# Novelty Assessment Report

**Paper**: FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction
**PDF URL**: https://openreview.net/pdf?id=3q9vHEqsNx
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2025-12-30

## Abstract

High-quality 3D world models are pivotal for embodied intelligence and Artificial General Intelligence (AGI), underpinning applications such as AR/VR content creation and robotic navigation. Despite the established strong imaginative priors, current video foundation models lack explicit 3D grounding capabilities, thus being limited in both spatial consistency and their utility for downstream 3D reasoning tasks. In this work, we present FantasyWorld, a geometry-enhanced framework that augments frozen video foundation models with a trainable geometric branch, enabling joint modeling of video latents and an implicit 3D field in a single forward pass. Our approach introduces cross-branch supervision, where geometry cues guide video generation and video priors regularize 3D prediction, thus yielding consistent and generalizable 3D-aware video representations. Notably, the resulting latents from the geometric branch can potentially serve as versatile representations for downstream 3D tasks such as novel view synthesis and navigation, without requiring per-scene optimization or fine-tuning. Extensive experiments show that FantasyWorld effectively bridges video imagination and 3D perception, outperforming recent geometry-consistent baselines in multi-view coherence and style consistency. Ablation studies further confirm that these gains stem from the unified backbone and cross-branch information exchange.



## Core Task Landscape

This paper addresses: **Geometry-Consistent World Modeling via Unified Video and 3D Prediction**

A total of **29 papers** were analyzed and organized into a taxonomy with **18 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Unified Video-3D Representation Learning**
- **Geometry-Conditioned Video Generation**
- **Geometry Estimation from Video**
- **World Models for Embodied Intelligence**
- **Controllable 3D Scene Generation**
- **Temporal Prediction with Spatial Reasoning**

### Complete Taxonomy Tree

- Geometry-Consistent World Modeling via Unified Video and 3D Prediction Survey Taxonomy
- Unified Video-3D Representation Learning
  - Cross-Modal Supervision for Joint Video-Geometry Learning ★ (3 papers)
  - [0] FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction (Anon et al., 2026) View paper
  - [2] Aether: Geometric-aware unified world modeling (Zhu, 2025) View paper
  - [14] Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling (Wu Haoyu, 2025) View paper
  - 4D Dynamic Scene Representation (3 papers)
  - [10] DeepVerse: 4D Autoregressive Video Generation as a World Model (Chen Junyi, 2025) View paper
  - [12] Gaussianprediction: Dynamic 3d gaussian prediction for motion extrapolation and free view synthesis (Boming Zhao, 2024) View paper
  - [21] WorldReel: 4D Video Generation with Consistent Geometry and Motion Modeling (Shaoheng Fang, 2025) View paper
  - Video Diffusion with Explicit 3D Constraints (2 papers)
  - [4] UniGeo: Taming Video Diffusion for Unified Consistent Geometry Estimation (Sun, 2025) View paper
  - [17] World-consistent Video Diffusion with Explicit 3D Modeling (Qihang Zhang, 2025) View paper
- Geometry-Conditioned Video Generation
  - 3D Proxy-Based Video Synthesis (3 papers)
  - [6] Gen3c: 3d-informed world-consistent video generation with precise camera control (Xuanchi Ren, 2025) View paper
  - [11] Shape-for-Motion: Precise and Consistent Video Editing With 3D Proxy (Liu Yuhao, 2025) View paper
  - [15] Human-VDM: Learning Single-Image 3D Human Gaussian Splatting from Video Diffusion Models (Zhibin Liu, 2024) View paper
  - Geometry-Guided Temporal Consistency (2 papers)
  - [24] Geometry-guided Online 3D Video Synthesis with Multi-View Temporal Consistency (Hyunho Ha, 2025) View paper
  - [27] Geometrically Consistent Light Field Synthesis Using Repaint Video Diffusion Model (Soyoung Yoon, 2024) View paper
  - Scene Composition with Geometric Realism (1 papers)
  - [16] Geosim: Realistic video simulation via geometry-aware composition for self-driving (Yun Chen, 2021) View paper
- Geometry Estimation from Video
  - Diffusion Prior-Based Geometry Prediction (1 papers)
  - [7] Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors (XU Tian-xing, 2025) View paper

    - Multi-View Consistent 3D Reconstruction (1 papers)
    - [23] ObjFiller-3D: Consistent Multi-view 3D Inpainting via Video Diffusion Models (Liu Jie, 2025) View paper
    - Video-to-3D Scene Understanding (2 papers)
    - [19] 3D Video Models through Point Tracking, Reconstructing and Forecasting (Chu, 2025) View paper
    - [25] Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors (Zheng Duo, 2025) View paper
- World Models for Embodied Intelligence
    - Action-Conditioned Video Prediction (2 papers)
    - [1] Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving (Yuqi Wang, 2024) View paper
    - [22] Pandora: Towards General World Model with Natural Language Actions and Video States (Xiang Jiannan, 2024) View paper
    - Physics-Informed World Modeling (2 papers)
    - [9] Robot Learning from a Physical World Model (Jiageng Mao, 2025) View paper
    - [20] RoboScape: Physics-informed Embodied World Model (Shang Yu, 2025) View paper
    - Data Generation for Embodied Learning (1 papers)
    - [5] Gigaworld-0: World models as data engine to empower embodied ai (GigaWorld Team, 2025) View paper
- Controllable 3D Scene Generation
    - Camera-Controlled Scene Synthesis (1 papers)
    - [8] Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation (Huang Tian-yu, 2025) View paper
    - Map-Conditioned Dynamic Scene Generation (1 papers)
    - [3] Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models (Lu, 2025) View paper
    - Multi-View Shared World Generation (1 papers)
    - [18] IC-World: In-Context Generation for Shared World Modeling (Fan Wu, 2025) View paper
- Temporal Prediction with Spatial Reasoning
    - Novel View Synthesis with Temporal Extrapolation (1 papers)
    - [13] Forecasting Future Videos from Novel Views via Disentangled 3D Scene Representation (Yarram, 2024) View paper
    - Human Motion Forecasting in 3D (2 papers)
    - [26] Learning Dynamic 3D Geometry and Texture for Video Face Swapping (C. Otto, 2022) View paper
    - [28] Predicting 3D Human Dynamics from Video (Jason Zhang, 2019) View paper
    - Scene Perception and Forecasting Systems (1 papers)
    - [29] Video Perception and Forecasting Models for Autonomous Systems (Yarram, n.d.) View paper

### Narrative

Core task: Geometry-consistent world modeling via unified video and 3D prediction. This emerging field seeks to bridge the gap between temporal video generation and explicit 3D geometric reasoning, enabling models that can predict future visual observations while maintaining coherent spatial structure. The taxonomy reveals several complementary research directions: Unified Video-3D Representation Learning explores joint embeddings and cross-modal supervision strategies that tie pixel-level dynamics to underlying geometry; Geometry-Conditioned Video Generation focuses on using depth, camera poses, or 3D scene representations to guide synthesis; Geometry Estimation from Video tackles the inverse problem of recovering spatial structure from temporal observations; World Models for Embodied Intelligence emphasizes predictive models for robotic planning and interaction; Controllable 3D Scene Generation addresses user-driven spatial content creation; and Temporal Prediction with Spatial Reasoning combines forecasting with geometric awareness. Representative works like Aether[2] and Infinicube[3] demonstrate how explicit 3D representations can scaffold video prediction, while approaches such as Driving Future[1] and Gigaworld[5] show the value of geometry-aware modeling in autonomous driving contexts.

A particularly active line of inquiry centers on cross-modal supervision, where video and geometry signals mutually constrain one another during training. FantasyWorld[0] exemplifies this approach by jointly learning video generation and 3D prediction through shared representations, closely aligning with Aether[2] and Geometry Forcing[14], which similarly enforce geometric consistency across modalities. In contrast, works like Gen3c[6] and GeometryCrafter[7] emphasize conditioning video synthesis on pre-extracted or user-specified geometry rather than learning both modalities end-to-end. Meanwhile, methods such as GaussianPrediction[12] and Novel View Forecasting[13] focus on predicting explicit 3D structures (e.g., Gaussian splats) to enable temporally coherent novel views. FantasyWorld[0] sits within the cross-modal supervision cluster, distinguishing itself by tightly coupling video and geometry learning rather than treating geometry as a fixed input or separate output, thereby enabling richer feedback between spatial and temporal reasoning.

## Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Aether: Geometric-aware unified world modeling

**Authors**: Zhu, Haoyi, Aether Team, Wang Yifan, Haoyi Zhu, et al. (23 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

The integration of geometric reconstruction and generative modeling remains a critical challenge in developing AI systems capable of human-like spatial reasoning. This paper proposes Aether, a unified framework that enables geometry-aware reasoning in world models by jointly optimizing three core capabilities: (1) 4D dynamic reconstruction, (2) action-conditioned video prediction, and (3) goal-conditioned visual planning. Through task-interleaved feature learning, Aether achieves synergistic kno...

#### Relationship Analysis

Both papers belong to the Cross-Modal Supervision for Joint Video-Geometry Learning category, employing bidirectional supervision where geometry guides video generation and video priors regularize 3D prediction. They overlap in their unified approach to video-3D modeling, with both augmenting video foundation models with geometric branches to achieve geometry-consistent generation. The key difference is that FantasyWorld focuses on augmenting frozen video models with a trainable geometric branch for implicit 3D field prediction, while Aether post-trains video diffusion models on synthetic 4D data to jointly optimize reconstruction, action-conditioned prediction, and visual planning using camera trajectories as action representations.

---

### 2. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

**Authors**: Wu Haoyu, Wu Diankun, Haoyu Wu, He Tianyu, Diankun Wu, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

### Abstract

Videos inherently represent 2D projections of a dynamic 3D world. However, our analysis suggests that video diffusion models trained solely on raw video data often fail to capture meaningful geometric-aware structure in their learned representations. To bridge this gap between video diffusion models and the underlying 3D nature of the physical world, we propose Geometry Forcing, a simple yet effective method that encourages video diffusion models to internalize latent 3D representations. Our key...

### Relationship Analysis

Both papers belong to the Cross-Modal Supervision for Joint Video-Geometry Learning category, using bidirectional supervision between video generation and 3D prediction. They overlap in their approach of coupling video diffusion models with geometric foundation models (VGGT) to achieve geometry-consistent video generation. However, FantasyWorld introduces a unified architecture with asymmetric dual branches (PCB and IRG blocks) that jointly predicts video latents and implicit 3D fields in a single forward pass, while Geometry Forcing takes a different approach by aligning intermediate diffusion features with frozen VGGT representations through Angular and Scale Alignment losses without modifying the video model architecture.

## Contributions Analysis

**Overall novelty summary.** FantasyWorld proposes a geometry-enhanced framework that augments frozen video foundation models with a trainable geometric branch, enabling joint video and 3D field modeling in a single forward pass. The paper resides in the 'Cross-Modal Supervision for Joint Video-Geometry Learning' leaf, which contains only three papers total. This represents a relatively sparse research direction within the broader taxonomy of 29 papers across multiple branches, suggesting the specific approach of bidirectional supervision between video and geometry modules is still emerging rather than saturated.

The taxonomy reveals that FantasyWorld's leaf sits within the larger 'Unified Video-3D Representation Learning' branch, which also includes sibling directions like '4D Dynamic Scene Representation' and 'Video Diffusion with Explicit 3D Constraints'. Neighboring branches pursue related but distinct goals: 'Geometry-Conditioned Video Generation' uses pre-extracted geometry as input rather than learning it jointly, while 'Geometry Estimation from Video' focuses on the inverse problem of recovering structure from observations. The scope note for FantasyWorld's leaf explicitly excludes unidirectional geometry-to-video guidance, positioning this work in a narrower methodological space where video priors actively regularize 3D prediction.

Among 30 candidates examined, the contribution-level analysis reveals mixed novelty signals. The core framework and cross-branch supervision mechanism each examined 10 candidates and found 2 potentially refutable prior works, suggesting some overlap with existing joint video-geometry approaches. However, the third contribution—generalizable 3D features for downstream tasks without fine-tuning—examined 10 candidates with zero refutable matches, indicating this aspect may be more distinctive. The limited search scope means these statistics reflect top-30 semantic matches rather than exhaustive coverage, so additional related work may exist beyond this sample.

Given the sparse population of the specific taxonomy leaf and the modest search scale, FantasyWorld appears to occupy a relatively novel position within cross-modal video-geometry learning, though the framework-level contributions show some prior art overlap among the examined candidates. The downstream generalization aspect seems less explored in the sampled literature, while the bidirectional supervision mechanism aligns with a small cluster of recent works pursuing similar joint optimization strategies.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

### Contribution 1: FantasyWorld: Geometry-enhanced framework for unified video and 3D modeling

**Description**: The authors introduce FantasyWorld, a framework that extends frozen video foundation models by adding a trainable geometric branch. This enables simultaneous prediction of video latents and an implicit 3D field in one forward pass, bridging video generation and 3D perception without per-scene optimization.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Aether: Geometric-aware unified world modeling

**URL**: View paper

**Prior Art Analysis**

Aether[2] demonstrates that similar prior work exists in augmenting video foundation models with geometric branches for joint video and 3D modeling. Both papers present frameworks that extend video diffusion models with trainable geometric components to enable simultaneous video generation and 3D reasoning. Aether[2] explicitly describes a unified framework that 'integrates reconstruction, prediction, and planning through multi-task learning' and uses 'video diffusion models' as a base, augmented with geometric modeling capabilities. The architectural approach of adding geometric branches to frozen or partially frozen video models, and the goal of producing both video outputs and 3D representations in a single forward pass, are core similarities that challenge the novelty claim of FantasyWorld being the first to propose this unified video-3D modeling approach.

**Evidence**

Evidence 1 - **Rationale**: Both papers describe unified frameworks that integrate video generation with geometric/3D modeling capabilities, suggesting Aether[2] demonstrates prior work in this unified approach. - **Original**: we present f antasy world , a geometry-enhanced framework that augments frozen video foundation models with a trainable geometric branch, enabling joint modeling of video latents and an implicit 3d field in a single forward pass. - **Candidate**: this work introduces aether , a unified world model that integrates reconstruction, prediction, and planning through multi-task learning on synthetic 4d data.

Evidence 2 - **Rationale**: Both frameworks build upon pre-trained video models and extend them with geometric capabilities, indicating Aether[2] pursued a similar architectural strategy before FantasyWorld. - **Original**: unified video-3d modeling: we propose f antasy world , a geometry-enhanced framework that jointly predicts video latents and an implicit 3d field through a single backbone, preserving imaginative priors while exposing explicit geometry. - **Candidate**: aether leverages pre-trained video generation models [28, 77] and is further refined via post-training with synthetic 4d data.

Evidence 3 - **Rationale**: Both papers describe augmenting pre-trained video diffusion models with additional geometric processing components, demonstrating that Aether[2] implemented this approach prior to FantasyWorld's submission. - **Original**: we split the backbone of video foundation models (i.e., wan2.1 in our case) into preconditioning blocks (pcb) that inject video priors and stabilize latents, and integrated reconstruction and generation blocks (irg) that fuse spatiotemporal tokens with a geometry co-encoder to predict a geometry-awa... - **Candidate**: we initialize aether with pre-trained cogvideo-5bi2v [77] weights, excluding the additional input and output projection layer channels for depth and raymap action trajectories, which are initialized to zero.

Evidence 4 - **Rationale**: Both frameworks emphasize producing 3D/4D representations alongside video in a single forward pass without per-scene optimization, indicating Aether[2] achieved this capability before FantasyWorld. - **Original**: as a result, our model generates camera-conditioned video features alongside an explicit 3d representation in a single forward pass, without relying on additional 3d reconstruction (e.g., nerf or 3dgs) or iterative memory refinement - **Candidate**: at their core, three capabilities stand out: first, perception equips the system with the ability to capture the intricate four-dimensional (4d) changes-integrating spatial and temporal information-that are essential for understanding the physical world

---

### 2. Controllable video generation: A survey

**URL**: View paper

**Brief Assessment**

Controllable Video Survey[36] is a comprehensive survey paper that reviews existing methods in controllable video generation. It does not present FantasyWorld or claim to introduce a geometry-enhanced framework for unified video and 3D modeling, and therefore does not challenge the novelty of this contribution.

---

### 3. Can Video Diffusion Model Reconstruct 4D Geometry?

**URL**: View paper

**Brief Assessment**

Video Diffusion Geometry[37] focuses on 4D reconstruction from monocular video using pointmap representations, not on augmenting video foundation models with geometric branches for joint video-3D generation as FantasyWorld does.

---

### 4. Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video

**URL**: View paper

**Brief Assessment**

Uni4D[31] focuses on 4D reconstruction from casual videos using multi-stage optimization of pretrained models without training. FantasyWorld trains a geometry branch within a video diffusion model for joint generation. The approaches differ fundamentally in methodology and objectives.

---

### 5. VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction

**URL**: View paper

**Brief Assessment**

VLM-3R[34] focuses on vision-language models for 3D spatial reasoning from monocular video, not on augmenting video foundation models with geometric branches for joint video-3D generation as in FantasyWorld.

---

### 6. V3d: Video diffusion models are effective 3d generators

**URL**: View paper

**Brief Assessment**

V3d[33] focuses on fine-tuning video diffusion models for multi-view image generation followed by 3D reconstruction, rather than jointly modeling video latents and implicit 3D fields in a unified forward pass with cross-branch supervision as in FantasyWorld.

---

### 7. Harnessing Foundation Models for Robust and Generalizable 6-DOF Bronchoscopy Localization

**URL**: View paper

**Brief Assessment**

Bronchoscopy Localization[32] focuses on medical bronchoscopy navigation using foundation models for depth estimation and landmark detection in surgical contexts, not on general video-to-3D world modeling or extending video foundation models with geometric branches for creative content generation.

---

### 8. Bridging the Gap Between Multimodal Foundation Models and World Models

**URL**: View paper

**Brief Assessment**

Foundation World Models[35] is a thesis overview covering discriminative and generative world modeling across multiple chapters. It does not present a unified video-3D framework with a trainable geometric branch like FantasyWorld.

---

### 9. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

**URL**: View paper

**Prior Art Analysis**

Geometry Forcing[14] demonstrates prior work that augments video diffusion models with geometric branches for joint video-3D modeling. Both papers extend frozen video foundation models with trainable geometric components to enable simultaneous video generation and 3D representation prediction in a single forward pass. Geometry Forcing[14] explicitly describes marrying video diffusion with 3D representation through alignment with a pretrained geometric foundation model (VGGT), achieving unified video-geometry modeling without per-scene optimization—the same core architectural principle claimed as novel in the original paper.

**Evidence**

Evidence 1 - **Rationale**: Both papers describe augmenting video diffusion models with geometric components. Geometry Forcing[14] aligns video diffusion features with a pretrained 3D foundation model (VGGT), while the original paper adds a trainable geometric branch. Both achieve joint video-3D modeling in a unified framework. - **Original**: we present f antasy world , a geometry-enhanced framework that augments frozen video foundation models with a trainable geometric branch, enabling joint modeling of video latents and an implicit 3d field in a single forward pass. - **Candidate**: we propose geometry forcing (gf), a simple yet effective approach that encourages video diffusion models to internalize 3d representations during training. inspired by recent advances in semantic representation alignment (repa) for image diffusion models (yu et al., 2024a), we align intermediate fea...

Evidence 2 - **Rationale**: Both papers describe mechanisms where geometric information guides video generation to produce 3D-consistent representations. Geometry Forcing[14] achieves this through alignment with geometric foundation model features, demonstrating the concept of geometry-guided video generation existed prior to the original submission. - **Original**: our approach introduces cross-branch supervision, where geometry cues guide video generation and video priors regularize 3d prediction, thus yielding consistent and generalizable 3d-aware video representations. - **Candidate**: to bridge this gap between video diffusion models and the underlying 3d nature of the physical world, we propose geometry forcing, a simple yet effective method that encourages video diffusion models to internalize latent 3d representations. our key insight is to guide the model's intermediate repre...

Evidence 3 - **Rationale**: Both papers claim that their unified video-3D representations can serve downstream tasks without per-scene optimization. Geometry Forcing[14] explicitly mentions this capability, demonstrating prior work with similar claims about reusable 3D-aware features. - **Original**: notably, the resulting latents from the geometric branch can potentially serve as versatile representations for downstream 3d tasks such as novel view synthesis and navigation, without requiring per-scene optimization or fine-tuning. - **Candidate**: moreover, the ability to reconstruct explicit geometry during inference provides a structured and interpretable form of memory, which can be further utilized to support long-term world modeling and reasoning.

---

### 10. Geovideo: Introducing geometric regularization into video generation model

**URL**: View paper

**Brief Assessment**

GeoVideo[30] focuses on introducing geometric regularization through depth prediction and multi-view consistency losses during video generation training. In contrast, FantasyWorld proposes a dual-branch architecture that jointly predicts video latents and an implicit 3D field in a single forward pass without per-scene optimization, representing a fundamentally different architectural approach to coupling video generation with 3D reasoning.

---

## Contribution 2: Cross-branch supervision mechanism between video and geometry

**Description**: The authors propose a cross-branch supervision strategy where geometric cues inform video generation while video priors regularize 3D prediction. This bidirectional constraint mechanism produces consistent and generalizable 3D-aware video representations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. CONSISTENCY ENHANCED DEEP LEARNING FOR VISUAL PERCEPTION DATA OF STRUCTURAL HEALTH MONITORING

**URL**: View paper

**Brief Assessment**

Consistency Enhanced Learning[56] focuses on structural health monitoring using visual perception data with mutual supervision between different modalities. The candidate does not address video generation or 3D-aware representations for world modeling, which are central to the original paper's contribution.

---

### 2. MGSR: 2D/3D Mutual-boosted Gaussian Splatting for High-fidelity Surface Reconstruction under Various Light Conditions

**URL**: View paper

**Brief Assessment**

MGSR[50] focuses on mutual supervision between 2D Gaussian Splatting (2DGS) and 3D Gaussian Splatting (3DGS) branches for surface reconstruction under varying light conditions, not on cross-branch supervision between video generation and geometric inference for 3D-aware world modeling.

---

### 3. mmHand: 3D hand pose estimation using millimeter-wave radar

**URL**: View paper

**Brief Assessment**

mmHand[52] focuses on hand pose estimation using millimeter-wave radar with cross-modal supervision between radar and camera modalities, not on video-geometry supervision for 3D world modeling.

---

### 4. GVLM: Geometry Grounded Vision Language Model with Unified 3D Reconstruction and Spatial Reasoning

**URL**: View paper

**Brief Assessment**

GVLM[53] focuses on a dual-expert architecture for 3D reconstruction and spatial reasoning tasks, not on cross-branch supervision between video generation and geometry prediction. The candidate's geometric and semantic experts interact via shared self-attention rather than the bidirectional cross-branch supervision mechanism described in the original paper.

---

### 5. AtlantaSDF: Neural 3D Indoor Scene Reconstruction with the Atlanta-world Assumption

**URL**: View paper

**Brief Assessment**

AtlantaSDF[55] focuses on neural 3D indoor scene reconstruction using the Atlanta-world assumption with MLP-based geometry representation. The candidate's mention of 'mutual supervision method' appears in a different context (ground/wall constraints) rather than bidirectional cross-branch supervision between video generation and 3D prediction branches as proposed in the original paper.

---

### 6. The 3D Modeling Technology for Substation Surface Temperature Distribution Based on Thermal-Geometric Joint

**URL**: View paper

**Brief Assessment**

Substation Temperature Modeling[54] focuses on thermal-geometric reconstruction for industrial inspection using infrared images and 3D Gaussian splatting, not video generation or 3D-aware video representations.

---

### 7. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera

**URL**: View paper

**Brief Assessment**

Geometric Constraints Monocular[49] focuses on self-supervised learning from monocular video using geometric constraints (epipolar geometry, depth-flow consistency) to jointly optimize depth, optical flow, and camera parameters. It does not propose a cross-branch supervision architecture between separate video generation and 3D geometry branches as in the original paper's dual-branch design with bidirectional cross-attention.

---

### 8. JOG3R: Towards 3D-Consistent Video Generators

**URL**: View paper

**Prior Art Analysis**

JOG3R[48] demonstrates prior work on bidirectional supervision between video generation and geometric reconstruction tasks. The candidate paper proposes joint training of video generation with 3D point map estimation using both photometric generation errors and 3D-aware reconstruction losses. This creates a mutual reinforcement mechanism where geometric supervision improves 3D-consistency of generated videos, while video features enhance reconstruction quality. The architecture unifies a video diffusion transformer (DiT) with a DUST3R-like reconstruction head, enabling cross-task feature sharing and joint optimization through combined loss functions.

**Evidence**

Evidence 1 - **Rationale**: Both papers describe bidirectional supervision between video generation and geometric tasks. The candidate explicitly proposes joint training with both photometric and 3D-aware losses, demonstrating the same core concept of mutual supervision. - **Original**: we introduce cross-branch supervision, where geometry cues guide video generation and video priors regularize 3d prediction, thus yielding consistent and generalizable 3d-aware video representations - **Candidate**: we propose to jointly train for the

two tasks, using photometric generation and 3d aware errors. specifically, we find that sota video generation and camera pose estimation (i.e., dust3r [79]) networks share common structures, and propose an architecture that unifies the two

Evidence 2 - **Rationale**: The candidate demonstrates the same bidirectional supervision mechanism through joint optimization with both generation and reconstruction losses, where each task supervises and improves the other. - **Original**: cross-branch supervision, where geometry cues guide video generation and video priors regularize 3d prediction - **Candidate**: we further fine-tune both the video generator and the reconstruction heads jointly to perform generation and reconstruction tasks. for training, we consider two losses: generation loss lgen and reconstruction loss lrec. the generation loss lgen is the common objective in training diffusion models th...

Evidence 3 - **Rationale**: The candidate validates that geometric and video tasks mutually reinforce each other without degradation, confirming the bidirectional supervision concept claimed in the original paper. - **Original**: we introduce constraints that let geometry supervise video features and video priors regularize 3d prediction, ensuring 3d-consistent frames inference - **Candidate**: this confirms that the two tasks are compatible and do not degrade each other's performance. our full method, jog3r, performs overall better than pre-trained dust3r

Evidence 4 - **Rationale**: Both papers formulate the cross-branch supervision through a combined loss function that integrates video generation loss with geometric supervision losses, demonstrating the same technical approach. - **Original**: the final objective is: ltotal = ez0,ϵ,t,c h ∥eθ(zt, t, c) - ϵ∥2 i + λlgeo combining the standard diffusion loss with geometry supervision which aggregates depth, point map, and camera supervision - **Candidate**: in our experiments, we study the effect of different combinations of these two losses. specifically, activating lrec alone analyzes the native 3d awareness of the video generation features while using them in conjunction, ltotal = lgen + λlrec (we empirically set λ = 1) investigates if the features ...

---

### 9. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling

**URL**: View paper

**Prior Art Analysis**

Geometry Forcing[14] demonstrates prior work on bidirectional supervision between video and geometry branches. The paper introduces angular alignment and scale alignment objectives that enforce geometric consistency in video features while using video priors to inform 3D prediction. This bidirectional constraint mechanism—where geometric representations guide video generation and video features are regularized by geometric priors—directly parallels the cross-branch supervision claimed as novel in the original paper.

**Evidence**

Evidence 1 - **Rationale**: Both papers describe bidirectional supervision mechanisms. Geometry Forcing[14] aligns video diffusion features with geometric representations through angular and scale alignment, achieving mutual reinforcement between video and geometry—the same principle as cross-branch supervision in the original paper. - **Original**: our approach introduces cross-branch supervision, where geometry cues guide video generation and video priors regularize 3d prediction, thus yielding consistent and generalizable 3d-aware video representations. - **Candidate**: to improve the geometric consistency of the learned representations, we introduce two complementary alignment objectives: angular alignment and scale alignment. these objectives are designed to align the latent features of the diffusion model with intermediate representations from a pretrained geome...

Evidence 2 - **Rationale**: Geometry Forcing[14] explicitly describes using geometric features (from VGGT) to supervise video diffusion hidden states through alignment losses, demonstrating the concept of geometry-to-video supervision existed prior to the original submission. - **Original**: we introduce constraints that let geometry supervise video features and video priors regularize 3d prediction, ensuring 3d-consistent frames inference. - **Candidate**: angular alignment enforces directional correspondence between the hidden states of the diffusion model, denoted by h, and specified target features, denoted by y. we select intermediate features from the transformer backbone of vggt (wang et al., 2025) as y, as these features preserve both local and...

---

### 10. 3D-Aware Video Stabilization via Reconstruction and Rendering

**URL**: View paper

**Brief Assessment**

3D Video Stabilization[51] focuses on video stabilization using 3D reconstruction and rendering techniques. The candidate's sparse context mentions geometry supervision but does not demonstrate a cross-branch mechanism between video generation and 3D prediction as proposed in the original paper.

---

## Contribution 3: Generalizable 3D features for downstream tasks without fine-tuning

**Description**: The geometric branch produces latent representations that can serve as versatile features for downstream 3D tasks like novel view synthesis and navigation, eliminating the need for per-scene optimization or task-specific fine-tuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d

**URL**: View paper

**Brief Assessment**

KITTI 360[44] is a dataset paper focused on urban scene understanding with semantic annotations and benchmarks. It does not address generalizable 3D features for novel view synthesis and navigation without per-scene optimization, which is the core contribution being evaluated.

---

### 2. Learning generalizable feature fields for mobile manipulation

**URL**: View paper

**Brief Assessment**

Generalizable Feature Fields[41] focuses on mobile manipulation tasks using generalizable neural feature fields for navigation and grasping, not on video generation and novel view synthesis from video foundation models as in the original paper.

---

### 3. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis

**URL**: View paper

**Brief Assessment**

GPS Gaussian[40] focuses on real-time human novel view synthesis using pixel-wise 3D Gaussian splatting, not on producing generalizable 3D features for diverse downstream tasks like navigation. The candidate's features are specific to human rendering tasks rather than general 3D reasoning applications.

---

### 4. Generalizable human gaussians for sparse view synthesis

**URL**: View paper

**Brief Assessment**

Generalizable Human Gaussians[43] focuses specifically on human rendering from sparse views using 2D UV space representations, not on general 3D scene features for navigation and novel view synthesis tasks as in the original paper.

---

### 5. Generalizable 3D Gaussian Splatting for novel view synthesis

**URL**: View paper

**Brief Assessment**

Generalizable Gaussian Splatting[39] focuses specifically on novel view synthesis using 3D Gaussian representations, whereas the original paper proposes a broader framework for joint video-3D modeling with versatile features for multiple downstream tasks including both novel view synthesis and navigation.

---

### 6. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations

**URL**: View paper

**Brief Assessment**

Scene Representation Transformer[38] focuses on novel view synthesis from posed/unposed images using set-latent representations and transformer architectures. While it produces scene representations, it does not explicitly demonstrate generalizable 3D features for diverse downstream tasks like navigation without fine-tuning, which is the core claim of the original paper's geometric branch.

---

### 7. Advances in feed-forward 3d reconstruction and view synthesis: A survey

**URL**: View paper

**Brief Assessment**

Feed Forward Survey[47] focuses on categorizing existing feed-forward 3D reconstruction methods by representation type (NeRF, 3DGS, pointmap, etc.) rather than proposing a novel architecture. The original paper introduces a specific geometric branch architecture that produces task-agnostic 3D features through cross-branch supervision, which is architecturally distinct from the survey's taxonomy of existing methods.

---

### 8. High-fidelity novel view synthesis via splatting-guided diffusion

**URL**: View paper

**Brief Assessment**

Splatting Guided Diffusion[42] focuses on novel view synthesis from single/sparse images using splatting-guided diffusion, not on producing generalizable 3D features for diverse downstream tasks like navigation.

---

### 9. View-invariant policy learning via zero-shot novel view synthesis

**URL**: View paper

**Brief Assessment**

View Invariant Policy[45] focuses on learning viewpoint-invariant policies for robotic manipulation using novel view synthesis for data augmentation, not on producing generalizable 3D features for downstream tasks like the original paper's geometric branch.

---

### 10. Fwd: Real-time novel view synthesis with forward warping and depth

**URL**: View paper

**Brief Assessment**

Forward Warping Depth[46] focuses on real-time novel view synthesis using forward warping and explicit depth representations. It does not claim to produce generalizable 3D features for diverse downstream tasks like navigation, nor does it address the broader goal of task-agnostic 3D representations without fine-tuning.

---

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References


- [0] FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction View paper
- [1] Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving View paper
- [2] Aether: Geometric-aware unified world modeling View paper
- [3] Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models View paper
- [4] UniGeo: Taming Video Diffusion for Unified Consistent Geometry Estimation View paper
- [5] Gigaworld-0: World models as data engine to empower embodied ai View paper
- [6] Gen3c: 3d-informed world-consistent video generation with precise camera control View paper
- [7] Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors View paper
- [8] Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation View paper
- [9] Robot Learning from a Physical World Model View paper
- [10] DeepVerse: 4D Autoregressive Video Generation as a World Model View paper
- [11] Shape-for-Motion: Precise and Consistent Video Editing With 3D Proxy View paper
- [12] Gaussianprediction: Dynamic 3d gaussian prediction for motion extrapolation and free view synthesis View paper
- [13] Forecasting Future Videos from Novel Views via Disentangled 3D Scene Representation View paper
- [14] Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling View paper
- [15] Human-VDM: Learning Single-Image 3D Human Gaussian Splatting from Video Diffusion Models View paper
- [16] Geosim: Realistic video simulation via geometry-aware composition for self-driving View paper
- [17] World-consistent Video Diffusion with Explicit 3D Modeling View paper
- [18] IC-World: In-Context Generation for Shared World Modeling View paper
- [19] 3D Video Models through Point Tracking, Reconstructing and Forecasting View paper
- [20] RoboScape: Physics-informed Embodied World Model View paper
- [21] WorldReel: 4D Video Generation with Consistent Geometry and Motion Modeling View paper
- [22] Pandora: Towards General World Model with Natural Language Actions and Video States View paper
- [23] ObjFiller-3D: Consistent Multi-view 3D Inpainting via Video Diffusion Models View paper
- [24] Geometry-guided Online 3D Video Synthesis with Multi-View Temporal Consistency View paper
- [25] Learning from Videos for 3D World: Enhancing MLLMs with 3D Vision Geometry Priors View paper

- [26] Learning Dynamic 3D Geometry and Texture for Video Face Swapping View paper
- [27] Geometrically Consistent Light Field Synthesis Using Repaint Video Diffusion Model View paper
- [28] Predicting 3D Human Dynamics from Video View paper
- [29] Video Perception and Forecasting Models for Autonomous Systems View paper
- [30] Geovideo: Introducing geometric regularization into video generation model View paper
- [31] Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video View paper
- [32] Harnessing Foundation Models for Robust and Generalizable 6-DOF Bronchoscopy Localization View paper
- [33] V3d: Video diffusion models are effective 3d generators View paper
- [34] VLM-3R: Vision-Language Models Augmented with Instruction-Aligned 3D Reconstruction View paper
- [35] Bridging the Gap Between Multimodal Foundation Models and World Models View paper
- [36] Controllable video generation: A survey View paper
- [37] Can Video Diffusion Model Reconstruct 4D Geometry? View paper
- [38] Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations View paper
- [39] Generalizable 3D Gaussian Splatting for novel view synthesis View paper
- [40] Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis View paper
- [41] Learning generalizable feature fields for mobile manipulation View paper
- [42] High-fidelity novel view synthesis via splatting-guided diffusion View paper
- [43] Generalizable human gaussians for sparse view synthesis View paper
- [44] Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d View paper
- [45] View-invariant policy learning via zero-shot novel view synthesis View paper
- [46] Fwd: Real-time novel view synthesis with forward warping and depth View paper
- [47] Advances in feed-forward 3d reconstruction and view synthesis: A survey View paper
- [48] JOG3R: Towards 3D-Consistent Video Generators View paper
- [49] Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera View paper
- [50] MGSR: 2D/3D Mutual-boosted Gaussian Splatting for High-fidelity Surface Reconstruction under Various Light Conditions View paper
- [51] 3D-Aware Video Stabilization via Reconstruction and Rendering View paper
- [52] mmHand: 3D hand pose estimation using millimeter-wave radar View paper
- [53] GVLM: Geometry Grounded Vision Language Model with Unified 3D Reconstruction and Spatial Reasoning View paper
- [54] The 3D Modeling Technology for Substation Surface Temperature Distribution Based on Thermal-Geometric Joint View paper
- [55] AtlantaSDF: Neural 3D Indoor Scene Reconstruction with the Atlanta-world Assumption View paper
- [56] CONSISTENCY ENHANCED DEEP LEARNING FOR VISUAL PERCEPTION DATA OF STRUCTURAL HEALTH MONITORING View paper