf{x}_i^{\text{before}} - \mathbf{x}_{r_{g(i)}}^{\text{before}} \big\| \Big)^2, \tag{6}$$

where $\mathbf{x}_i^{\text{before}}$ denotes the Gaussians center position before applying the predicted by $f_{\text{def}}$ and TRD displacement and $\mathbf{x}_i^{\text{after}}$ are those positions after deformation. The loss penalizes changes in distance from each Gaussian to its object representative. This soft constraint complements the hard rigid propagation at inference.

*b) Semantic Majority Consistency:* For maintaining semantic coherence, we found that soft KL-divergence regularization is not enough, so add extra penalty to minimize MSE between a query Gaussian's predicted class distribution ($p_q$) and the mean of its $k$ spatial neighbors ($p_n$):

$$\mathcal{L}_{\text{major}} = \frac{1}{M} \sum_{q=1}^{M} \Big\| \mathbf{p}_q - \frac{1}{k} \sum_{n \in \mathcal{N}_k(q)} \mathbf{p}_n \Big\|^2. \tag{7}$$

Losses $\mathcal{L}_{\text{3d}}$ and $\mathcal{L}_{\text{major}}$ are computed every $\tau_{\text{reg}}$ iterations, $\mathcal{L}_{\text{rigid}}$ activates after iteration $t_{\text{rigid}}$ once the deformation field stabilizes.

### D. CLIP-like Open-Vocabulary Querying

To enable text-driven scene interaction, we construct an offline lookup table. For each object group $g \in \{1, \ldots, K\}$, we extract a representative masked view, encode it with CLIP-like [21] model, and store the text-aligned embedding $\mathbf{t}_g$. At inference, a natural language prompt $\mathbf{q}_{\text{text}}$ is encoded via the same text encoder, and object groups are retrieved by maximizing cosine similarity:

$$g^* = \arg\max_g \cos\big(\text{CLIP}_{\text{text}}(\mathbf{q}_{\text{text}}), \mathbf{t}_g\big). \tag{8}$$

The retrieved Gaussian subset $\mathcal{G}_{g^*}$ can be independently rendered at arbitrary timestamps $t$ and camera poses $\mathbf{C}$, supporting object isolation, editing, and selective visualization without retraining.

## III. Experiments

### A. Implementation Details

For data preparation, we employ SAM2 [21] combined with DEVA [22] tracking to generate consistent object masks across all views and timestamps. This ensures temporal and multi-view consistency required for training identity encoding vectors. We benchmark our method on the Dynamic Indoor Scene dataset [23], which contains four scenes with multiple objects undergoing independent rigid body motions.

Our training configuration includes the following hyperparameters: for 3D semantic regularization, we set $k = 5$ nearest neighbors with $\lambda_{\text{3d}} = 2.0$, computed every 2 iterations on samples of 1000 points (from maximum 300,000). For majority consistency loss, we use $k = 5$ neighbors with $\lambda_{\text{maj}} = 0.5$, computed on 1000 sampled points. The rigid distance loss is activated after iteration 10,000 with $\lambda_{\text{rigid}} = 0.5$.

### B. Notation and Method Variants.

To systematically evaluate the contribution of each component, we introduce the following naming convention. Each variant is defined by two independent choices: (1) the set of loss functions used during *training*, and (2) the extrapolation strategy applied during *inference*.

| Variant | Training losses | Inference |
|---|---|---|
| **5DGS** | $\mathcal{L}_{\text{5DGS}}$ | Standard |
| **R5DGS** | $\mathcal{L}_{\text{5DGS}}$ | Rigid-body |
| **R5DGS w/ extra loss** | $\mathcal{L}_{\text{5DGS}} + \mathcal{L}_{\text{rigid}} + \mathcal{L}_{\text{major}}$ | Rigid-body |

TABLE I: Summary of method variants. All variants share the same Identity Encoding architecture and CLIP-based lookup table.

Based on the loss function we use, we define three method variants evaluated in our experiments, as shown in Table I.

Inference strategies:

- *Standard*: per-Gaussian dynamics prediction via the full TRD module (as in TRACE [19]).
- *Rigid-body*: group-level dynamics prediction for representative Gaussians only, with rigid propagation to the rest of the object (Sec. II-B).

### C. Benchmark Results

We evaluate our method across theree key dimensions: rendering quality (PSNR, SSIM, LPIPS), segmentation accuracy (mIoU) and novel view synthesis speed (FPS) on unseen timestamps (extrapolation).

*a) Results Analysis:* Table II and Table III demonstrate the performance characteristics of our approach. By restricting physics prediction to a small set of representative Gaussians per object (typically 9-12) rather than the full set ($\sim$40,000), our approach achieves a consistent 11 FPS speedup during extrapolation novel view synthesis. While this group-level constraint introduces a reduction in photometric fidelity, mainly stemming from a small number of misclassified Gaussians near object boundaries and the exclusion of shadows from trajectory propagation, as they are not classified as part of the object, it preserves physically plausible motion trajectories and maintains structural coherence across future frames, as visualized in Figure 1. Furthermore, segmentation accuracy remains nearly unaffected, confirming that the rigid propagation effectively preserves object-level structure.

*b) Segmentation Failures Analysis:* We observe notably lower mIoU on Darkroom and Factory scenes. This is primarily attributed to tracking errors during the offline mask preparation stage with SAM2+DEVA, where occlusions and rapid motions caused identity switches. Figure 1 visualizes representative failure cases where the tracker loses object consistency across frames.

### D. Open-Vocabulary Grounding

Beyond reconstruction and segmentation, our framework enables semantic interaction with the 4D scene. By constructing an offline CLIP-based lookup table that maps object Gaussian groups to text embeddings, we support natural language queries for object retrieval and selective rendering. Specifically, we used Perception Encoder [24] for extracting image and text embeddings. Figure 2 demonstrates successful grounding of queries like "donut" and "apple", where the system retrieves the corresponding Gaussian subset and renders it at arbitrary timestamps and viewpoints. This capability opens new possibilities for robot instruction following and interactive scene editing without retraining.

## IV. Conclusion

In this work, we introduced a semantics-augmented, physics-informed 4D Gaussian framework that bridges the gap between high-fidelity dynamic scene extrapolation and real-time interactive deployment. By integrating compact Identity Encodings with a Translation-Rotation Dynamics system, we enable precise Gaussian-to-object grouping and open-vocabulary text-driven querying via an offline CLIP lookup table. Crucially, our rigid-body inference constraint replaces per-Gaussian physics prediction with centroid-driven integration and rigid propagation, yielding an 11 FPS speedup during extrapolation while preserving physically plausible motion trajectories and structural coherence.

Despite these advances, the rigid-body assumption introduces a moderate trade-off in photometric fidelity, primarily affecting highly deformable regions or scenes with imperfect mask tracking. The current pipeline also relies on offline SAM2+DEVA mask association, which can suffer from identity switches under severe occlusions or rapid motion. Future work will focus on enhancing the semantic supervision and extending our rigid-body formulation to articulated and composite objects.

Finally, our approach demonstrates that semantic-aware, physics-constrained Gaussian representations offer a practical path toward efficient, controllable 4D scene understanding for embodied applications.

| Methods | Dining Table | | | Chessboard | | | Darkroom | | | Factory | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| TRACE | 35.580 | 0.962 | 0.0497 | 34.630 | 0.963 | 0.055 | 37.774 | 0.961 | 0.067 | 36.488 | 0.965 | 0.049 |
| 5DGS (Ours) | 35.428 | 0.956 | 0.055 | 33.991 | 0.956 | 0.063 | 36.600 | 0.955 | 0.074 | 35.926 | 0.958 | 0.055 |
| R5DGS (Ours) | 28.844 | 0.942 | 0.066 | 28.805 | 0.932 | 0.086 | 31.181 | 0.939 | 0.091 | 29.798 | 0.924 | 0.075 |
| R5DGS w/ extra loss (Ours) | 28.688 | 0.939 | 0.067 | 29.153 | 0.929 | 0.089 | 31.537 | 0.943 | 0.087 | 30.749 | 0.928 | 0.073 |

TABLE II: Reconstruction metric results for different model variants compared with TRACE [19].

| Methods | Dining Table | | Chessboard | | Darkroom | | Factory | | Overall | |
|---|---|---|---|---|---|---|---|---|---|---|
| | FPS↑ | mIoU↑ | FPS↑ | mIoU↑ | FPS↑ | mIoU↑ | FPS↑ | mIoU↑ | FPS↑ | mIoU↑ |
| 5DGS | 66.9 | **0.78** | 67.3 | **0.75** | 49.4 | 0.37 | 64.9 | **0.47** | 62.1 | 0.59 |
| R5DGS | **76.3** | 0.77 | **76.9** | 0.73 | **66.2** | **0.38** | **75.0** | 0.46 | **73.6** | 0.59 |

TABLE III: mIoU and FPS results for different model variants to validate rigid constraints effect. The best results are denoted with **bold**. FPS is calculated on NVIDIA RTX 4090

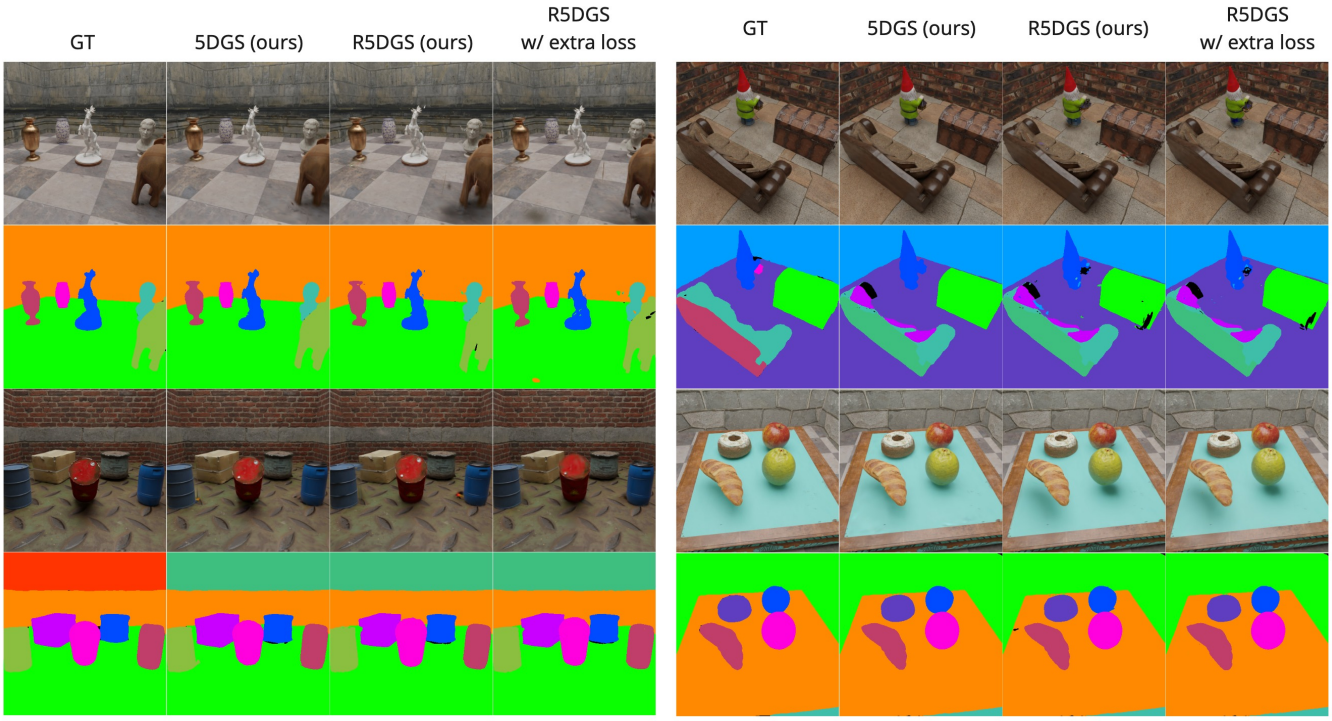


Fig. 1: Visual comparison of different model variants with ground truth rgb and semantics.

Fig. 2: Grounding result visualization with prompts “donut”, “lemon” and “apple”.

## References


[1] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, *Deepsdf: Learning continuous signed distance functions for shape representation*, 2019. arXiv: 1901.05103 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1901.05103.

[2] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, *Occupancy networks: Learning 3d reconstruction in function space*, 2019. arXiv: 1812.03828 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1812.03828.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, *Nerf: Representing scenes as neural radiance fields for view synthesis*, 2020. arXiv: 2003.08934 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2003.08934.

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, *3d gaussian splatting for real-time radiance field rendering*, 2023. arXiv: 2308.04079 [cs.GR]. [Online]. Available: https://arxiv.org/abs/2308.04079.

[5] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, *D-nerf: Neural radiance fields for dynamic scenes*, 2020. arXiv: 2011.13961 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2011.13961.

[6] K. Park et al., *Nerfies: Deformable neural radiance fields*, 2021. arXiv: 2011.12948 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2011.12948.

[7] J. Fang et al., "Fast dynamic radiance fields with time-aware neural voxels," in *SIGGRAPH Asia 2022 Conference Papers*, ser. SA '22, ACM, Nov. 2022, pp. 1–9. DOI: 10.1145/3550469.3555383. [Online]. Available: http://dx.doi.org/10.1145/3550469.3555383.

[8] A. Cao and J. Johnson, *Hexplane: A fast representation for dynamic scenes*, 2023. arXiv: 2301.09632 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2301.09632.

[9] Z. Li, S. Niklaus, N. Snavely, and O. Wang, *Neural scene flow fields for space-time view synthesis of dynamic scenes*, 2021. arXiv: 2011.13084 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2011.13084.

[10] G. Wu et al., *4d gaussian splatting for real-time dynamic scene rendering*, 2024. arXiv: 2310.08528 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2310.08528.

[11] Z. Yang, X. Gao, W. Zhou, S. Jiao, Y. Zhang, and X. Jin, "Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction," *arXiv preprint arXiv:2309.13101*, 2023.

[12] "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.

[13] D. Baieri, S. Esposito, F. Maggioli, and E. Rodolà, *Fluid dynamics network: Topology-agnostic 4d reconstruction via fluid dynamics priors*, 2023. arXiv: 2303.09871 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2303.09871.

[14] J. Qiu, R. Cen, Z. Li, H. Yan, M.-M. Cheng, and B. Ren, "Neusmoke: Efficient smoke reconstruction and view synthesis with neural transportation fields," in *SIGGRAPH Asia Conference Proceedings*, 2024.

[15] X. Li et al., *Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification*, 2023. arXiv: 2303.05512 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2303.05512.

[16] L. Zhong, H.-X. Yu, J. Wu, and Y. Li, *Reconstruction and simulation of elastic objects with spring-mass 3d gaussians*, 2024. arXiv: 2403.09434 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2403.09434.

[17] T. Zhang et al., *Physdreamer: Physics-based interaction with 3d objects via video generation*, 2024. arXiv: 2404.13026 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2404.13026.

[18] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia, *Learning to simulate complex physics with graph networks*, 2020. arXiv: 2002.09405 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2002.09405.

[19] J. Li, Z. Song, and B. Yang, "Trace: Learning 3d gaussian physical dynamics from multi-view videos," *ICCV*, 2025.

[20] M. Ye, M. Danelljan, F. Yu, and L. Ke, "Gaussian grouping: Segment and edit anything in 3d scenes," in *ECCV*, 2024.

[21] A. Radford et al., *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.00020.

[22] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *ICCV*, 2023.

[23] J. Li, Z. Song, and B. Yang, "Nvfi: Neural velocity fields for 3d physics learning from dynamic videos," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 723–34 751, 2023.

[24] D. Bolya et al., *Perception encoder: The best visual embeddings are not at the output of the network*, 2025. arXiv: 2504.13181 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2504.13181.