FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction

ICLR 2026 · Poster

Yixiang Dai*[1] Fan Jiang*†‡[1] Chiyu Wang*[1] Mu Xu[1] YongGang Qi‡[2]

[1]AMAP, Alibaba Group [2]Beijing University of Posts and Telecommunications

*Equal Contribution †Project Leader ‡Corresponding Author

Motivation

Imaginative

Video Generation Models

➢ Strength: Strong 2D visual priors.

➢ Limitation: Weak 3D consistency.

Geometric

Feed-Forward 3D Reconstruction Models

➢ Strength: Fast and High-quality reconstruction.

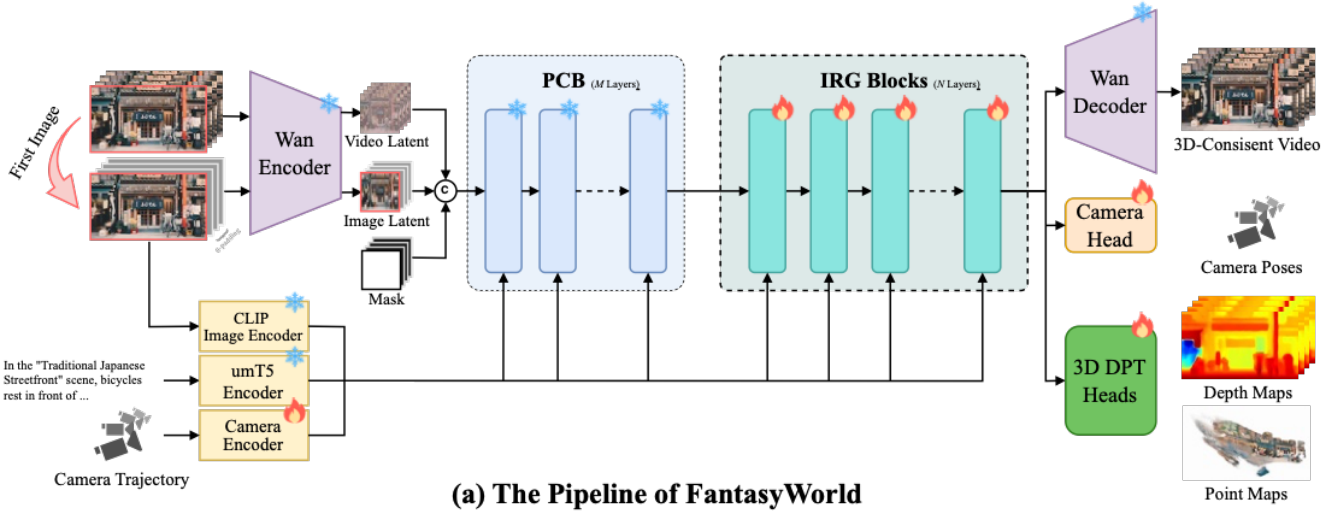
➢ Limitation: Reliant on acquiring high-quality 2D source data.

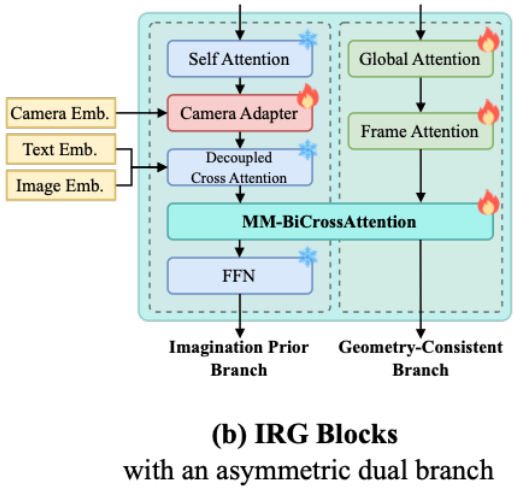
Our goal: Bidirectional mutual promotion in unified generation framework.

3D → 2D: Improve the geometric consistency of the 2D branch.

2D → 3D: Provide visual priors for the 3D branch.

Overview

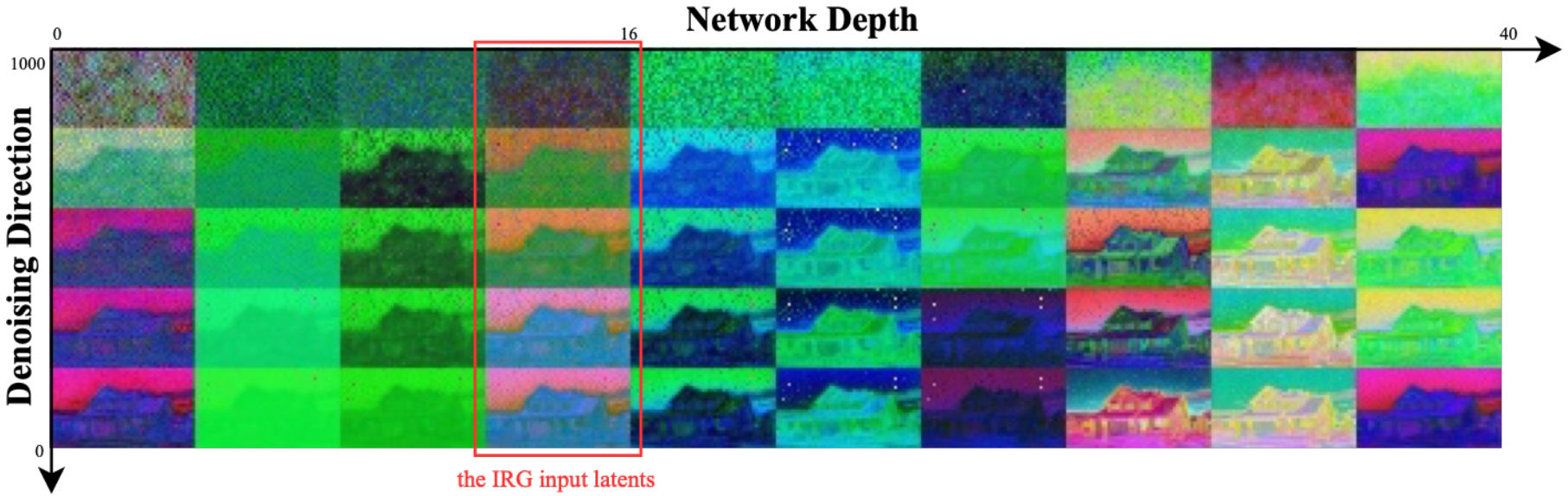




Given image + text + camera trajectory → video + implicit 3D field (single forward pass)

Preconditioning Blocks (PCB)

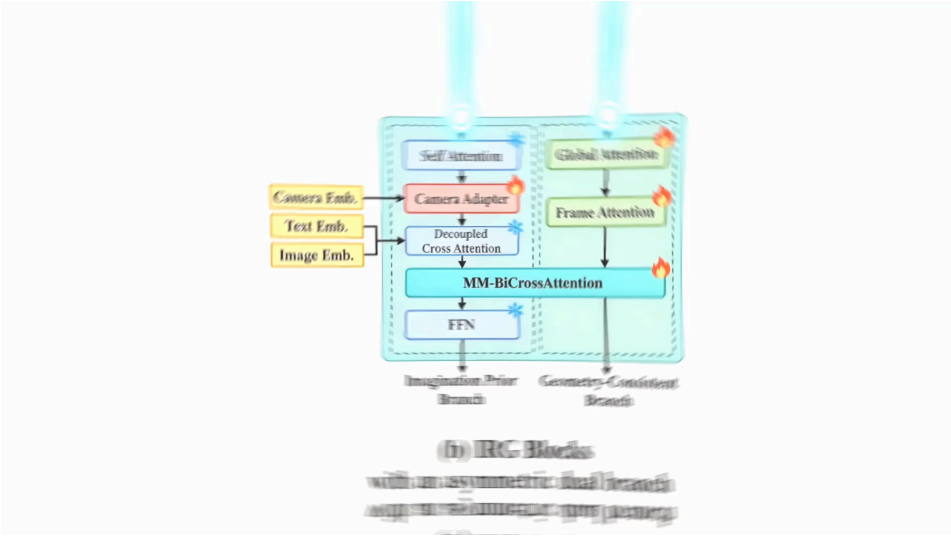
The features extracted by the PCB module ensures that the geometry branch is guided by meaningful signals rather than pure noise



Integrated Reconstruction & Generation Blocks (IRG)

Asymmetric dual-branch structure:

- ➢ Imaginative Branch: Frozen weights, Appearance synthesis
- ➢ Geometric Branch: Trainable, Explicit 3D reasoning
- ➢ Coupled via lightweight adapters + MM-BiCrossAttention

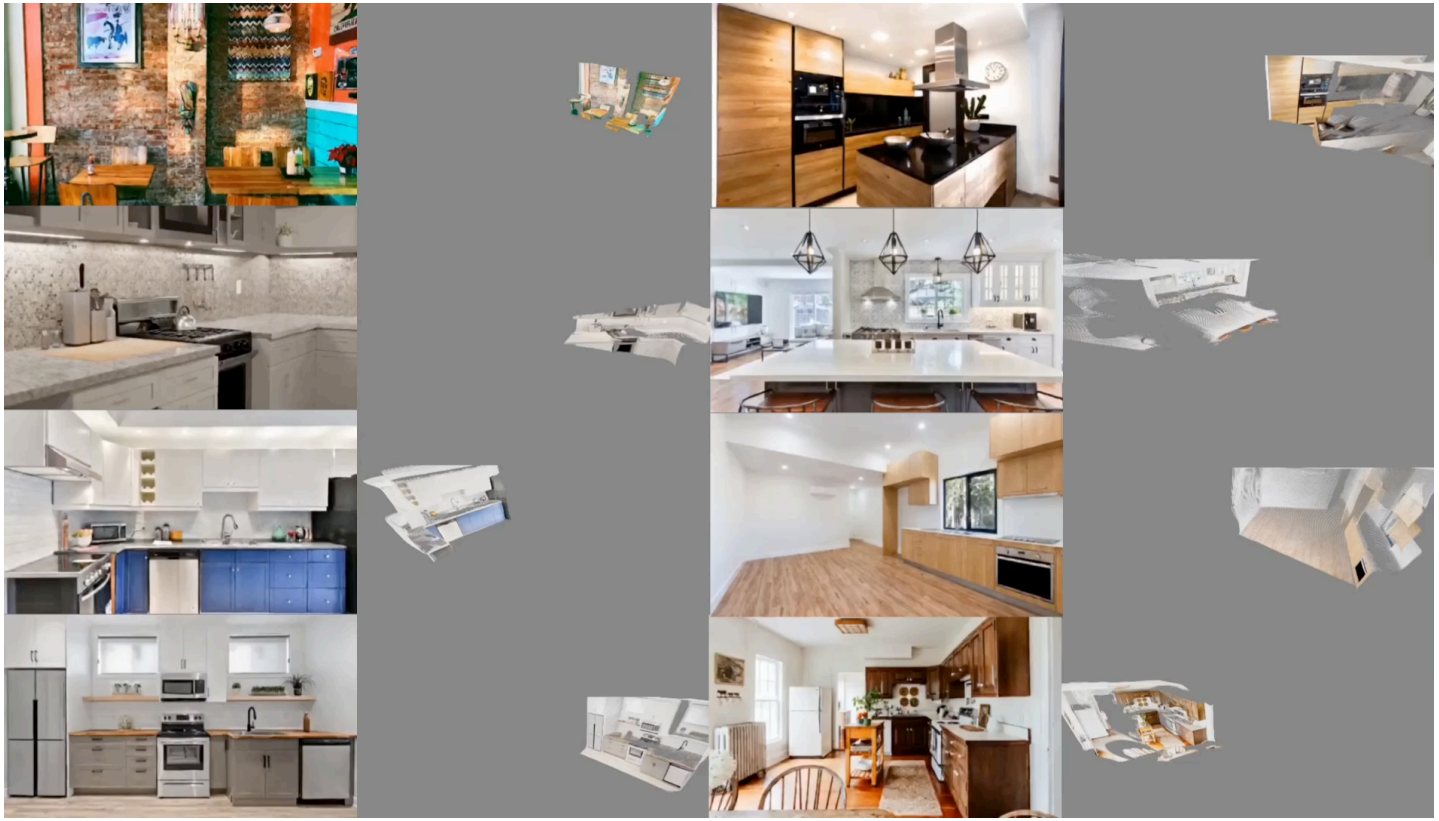


Results on WorldScore Leaderboard

Method	Motion	3D Consist.	Photo Consist.	Style Consist.	Camera Ctrl.	Object Ctrl.	Content Align.	Subjective Qual.
WonderWorld	Small	82.85±19.69	67.86±23.56	55.79±34.89	**92.32**	47.63	**79.09**	**69.03**
AETHER	Small	79.84±14.68	58.68±38.59	72.09±32.62	57.44	52.26	28.06	41.11
Uni3C	Small	78.59±21.08	85.48±20.98	88.32±18.47	62.94	45.83	47.40	57.00
Voyager	Small	56.00±26.32	80.68±16.32	72.89±29.78	45.92	**57.69**	48.36	44.74
Ours w/o 3D	Small	79.77±16.06	83.86±8.73	92.54±12.90	57.94	37.33	43.31	55.85
Ours w/ 3D	Small	**83.31**±14.24	**86.11**±7.97	**94.22**±9.11	57.05	34.46	38.45	57.40
WonderWorld	Large	63.70±24.37	3.22±8.47	35.95±33.47	**96.28**	38.61	**97.10**	**72.46**
AETHER	Large	63.97±17.39	33.11±23.99	61.99±32.24	4.43	34.78	33.69	35.09
Uni3C	Large	73.95±17.55	46.78±32.64	71.43±29.38	8.69	34.28	77.88	51.12
Voyager	Large	13.82±19.96	9.52±17.17	61.34±35.29	0.00	**49.23**	64.10	39.21
Ours w/o 3D	Large	72.06±20.14	56.98±23.60	81.59±22.23	9.32	34.44	75.85	46.96
Ours w/ 3D	Large	**74.83**±16.31	**60.61**±21.39	**82.02**±19.56	11.24	31.96	77.20	50.46

Qualitative Results

More Results



Summary

- FantasyWorld: Unified feed-forward model for joint video + 3D generation
- Cross-branch supervision between video priors and geometry cues
- Ranked 1st on WorldScore Leaderboard
- No per-scene optimization — generalizable 3D features for downstream tasks

---

Code & Models: https://github.com/Fantasy-AMAP/fantasy-world

Paper: arxiv.org/abs/2509.21657
