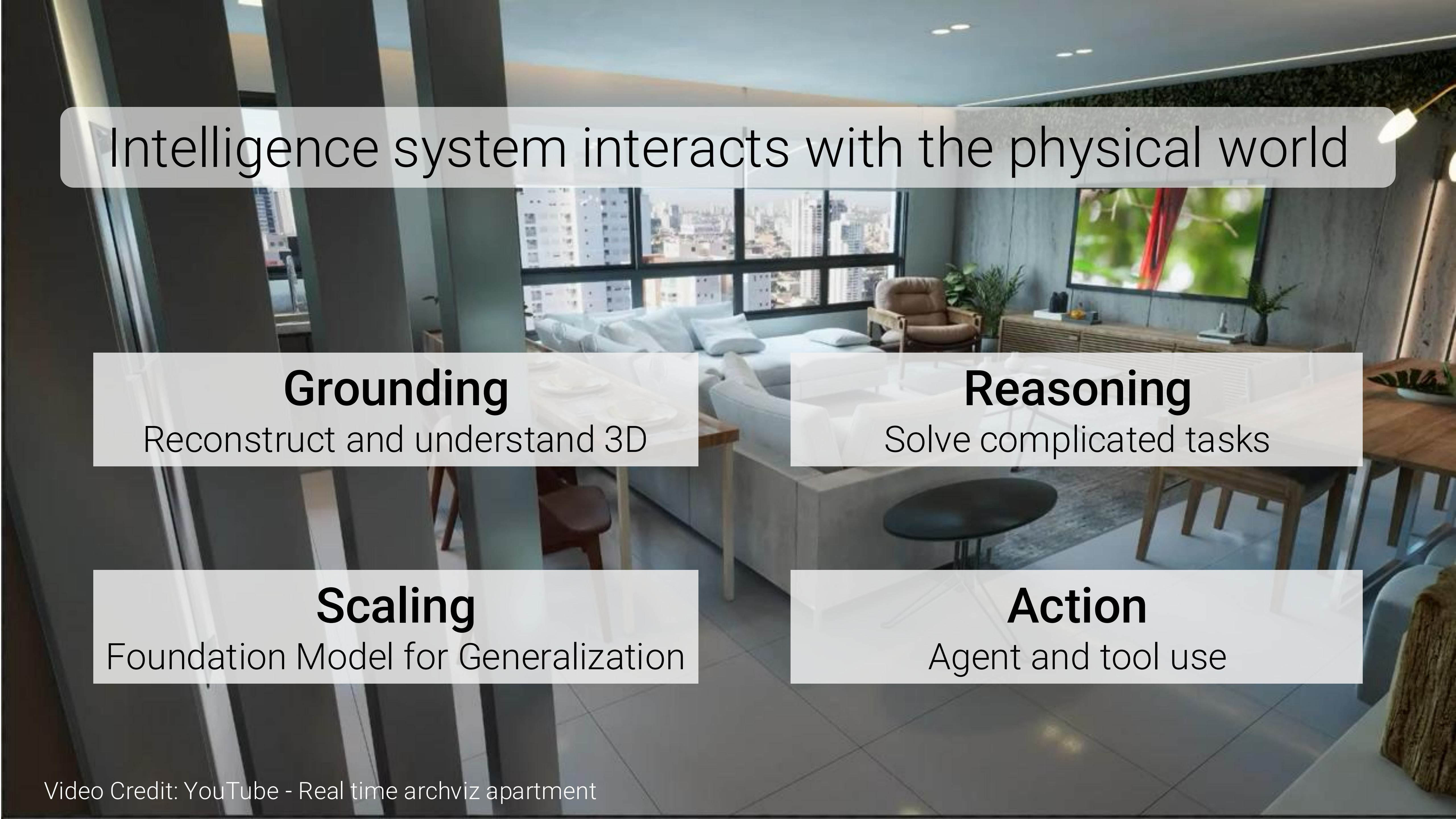
Building Visual Intelligence

Songyou Peng

Google DeepMind

Meta

Sep 29, 2025



Intelligence system interacts with the physical world

Grounding
Reconstruct and understand 3D

Reasoning
Solve complicated tasks

Scaling
Foundation Model for Generalization

Action
Agent and tool use

Video Credit: YouTube - Real time archviz apartment

Building Visual Intelligence

Grounding
Reconstruct and understand 3D

Reasoning
Solve complicated tasks

Scaling
Foundation Model for Generalization

Action
Agent and tool use

Building Visual Intelligence

Grounding
Reconstruct and understand 3D

Reasoning
Solve complicated tasks

Scaling
Foundation Model for Generalization

Action
Agent and tool use

My PhD

Learn to **Reconstruct** and **understand** 3D World

ConvOccNet
ECCV 2020 (Spotlight)

MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

NICE-SLAM
CVPR 2022

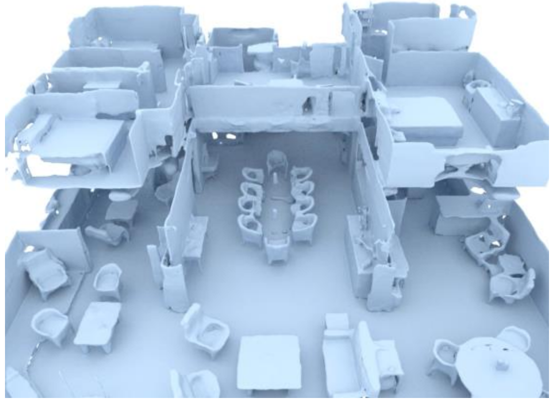
NICER-SLAM
3DV 2024 (Oral)

UNISURF
ICCV 2021 (Oral)

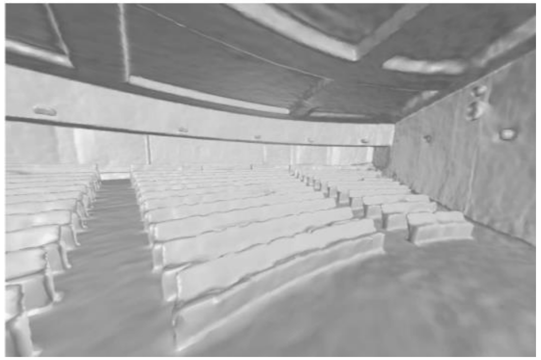
OpenScene
CVPR 2023

My PhD

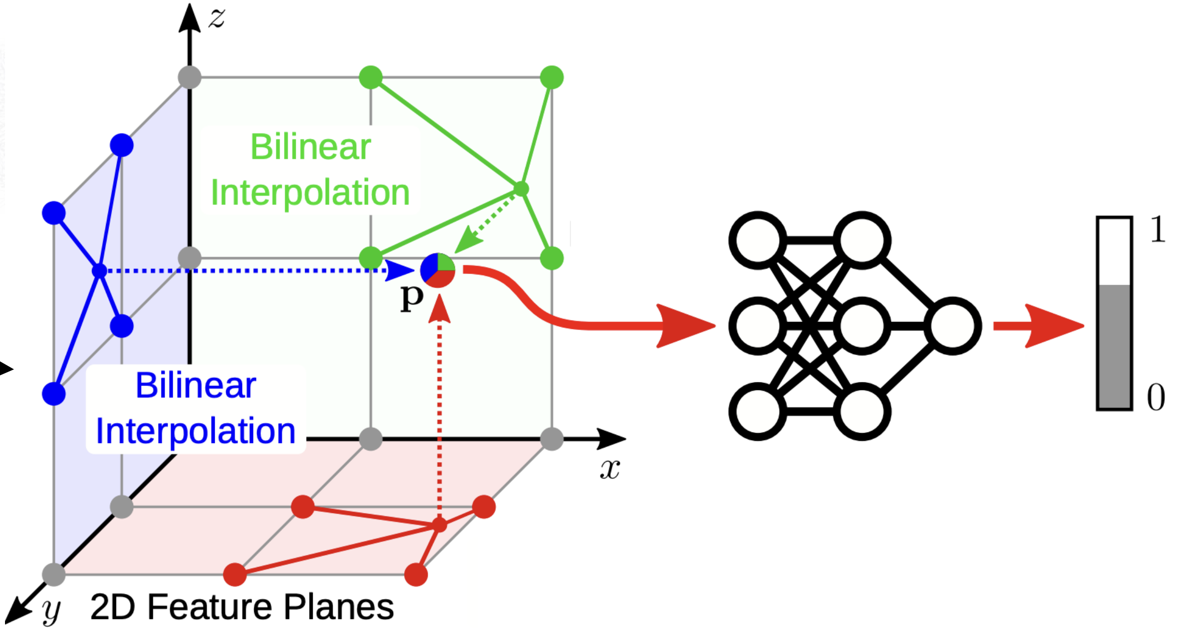
Learn to Reconstruct and Understand 3D World



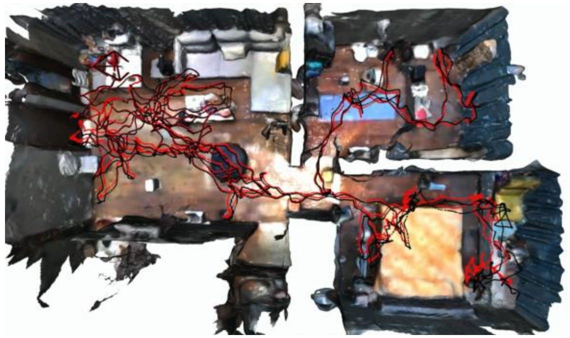
ConvOccNet
ECCV 2020 (Spotlight)



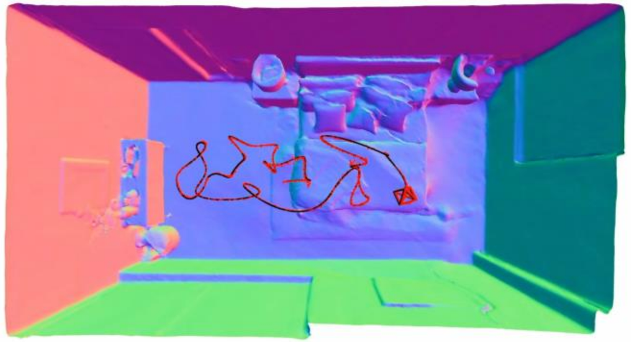
MonoSDF
NeurIPS 2022



The “Tri-plane”



NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Best Honor. Men.)

UNISURF
ICCV 2021 (Oral)

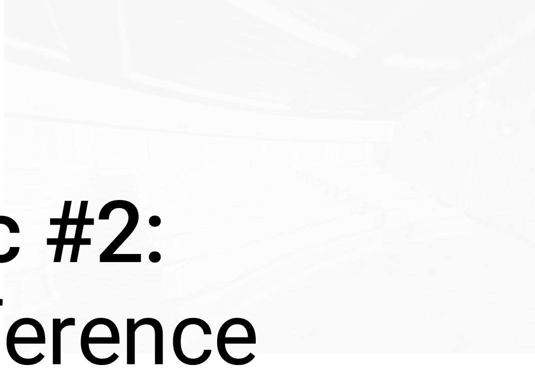
OpenScene
CVPR 2023

My PhD

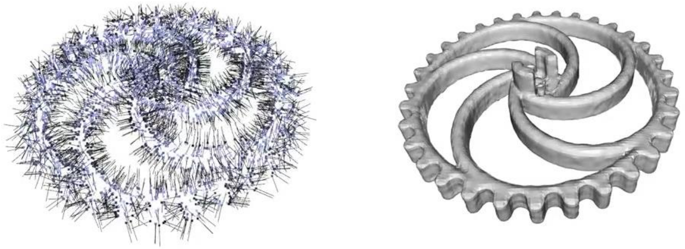
Learn to Reconstruct and Understand 3D World

Topic #2:
Fast Inference

ConvOccNet
ECCV 2020 (Spotlight)



MonoSDF
NeurIPS 2022



Shape As Points
NeurIPS 2021 (Oral)

runs now at 50 fps on a GTX 1080 Ti

KiloNeRF
ICCV 2021

NICE-SLAM
CVPR 2022

NICER-SLAM
3DV 2024 (Oral)

UNISURF
ICCV 2021 (Oral)

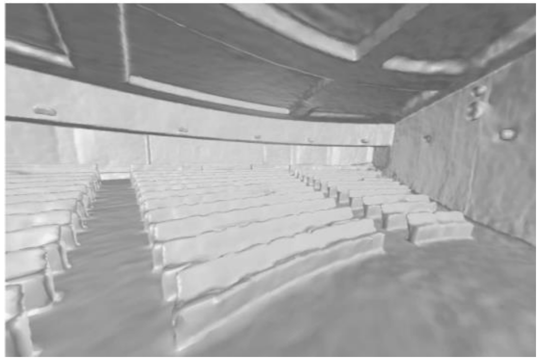
OpenScene
CVPR 2023

My PhD

Learn to Reconstruct and Understand 3D World

Topic #3:
Reconstruct from 2D Observations

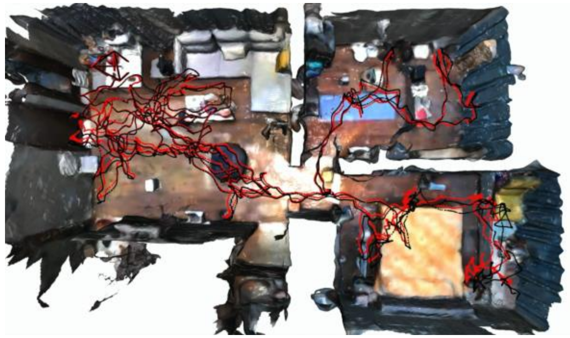
ConvOccNet
ECCV 2020 (Spotlight)



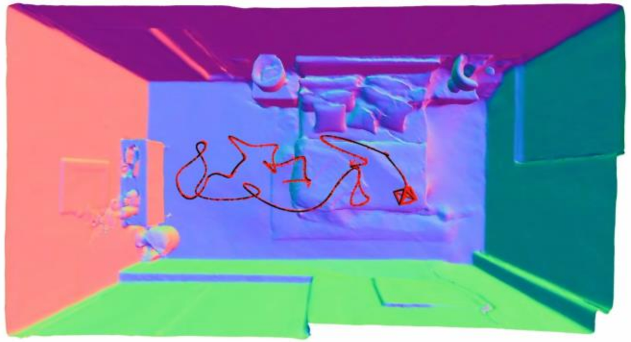
MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

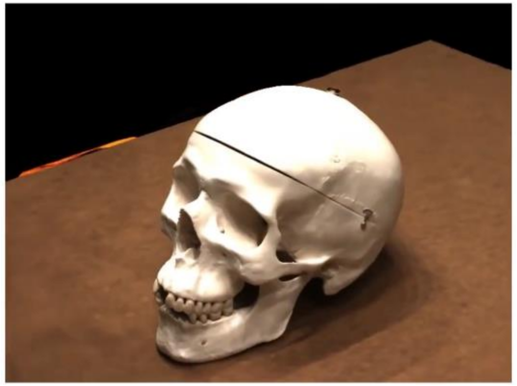
KiloNeRF
ICCV 2021



NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Best Paper Honorable)



UNISURF
ICCV 2021 (Oral)

OpenScene
CVPR 2023

My PhD

Learn to Reconstruct and Understand 3D World

ConvOccNet
ECCV 2020 (Spotlight)

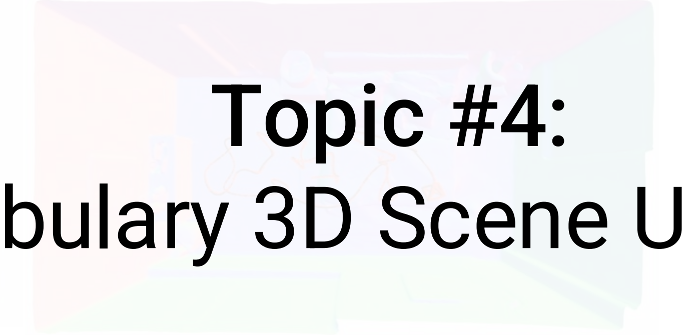
MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

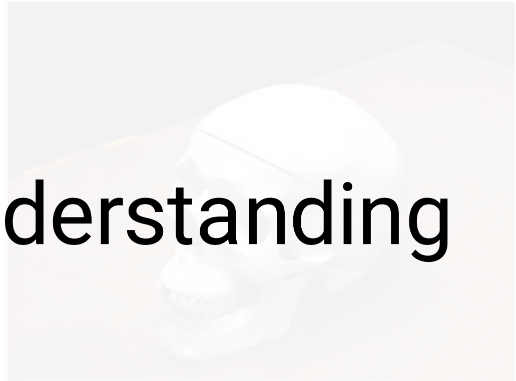
KiloNeRF
ICCV 2021

Topic #4:
Open-vocabulary 3D Scene Understanding

NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)

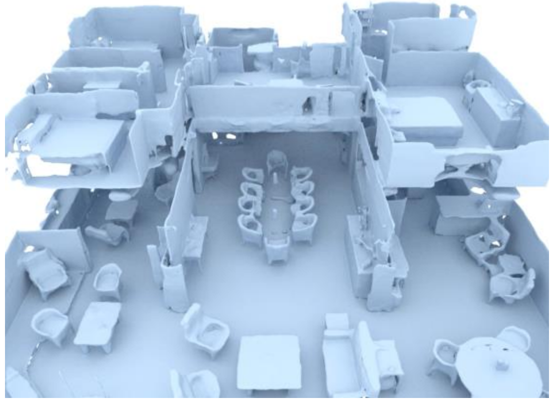


UNISURF
ICCV 2021 (Oral)

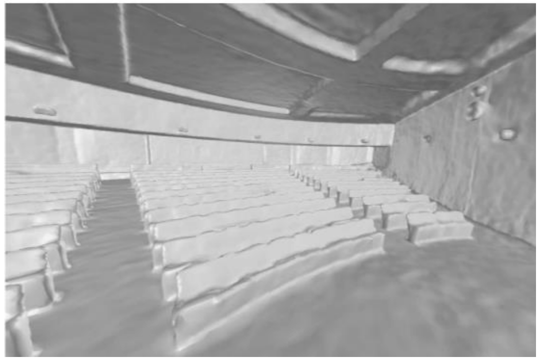
OpenScene
CVPR 2023

My PhD

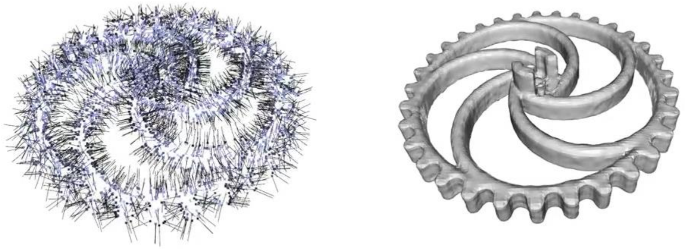
Learn to Reconstruct and Understand 3D World



ConvOccNet
ECCV 2020 (Spotlight)

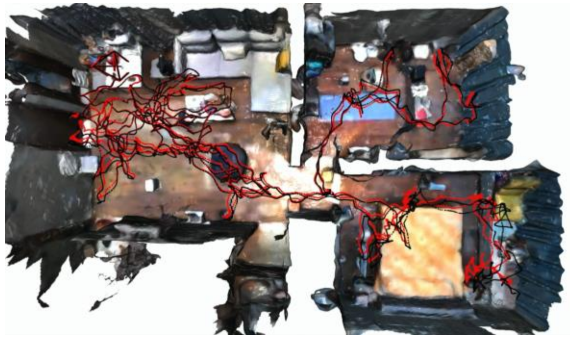


MonoSDF
NeurIPS 2022

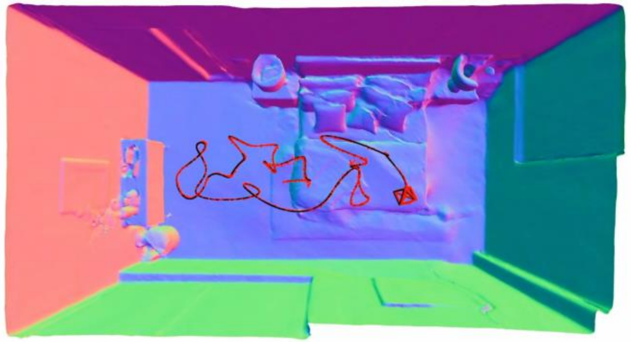


Shape As Points
NeurIPS 2021 (Oral)

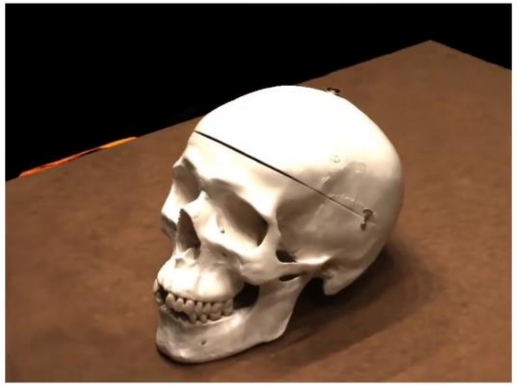
KiloNeRF
ICCV 2021



NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Best Paper Honorable)



UNISURF
ICCV 2021 (Oral)

OpenScene
CVPR 2023

Building Visual Intelligence

Grounding

Reconstruct and understand 3D

Reasoning

Solve complicated tasks

Scaling

Foundation Model for Generalization

Action

Agent and tool use

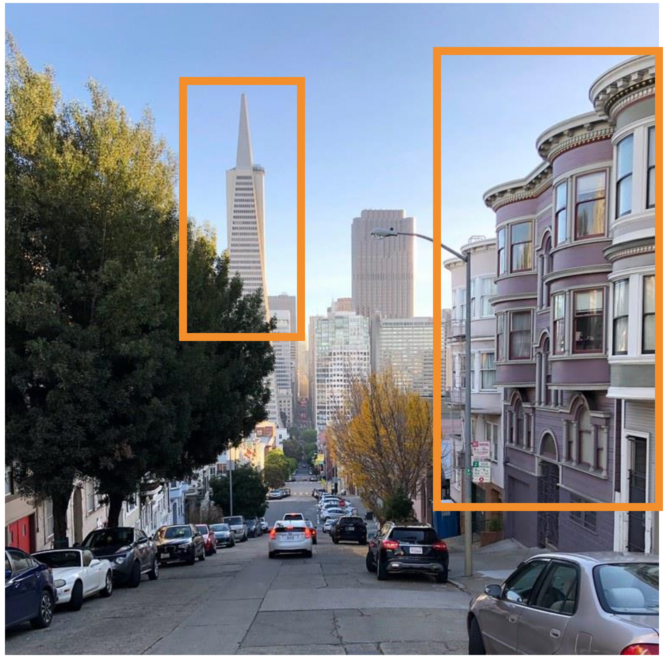
Current Focus at GDM

Teaching Multimodal LLMs to Think in Space

Pre-training for ✦ Gemini

X billion tokens for **spatial grounding**, **multi-view consistency**, **high-level semantics**, etc

Post-training for ✦ Gemini



The model can **think with images**, and actively conduct information seeking

Building Visual Intelligence

Grounding

Reconstruct and understand 3D

Reasoning

Solve complicated tasks

Scaling

Foundation Model for Generalization

Action

Agent and tool use

Foundation Model for Visual Intelligence

From 2 Views to 10 Million

NoPoSplat

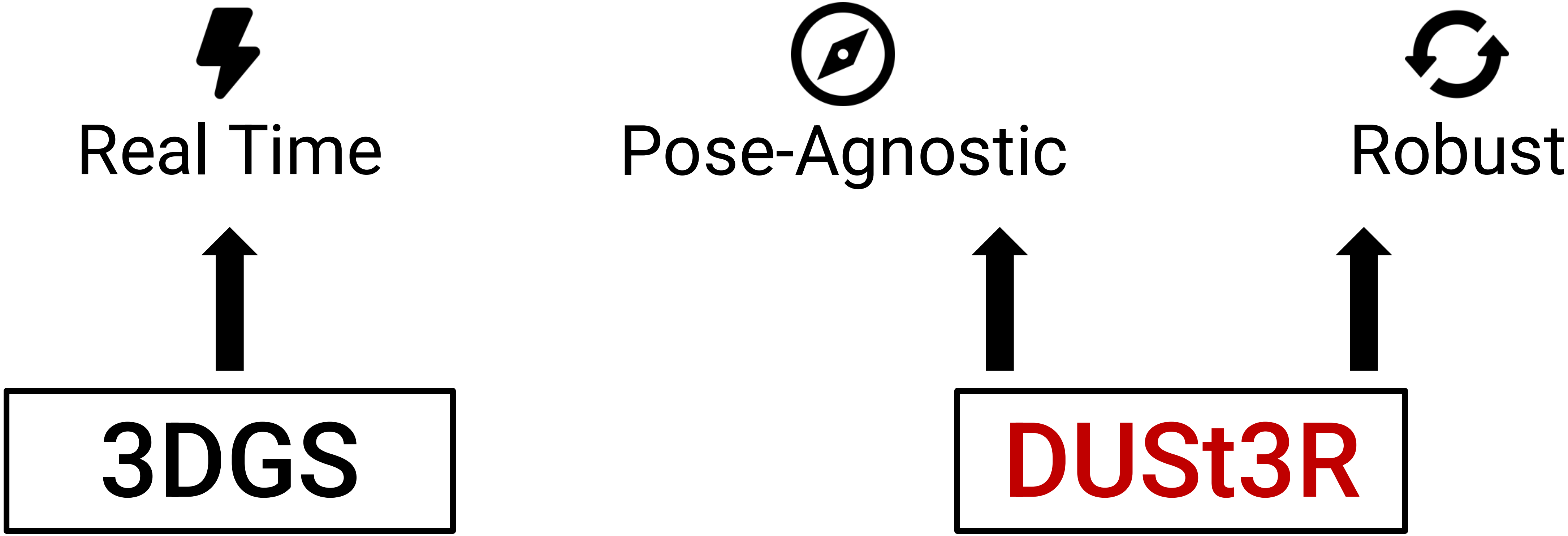
ICLR 2025 (**Oral**)

Visual Chronicles

ICCV 2025 (**Highlight**)

An Ideal 3D Modelling Pipeline

Instant, Pose-Free, Real-World 3D Everywhere



Real Time

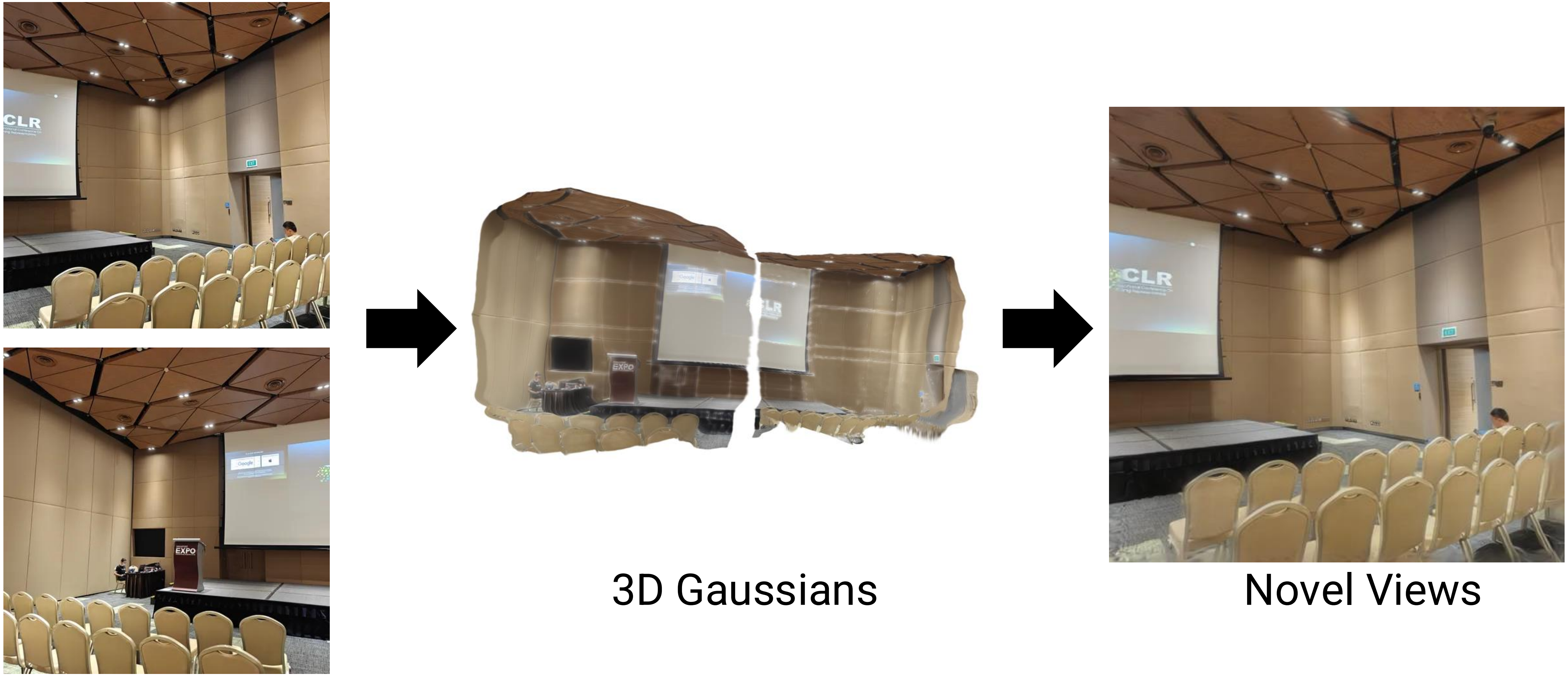
Pose-Agnostic

Robust

3DGS

DUSt3R

Goal: Unposed Feedforward 3DGS



3D Gaussians

Novel Views

Input Images **w/o poses**

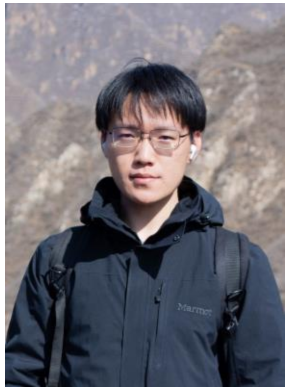
ETH zürich ETH AI CENTER ICLR NVIDIA

No Pose, No Problem 👌

Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images

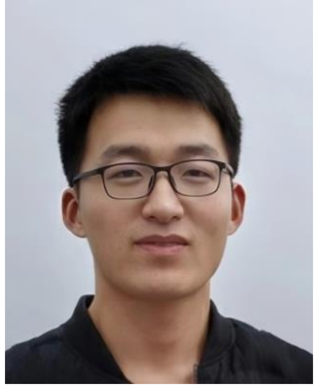
(a.k.a **NoPoSplat**)

ICLR 2025 (Oral, top 1.8%)

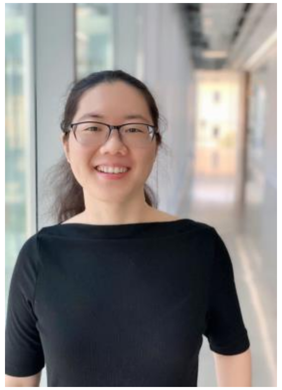


Botao Ye

Sifei Liu



Haofei Xu



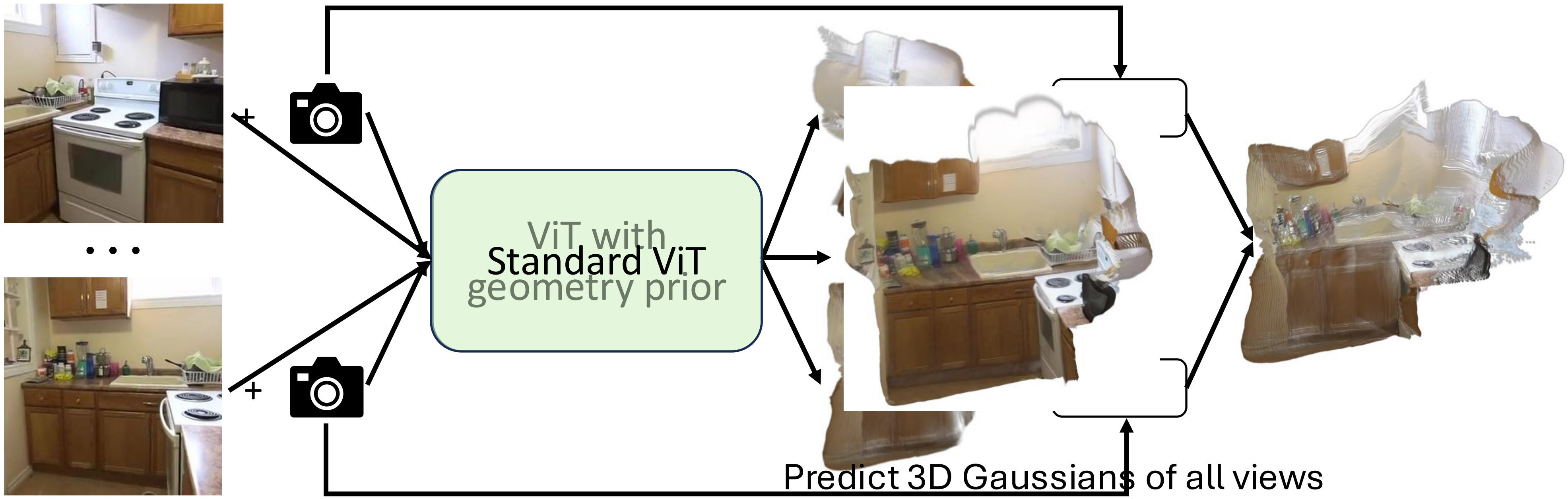
Xueting Li

Marc Pollefeys

Ming-Hsuan Yang

Songyou Peng

Previous Feed-forward 3DGS



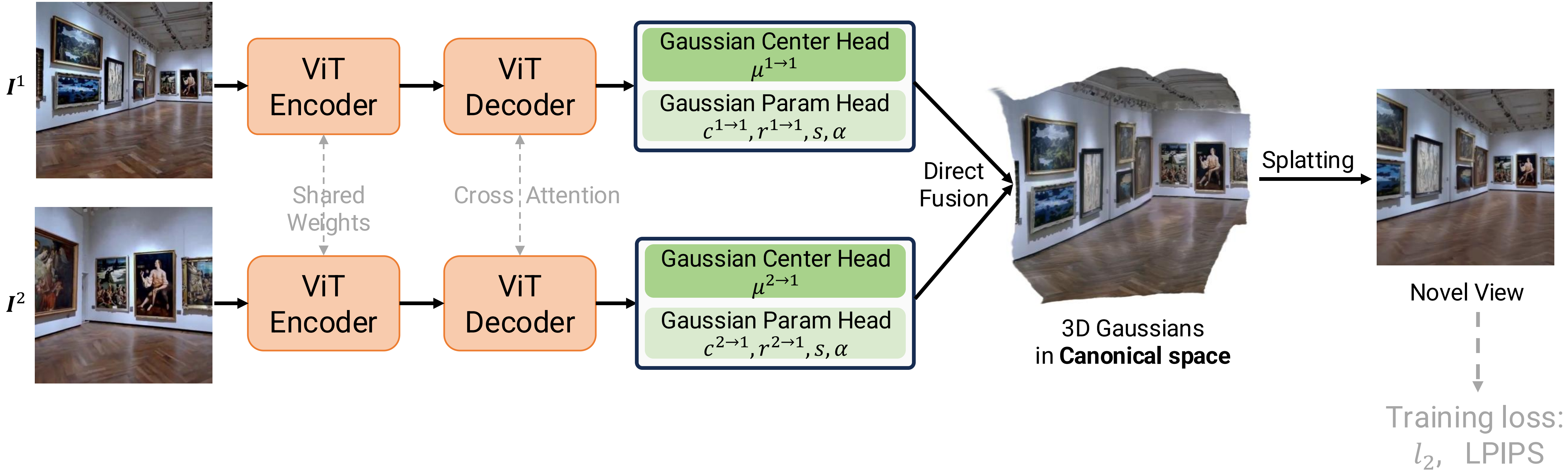
ViT with geometry prior

Standard ViT

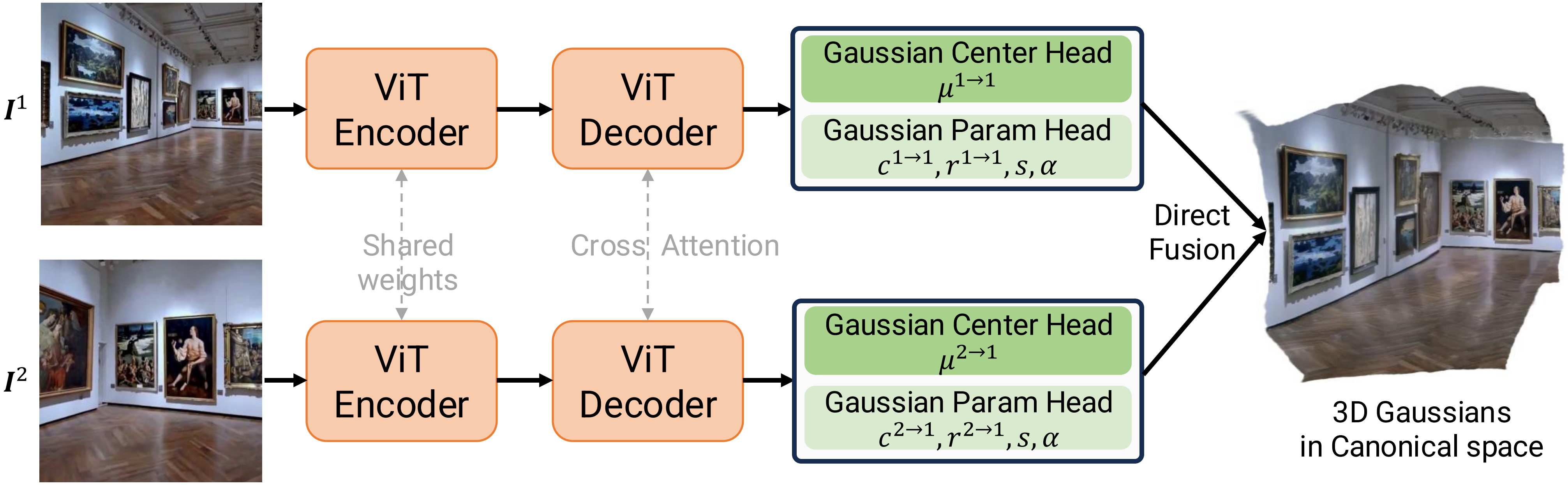
Predict 3D Gaussians of all views in a **canonical space**

Charatan, Li, Tagliasacchi, Sitzmann: pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. CVPR 2024

Architecture

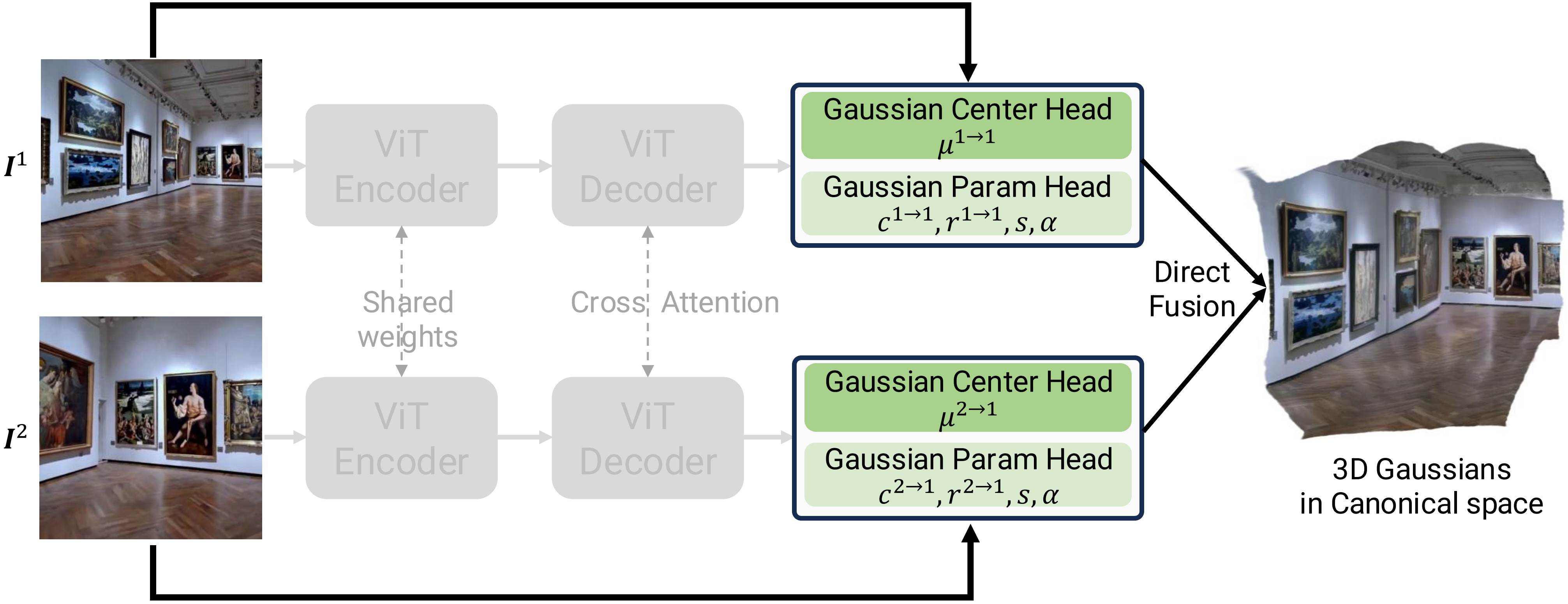


Issue 1: Blurry Rendering



Issue 1: Blurry Rendering

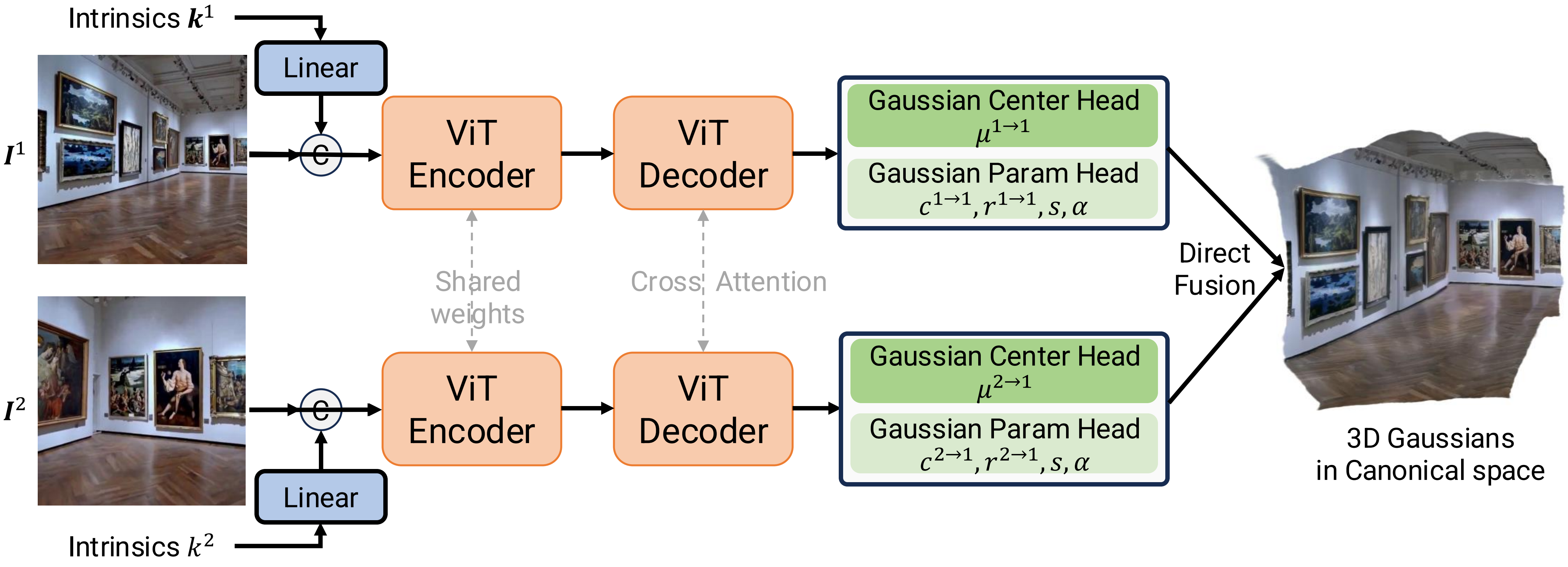
Solution: Add a shortcut!



Issue 2: Scale Ambiguity

Solution: Add the intrinsic embeddings!

$$\boxed{p = K(RP + t)}$$



Issue 3: Inaccurate Pose Estimation

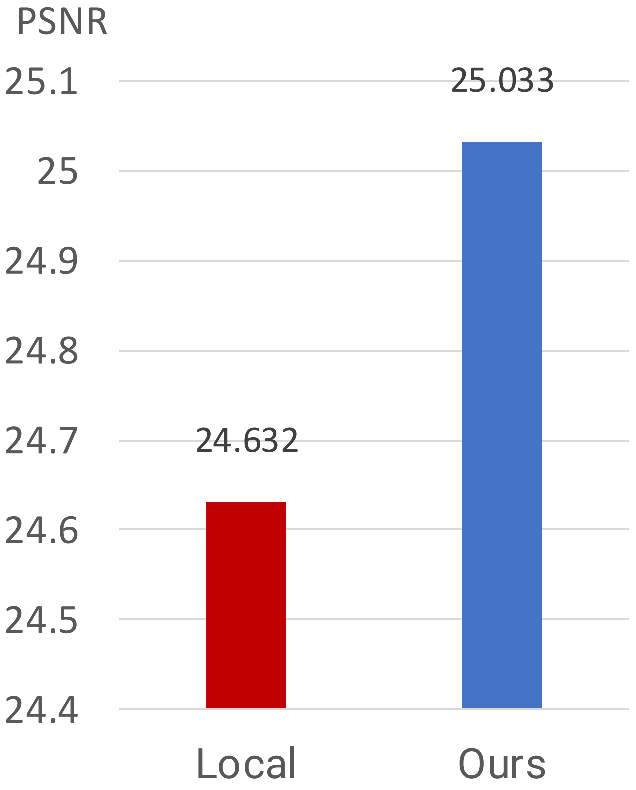
Solution: coarse-to-fine estimation

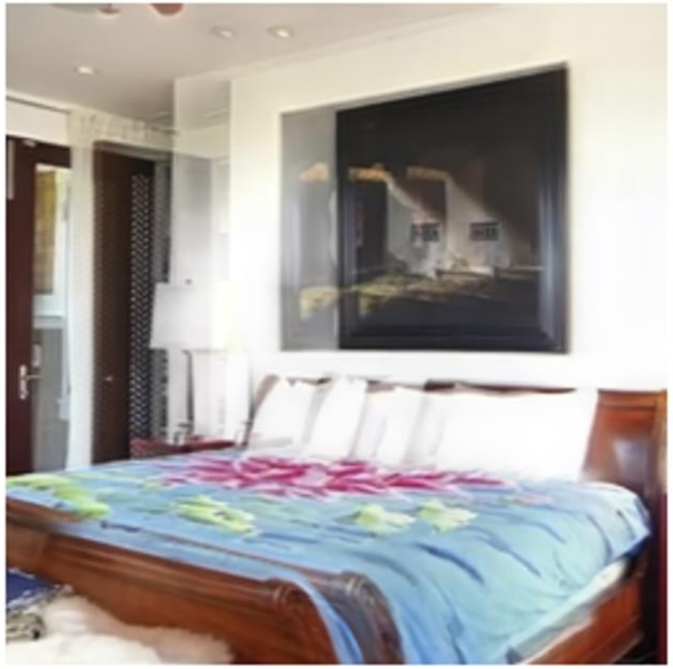
- Coarse stage: run RANSAC-PnP on Gaussian centers
- Refine stage: optimize with photometric loss

PnP	Photometric	5°	10°	20°
✓	✓	**0.318**	**0.538**	**0.717**
✓		0.287	0.506	0.692
	✓	0.017	0.027	0.051

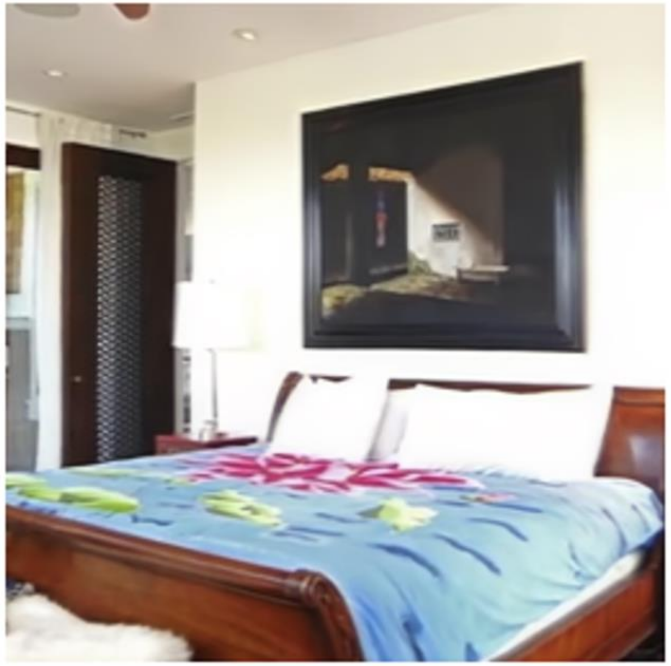
Ablation

Canonical Gaussian prediction





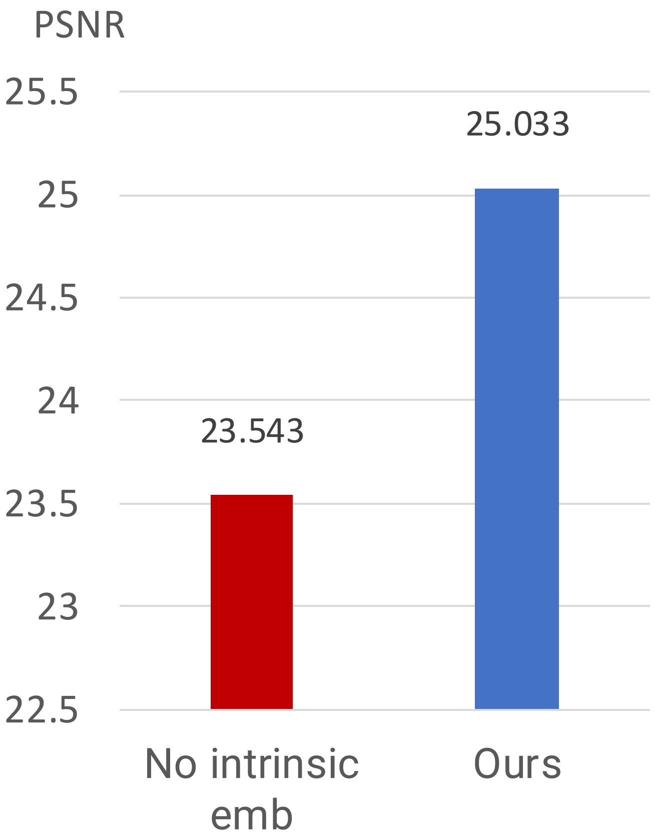
Local

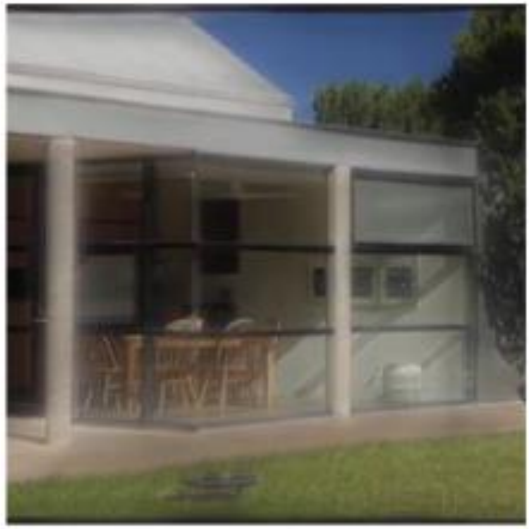


Canonical

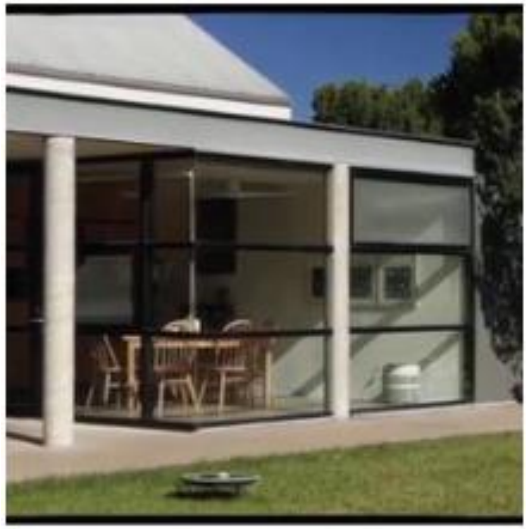
Ablation

Intrinsic embedding





No Intrinsic Emb



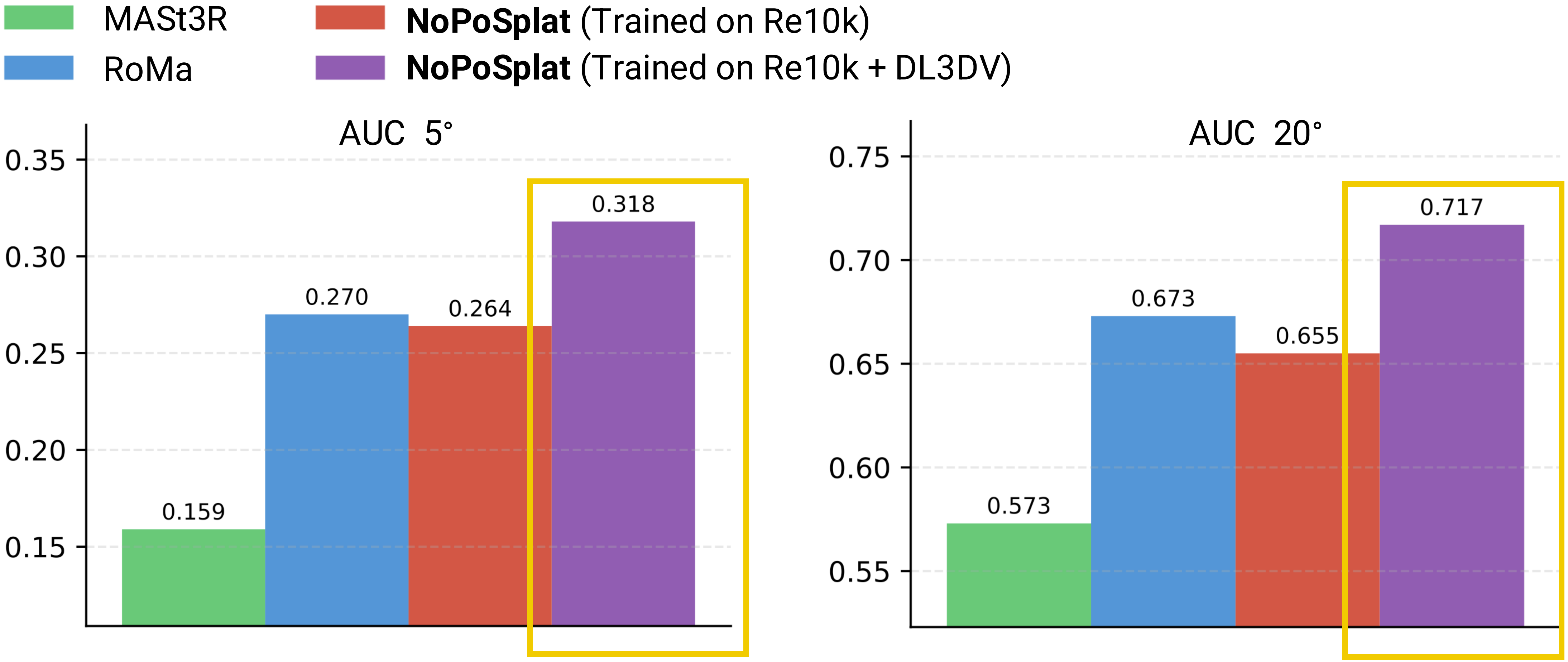
Ours

GT

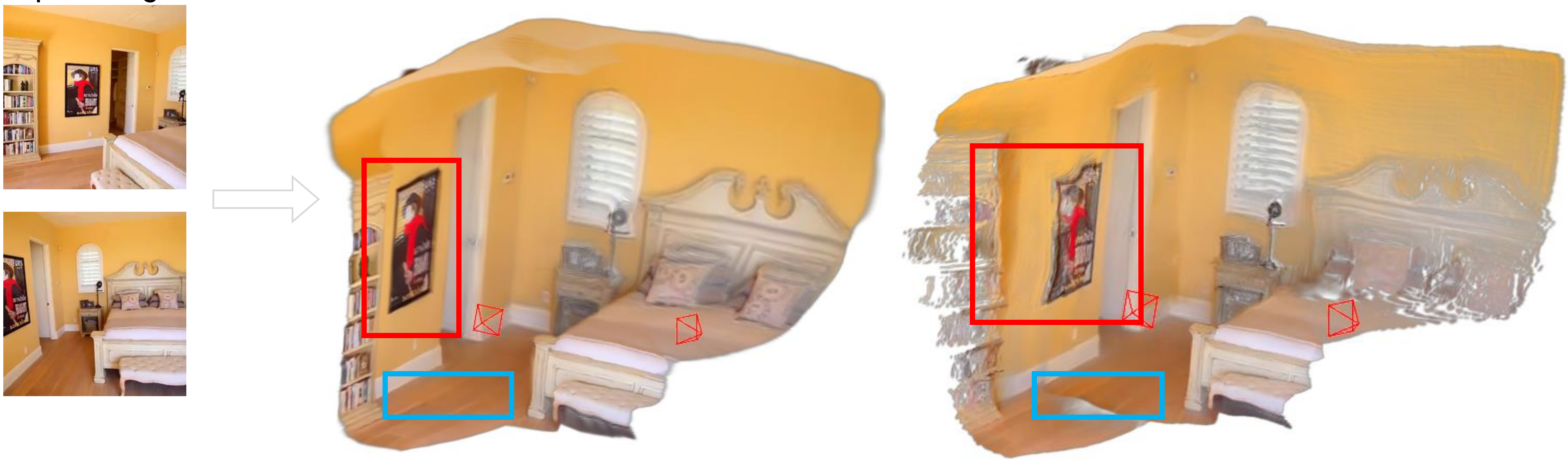
What is More…

Accurate Pose Estimation

Evaluation on ScanNet



High Quality Geometry



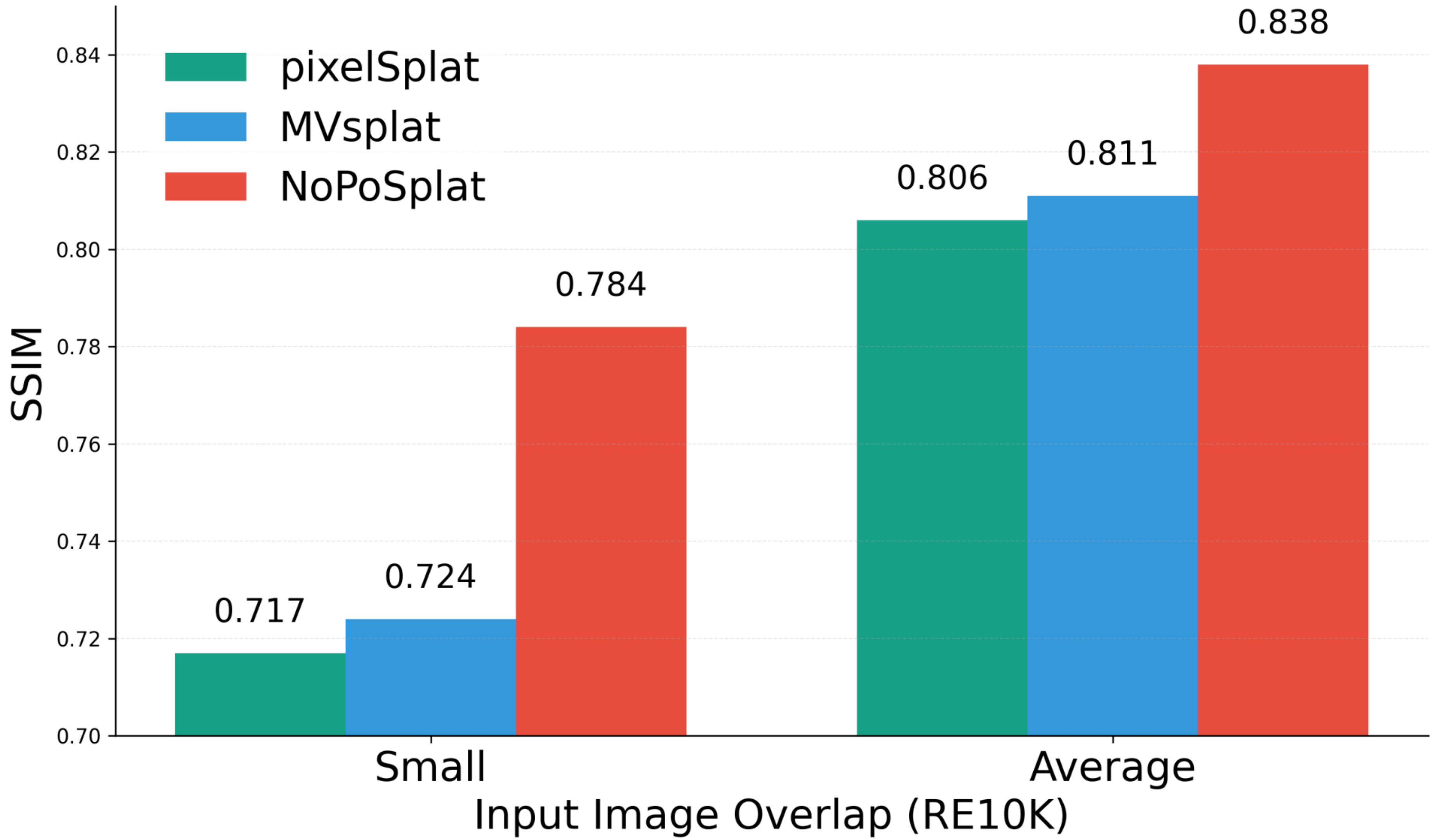
Input Images

NoPoSplat (pose-free)

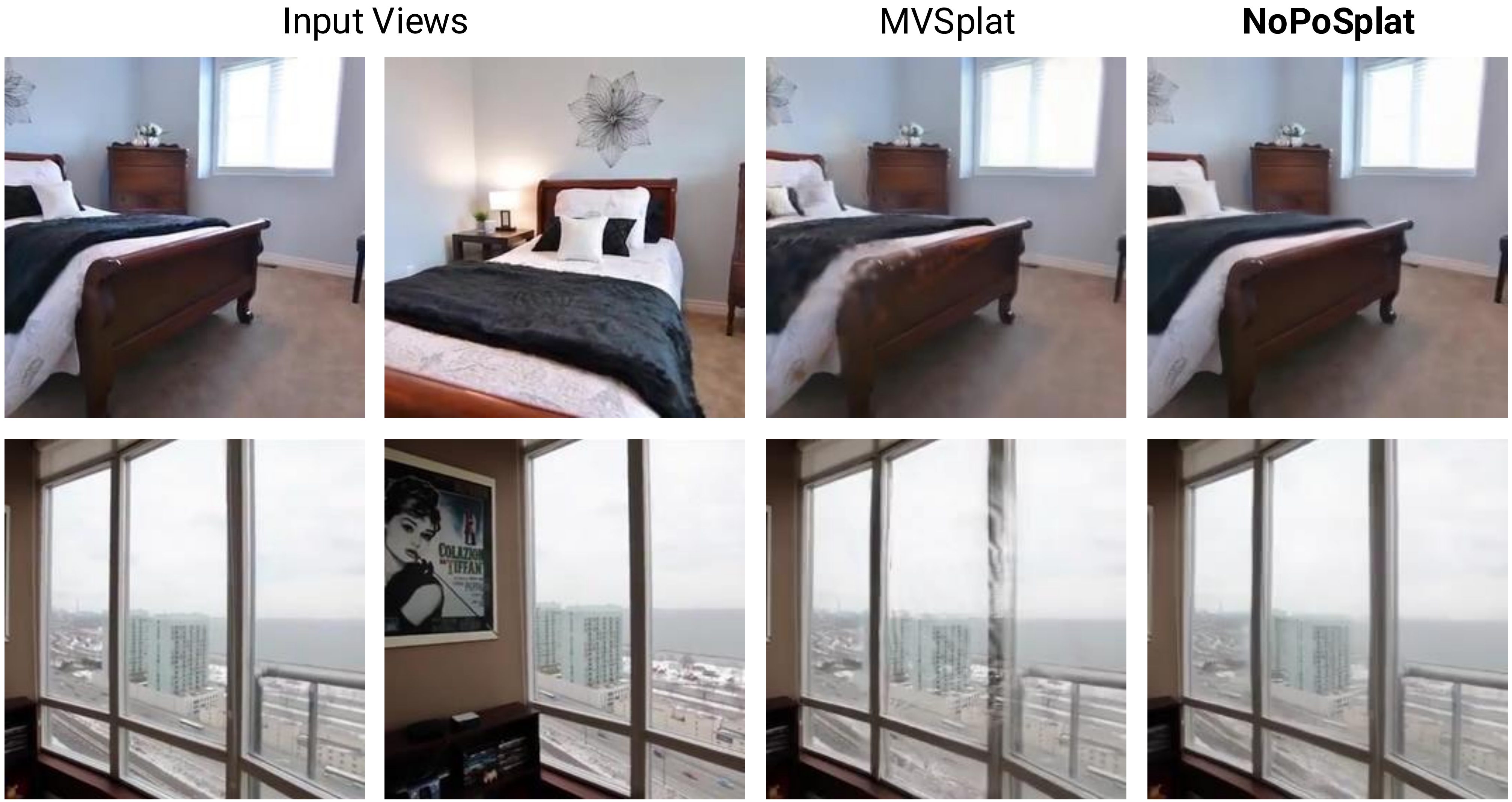
MVSplat (pose-required)

Appearance Quality

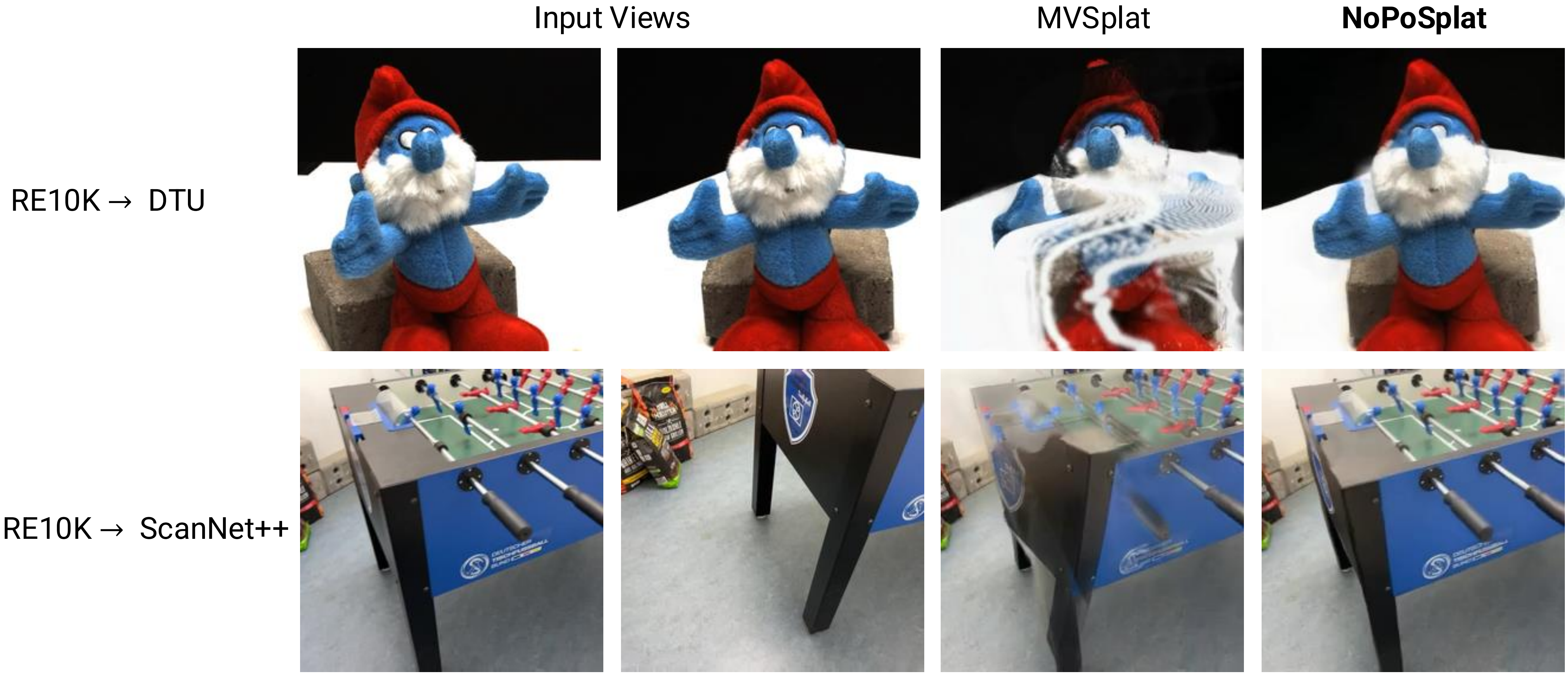
Better even than pose-required methods!



Appearance Quality



Cross-Dataset Generalization



Input Views

MVSplat

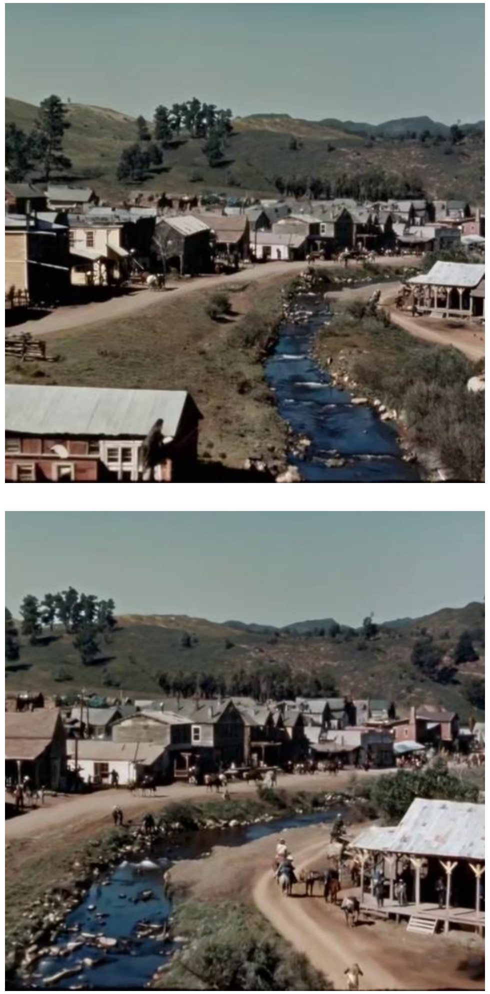
NoPoSplat

RE10K → DTU

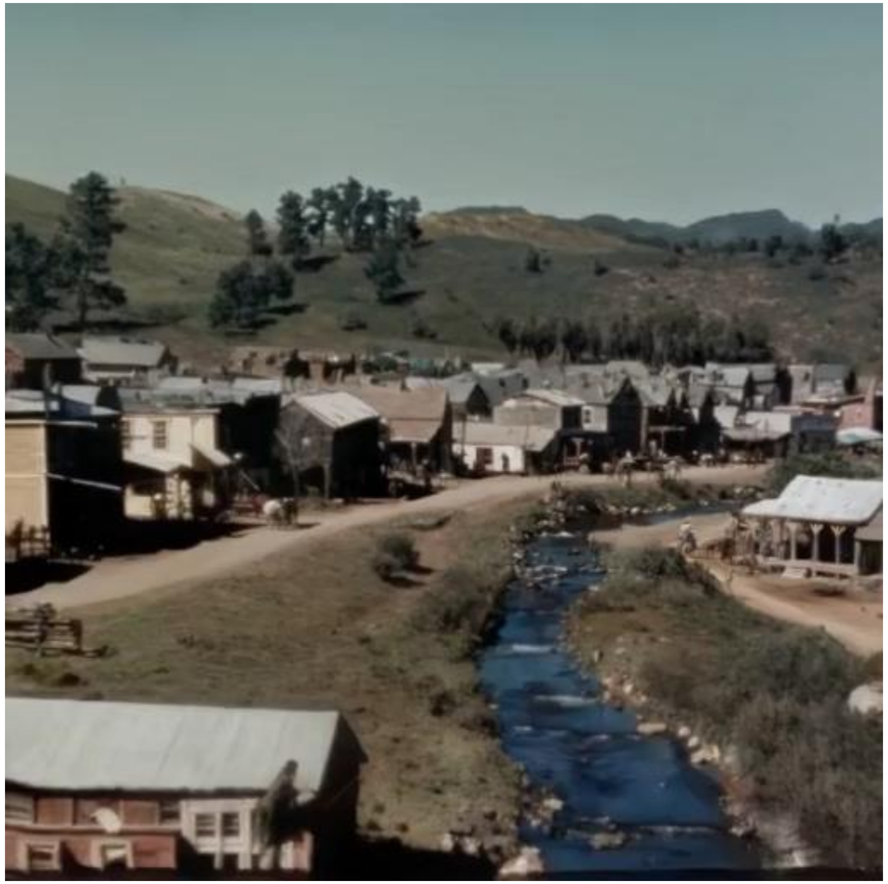
RE10K → ScanNet++

In-the-Wild Data

Images extracted from OpenAI Sora



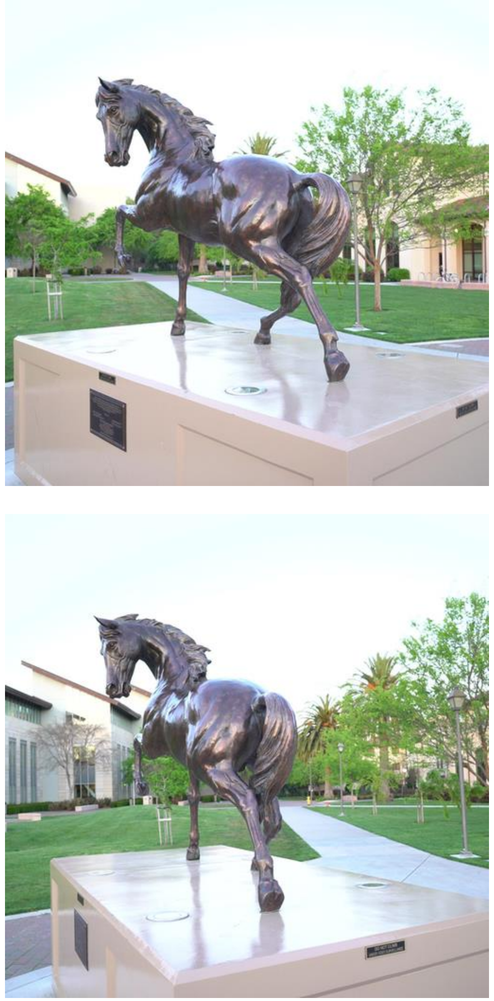
Input Images



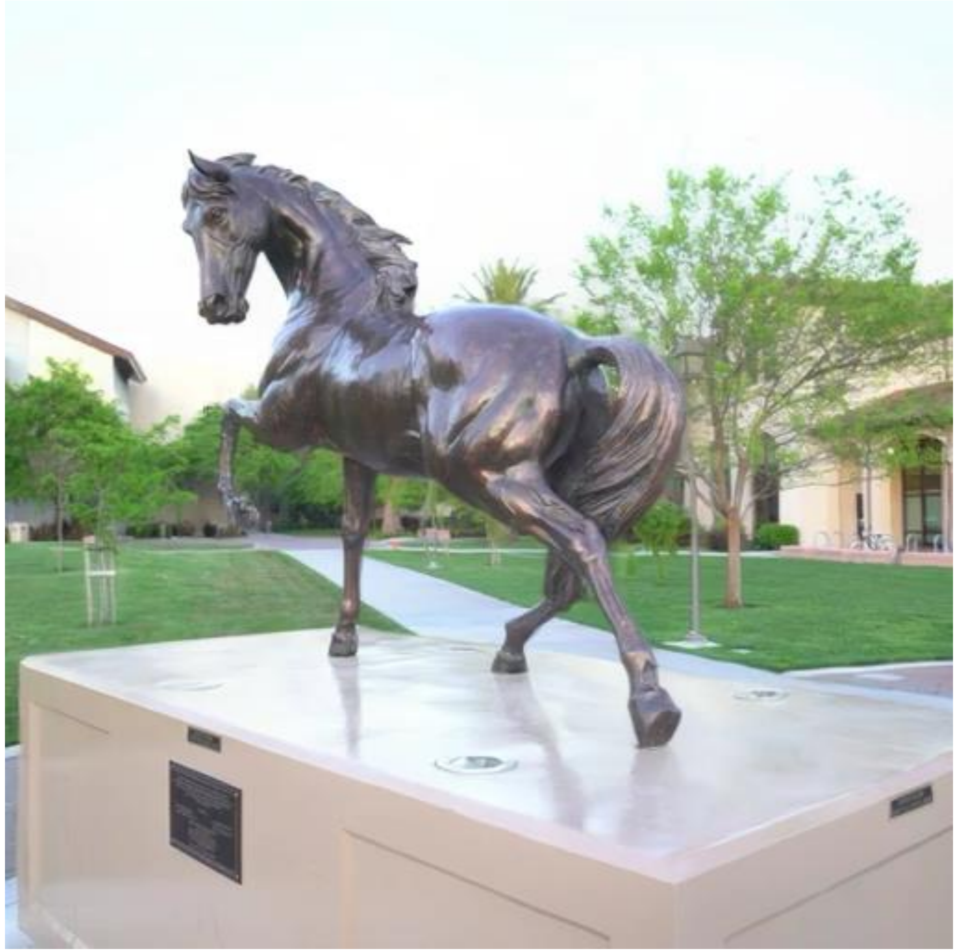
Novel Views

In-the-Wild Data

Images from Tanks & Temples



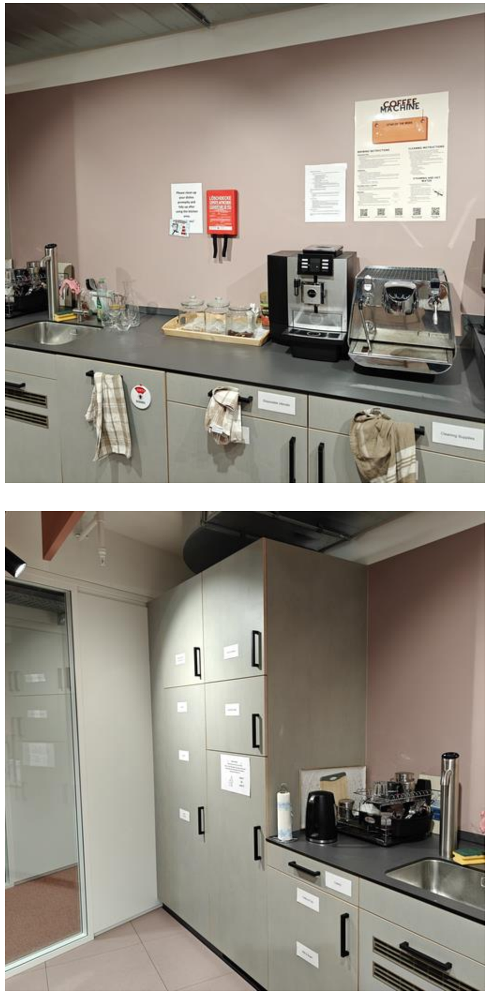
Input Images



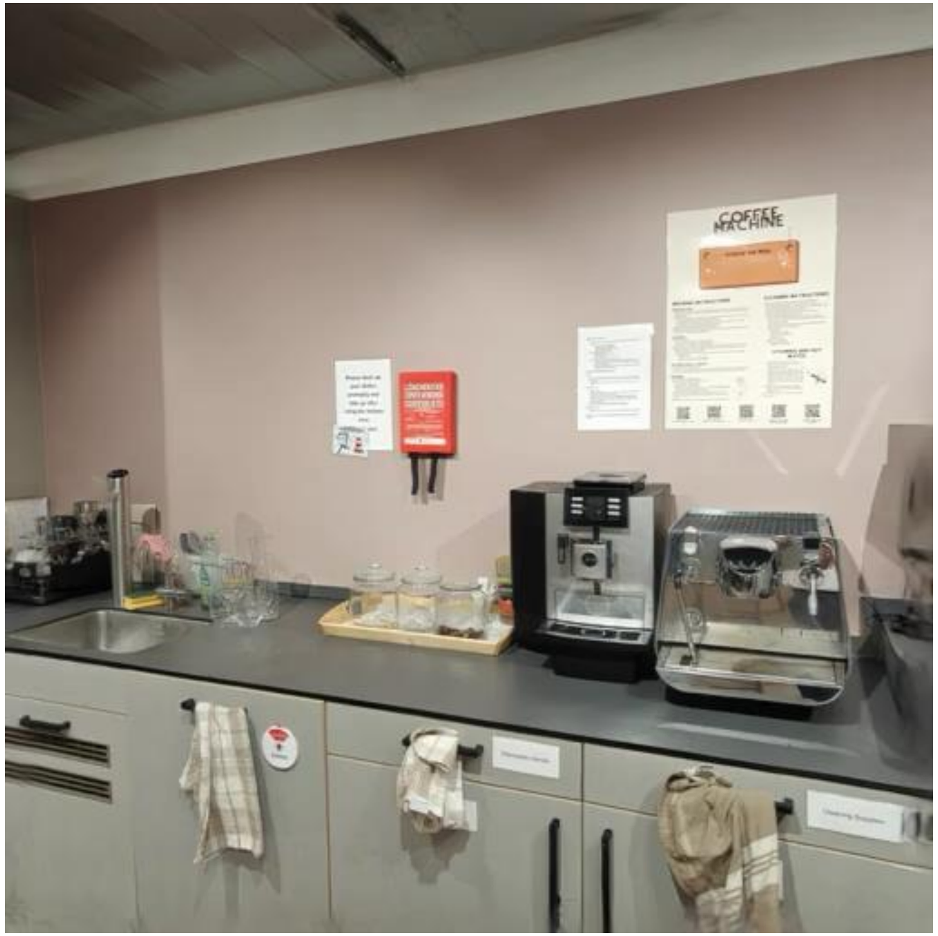
Novel Views

In-the-Wild Data

iPhone images



Input Images



Novel Views

Take-home Messages

- Feedforward NVS can be surprisingly simple!

- Side product: SoTA relative pose estimation

- Foundation model rocks!

Foundation Model for Visual Intelligence

From 2 Views to 10 Million

NoPoSplat

ICLR 2025 (**Oral**)

Visual Chronicles

ICCV 2025 (**Highlight**)

Google DeepMind

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Stanford

Visual Chronicles

Using Multimodal LLMs to Analyze Massive Collections of Images

ICCV 2025 (Highlight)

Boyang Deng

Songyou Peng

Kyle Genova

Gordon Wetzstein

Noah Snavely

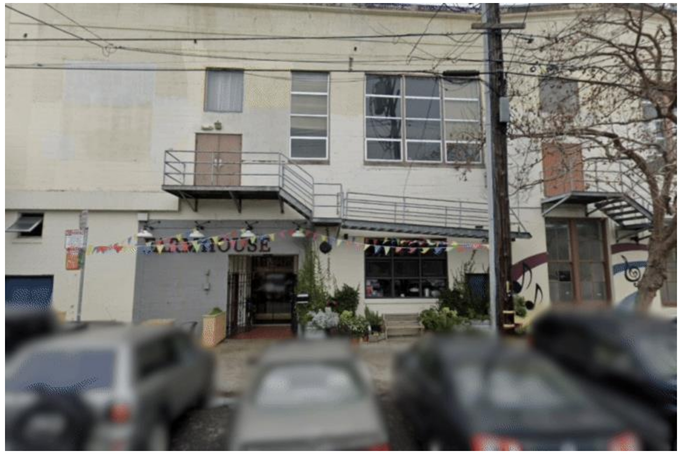
Leo Guibas

Tom Funkhouser

Motivation

What are the **interesting changes** happened in the time-lapses?

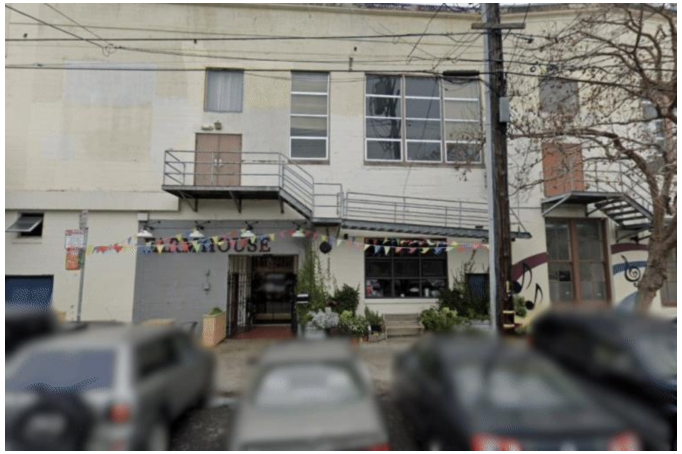
The bridge was painted in a bright blue color.



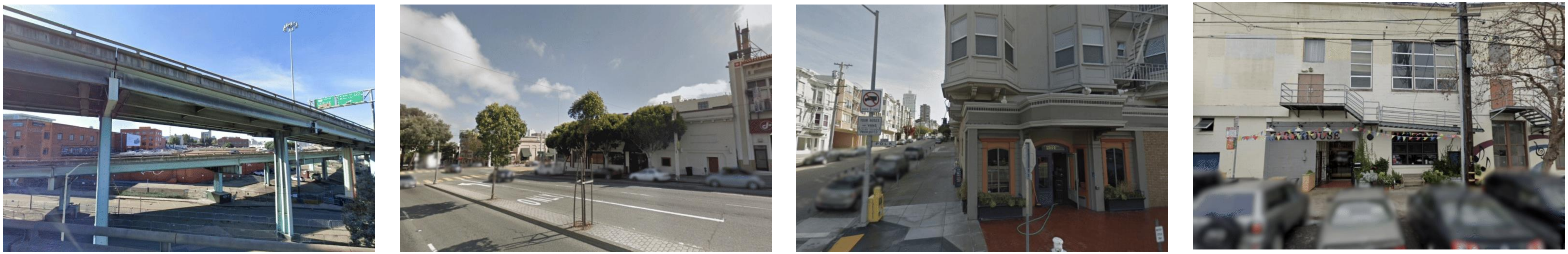
The restaurant **extended a dining structure outside**.

Motivation

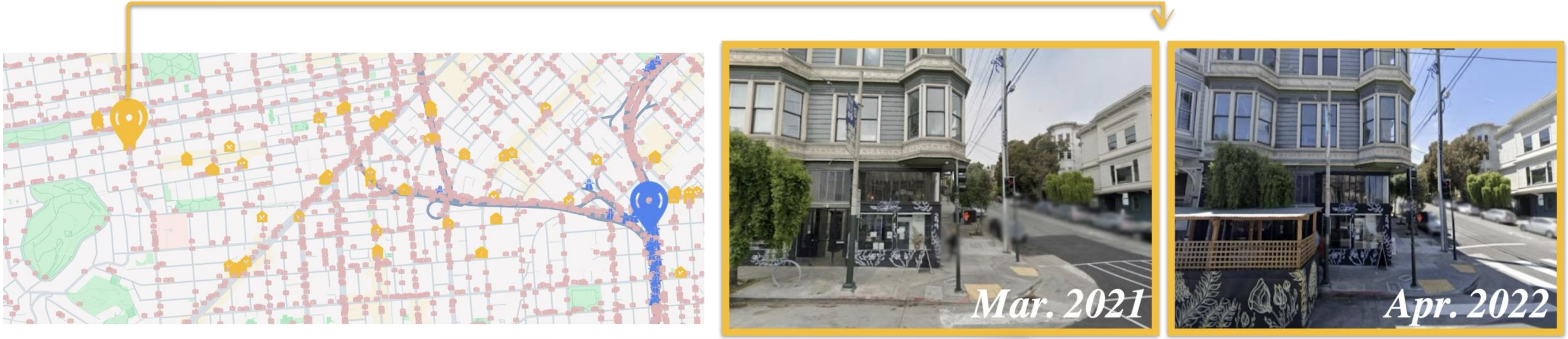
What are the **interesting changes** happened in the time-lapses?



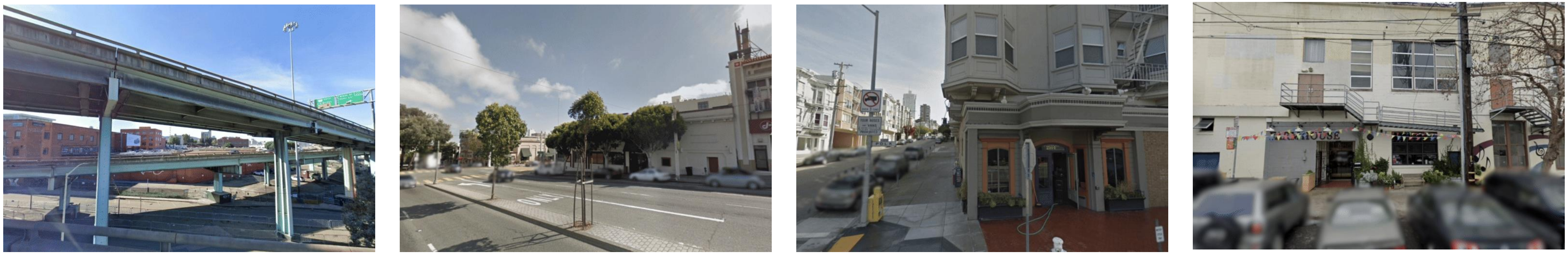
- **Open-ended queries**
- Not too challenging for humans
- What if we have **millions** of time-lapses?
- What if we want to know **trends of changes**?
- Quite challenging for any CV models!
  - No “interesting change” detectors.
  - No ImageNet of interesting changes.



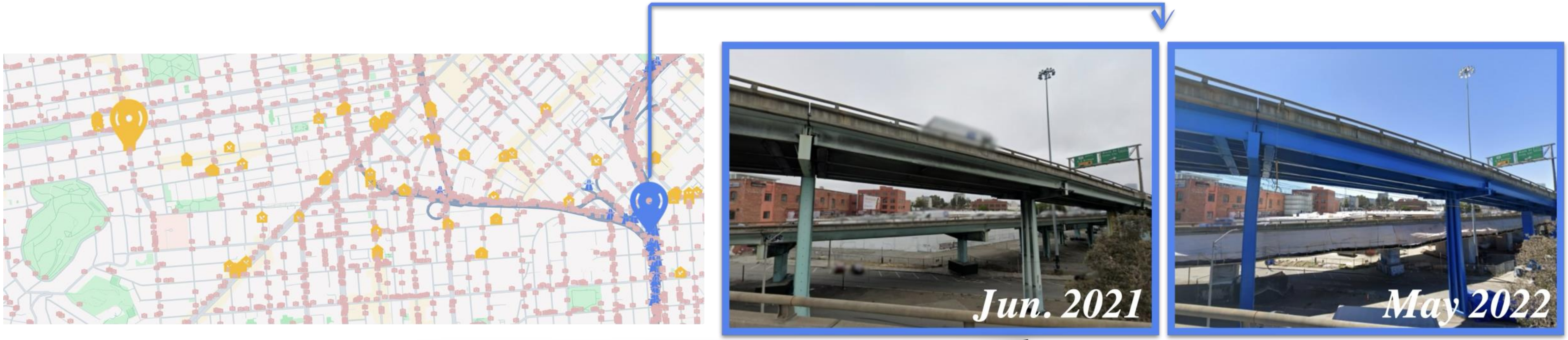
Massive Collections of Images 📷 (**20M** per City)



…added ***outdoor dining*****.** (seen 1482 times in 🏠)



Massive Collections of Images 📷 (**20M** per City)



…overpass painted ***blue****. (seen 481 times in* 🛣 *)*

How to Approach Trend Discovery?

MLLMs as an essential tool

Brute Force #1: Directly ask LLMs w/o any data?

- Abstract answers, e.g. “Increased focus on sustainability”.
- No evidence — Hard to verify any trends.

Brute Force #2: Feed all images to MLLMs and ask?

- Gemini could take up to 8K images at a time
- Boring output: Half of the output is about addition / removal of scaffolding

Visual Chronicles

Step 1: Use MLLMs for Local Change Detection

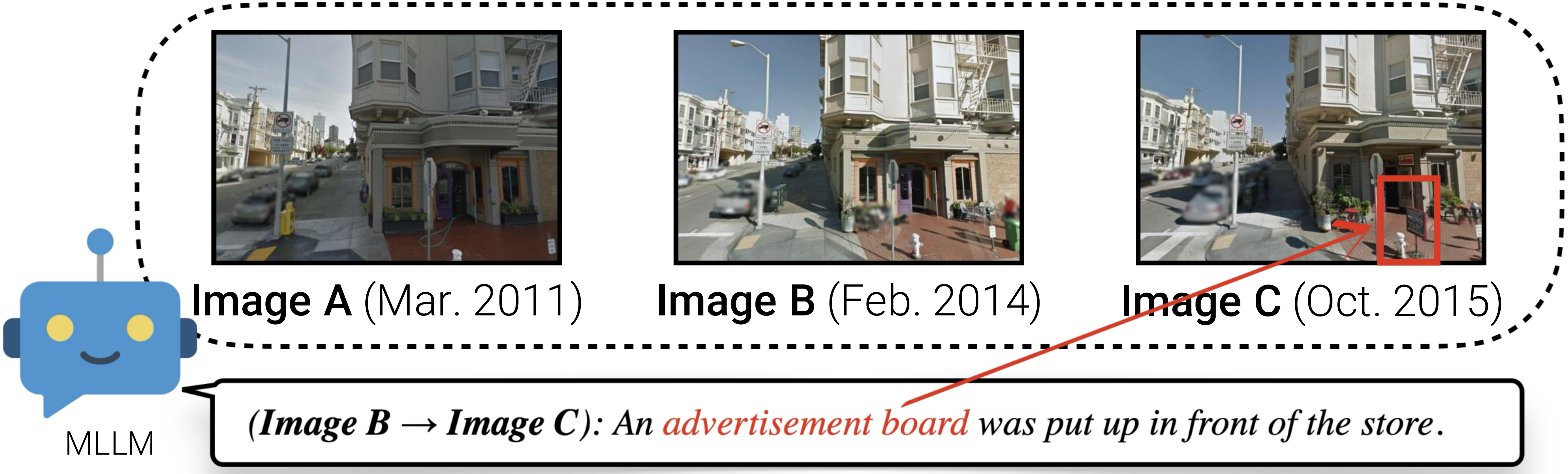
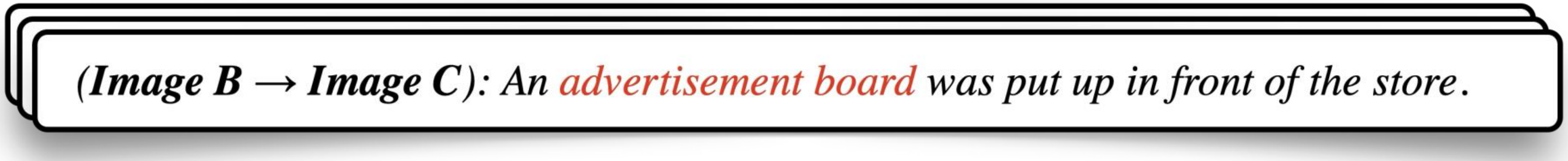


Image A (Mar. 2011) **Image B** (Feb. 2014) **Image C** (Oct. 2015)

***(Image B → Image C)**: An advertisement board was put up in front of the store.*

Visual Chronicles

Step 2: find trends among local changes (**3M** per city)



*(**Image B** → **Image C**): An advertisement board was put up in front of the store.*

Brute Force: Feed all changes to LLMs?

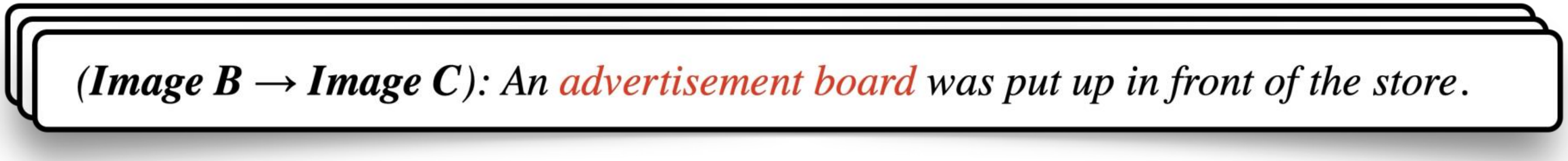
- Very limited input and output

Ours: Two-step hybrid approach

1. Produce **visual trend proposals**
2. Verify which proposed trends are supported by N changes

Visual Chronicles – Trend Discovery

How to produce visual trend proposals?



*(**Image B** → **Image C**): An advertisement board was put up in front of the store.*

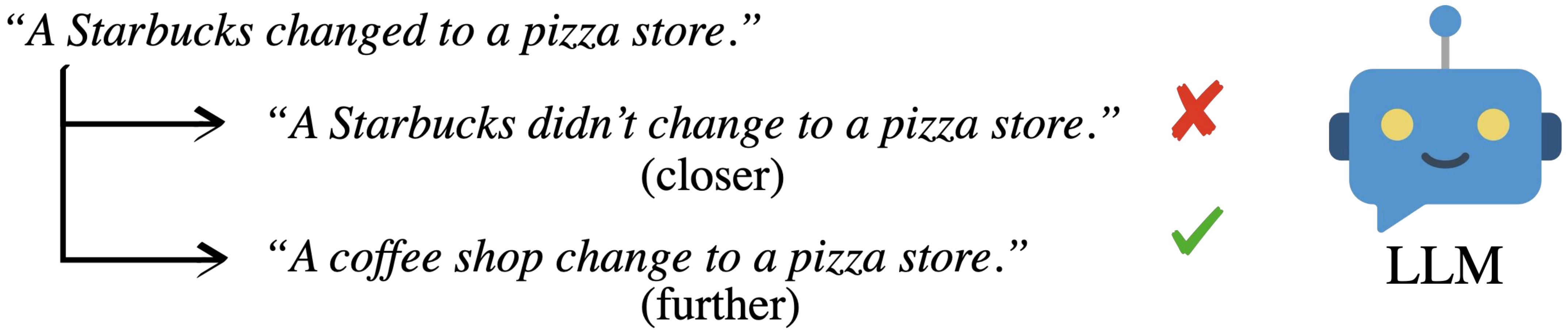
1. Encode local changes to text embedding
2. Sort them based on the lengths
3. NMS with to find the top 500 trend proposals

Visual Chronicles – Trend Discovery

How to verify which proposal are supported?

Use <u>distance in the text embedding</u> space with a tighter threshold

- **It cannot capture subtle similarities**!



“A Starbucks changed to a pizza store.”

“A Starbucks didn’t change to a pizza store.” ✗
(closer)

“A coffee shop change to a pizza store.” ✓
(further)

LLM

Ours: Pick top 1,500 changes for each proposal, **use LLMs to verify**

Visual Chronicles

First use of MLLMs for massive scale analysis of images

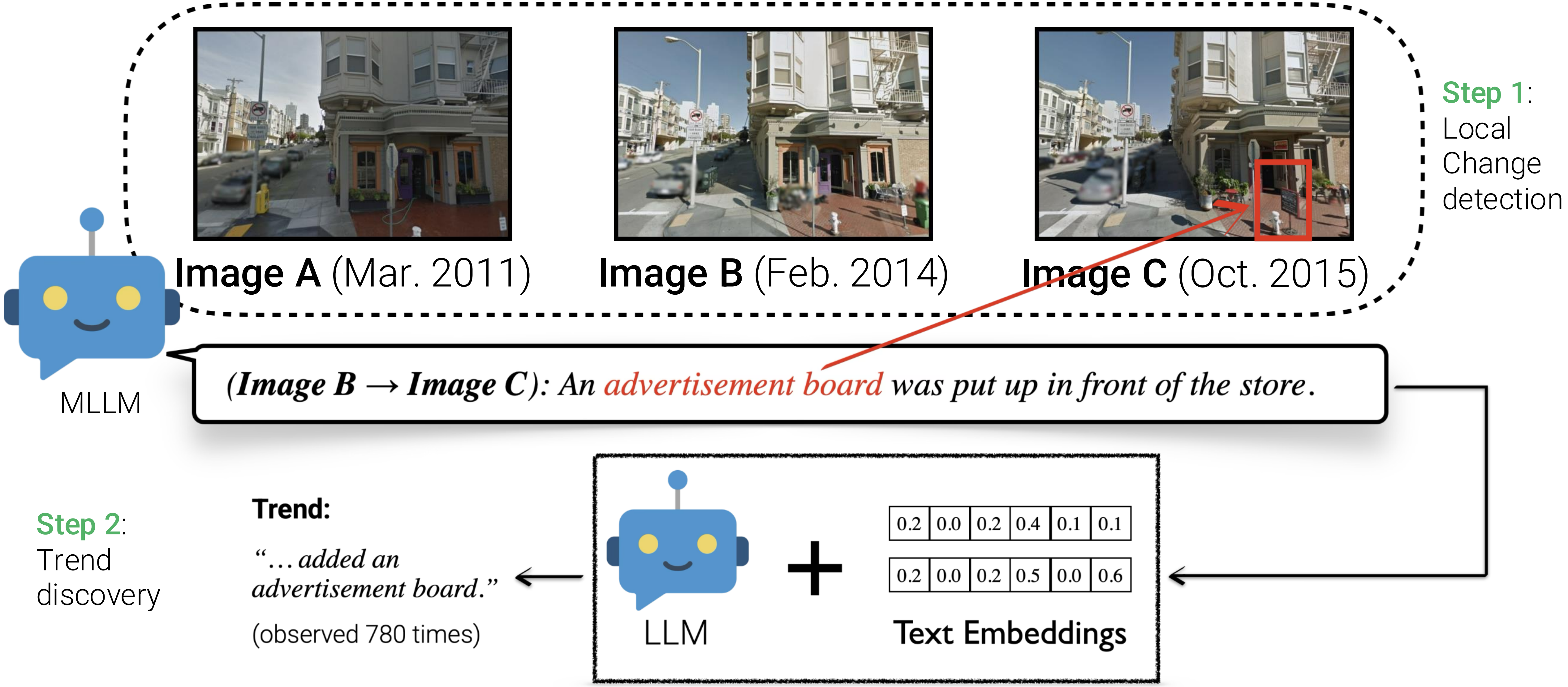


Image A (Mar. 2011)

Image B (Feb. 2014)

Image C (Oct. 2015)

Step 1: Local Change detection

MLLM

(Image B → Image C): *An advertisement board was put up in front of the store.*

Step 2: Trend discovery

Trend:

"… added an advertisement board."

(observed 780 times)

LLM

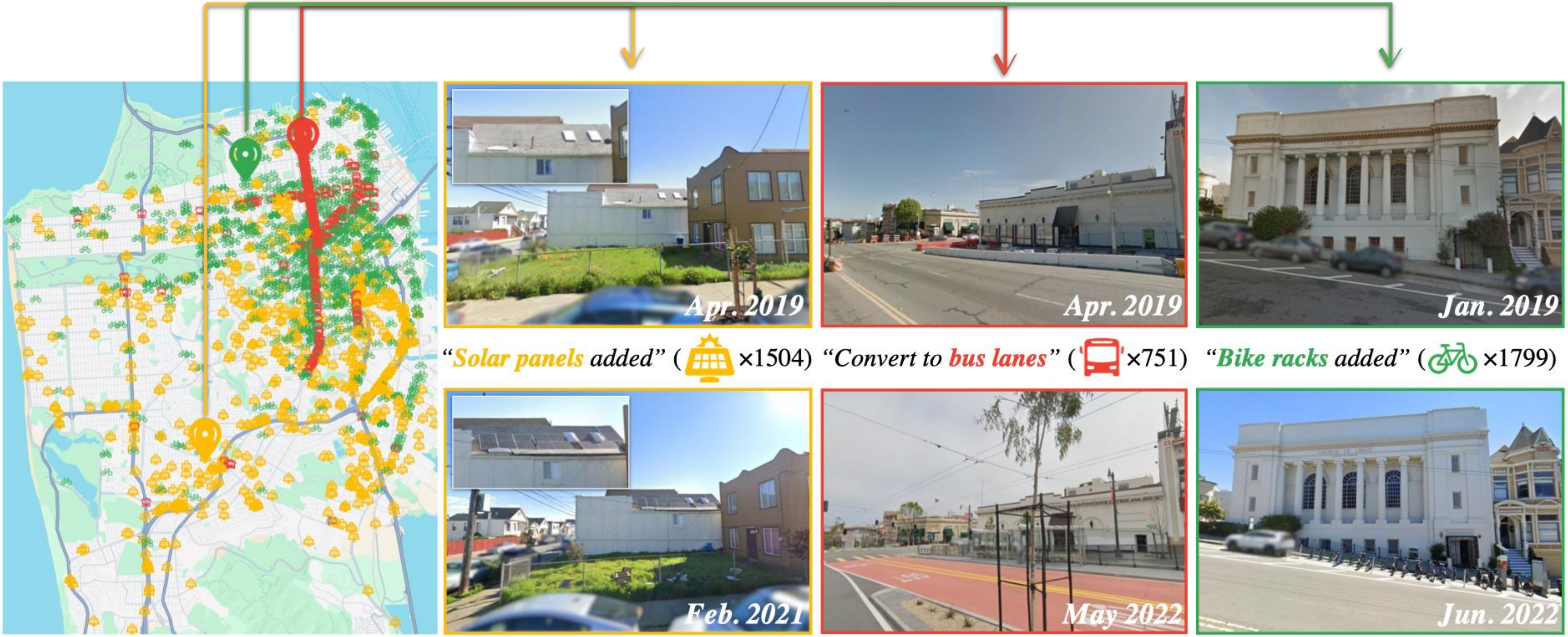
+

0.2	0.0	0.2	0.4	0.1	0.1
0.2	0.0	0.2	0.5	0.0	0.6

Text Embeddings

Results

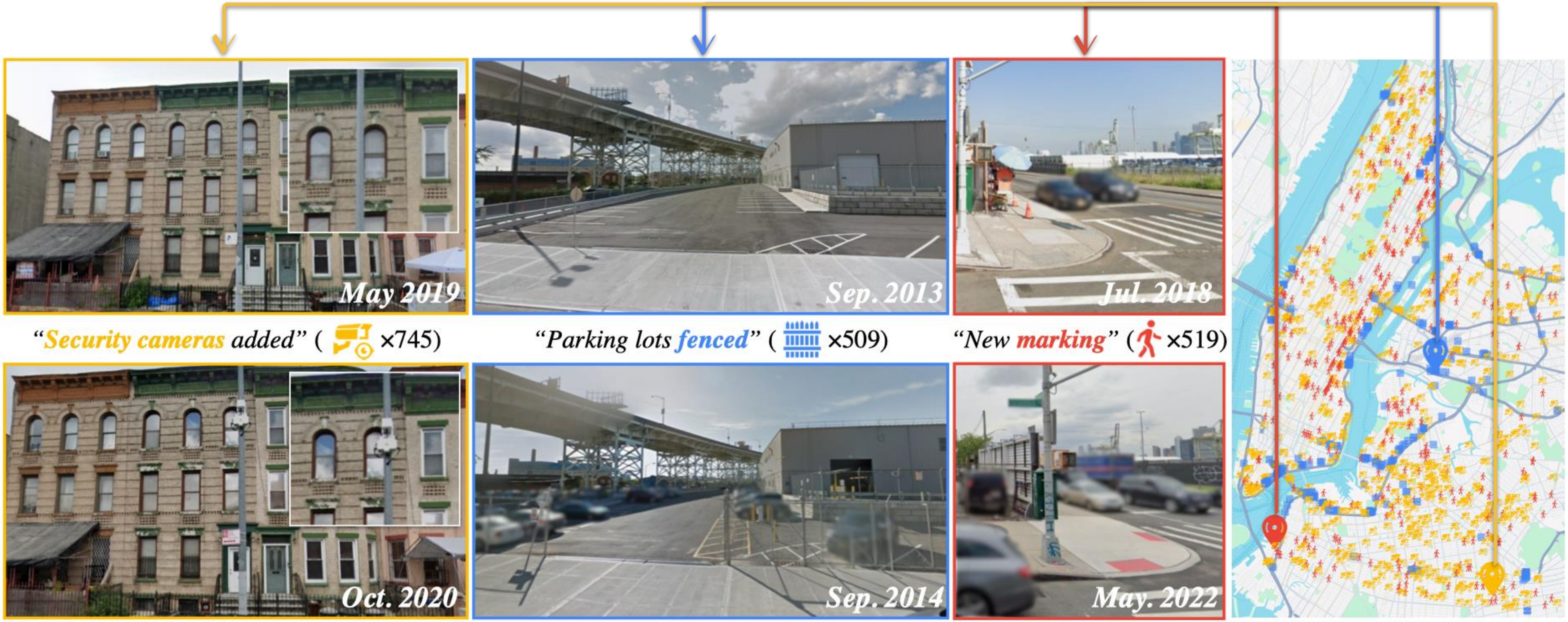
Discover fascinating trends in **San Francisco**



"*Solar panels added*" (☀ ×1504) "*Convert to bus lanes*" (🚌 ×751) "*Bike racks added*" (🚲 ×1799)

Results

Discover fascinating trends in **New York**



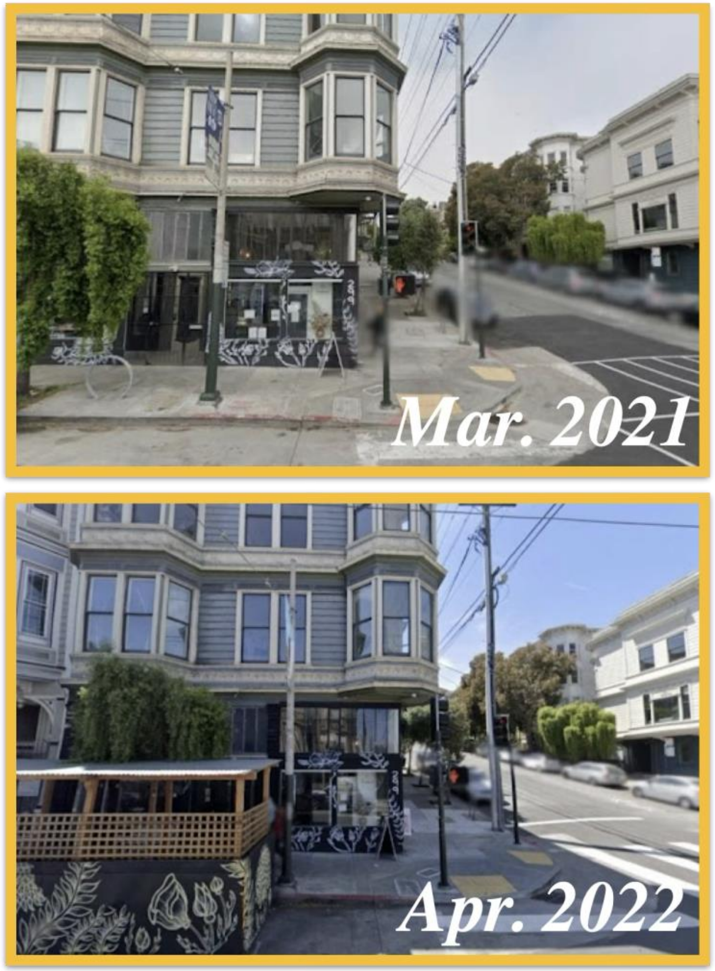
“Security cameras added” (×745)

“Parking lots fenced” (×509)

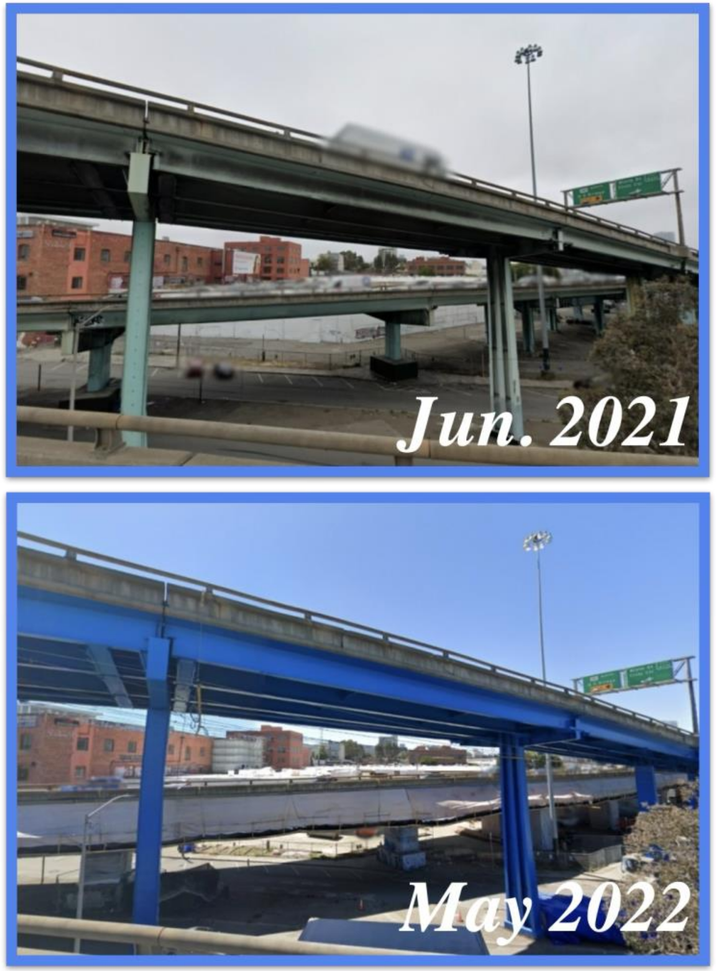
“New marking” (×519)

Results

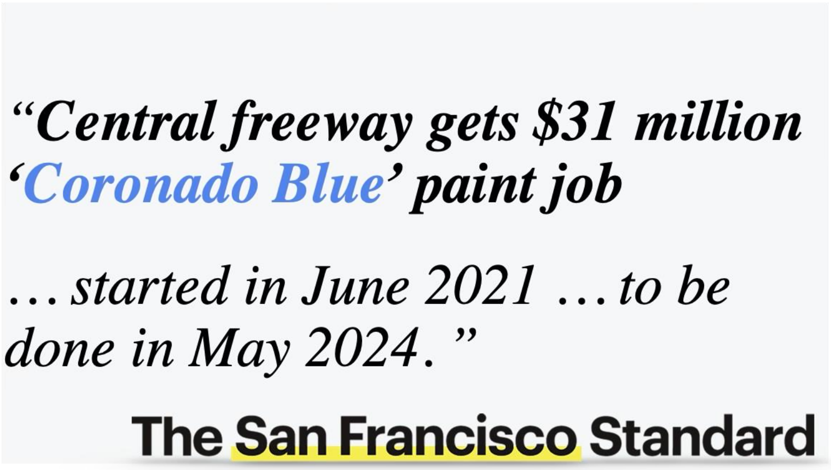
Support **temporally conditioned** search, e.g. “since 2020”



Outdoor Dining
(seen 1482 times)



Blue Overpass
(seen 481 times)



“Central freeway gets $31 million ‘Coronado Blue’ paint job

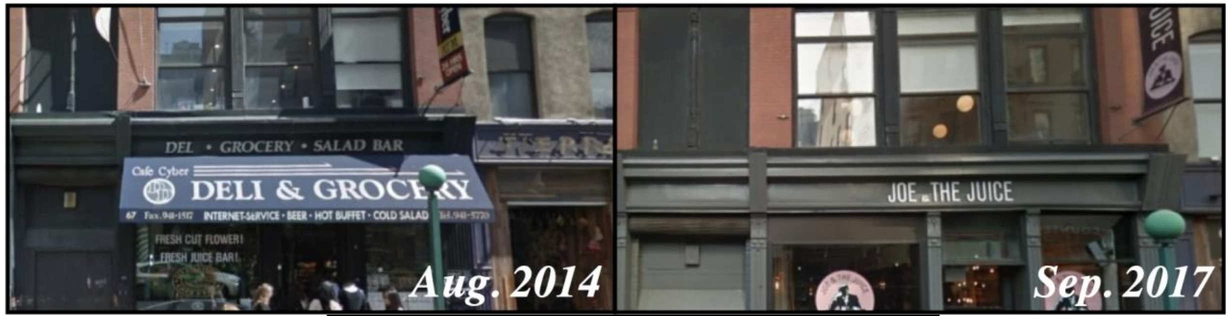
… started in June 2021 … to be done in May 2024.”

The San Francisco Standard

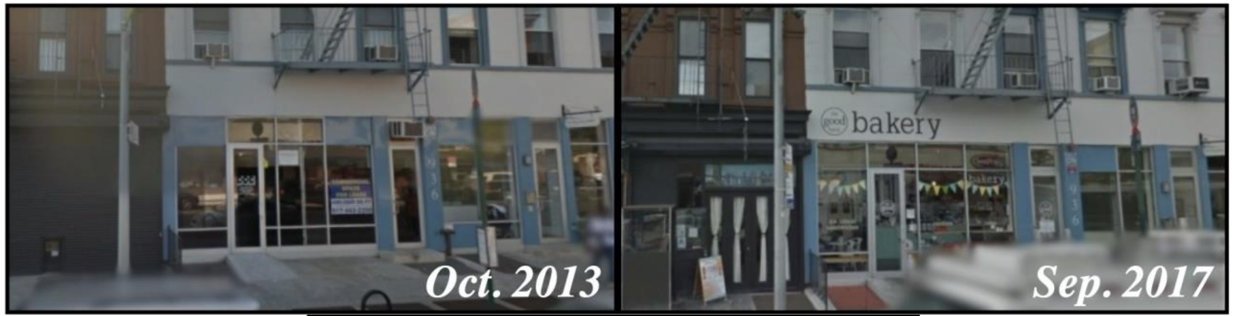
Results

Support **semantically conditioned** search, e.g. “retail store”

Some retail stores opened in NYC, 2011 - 2023.

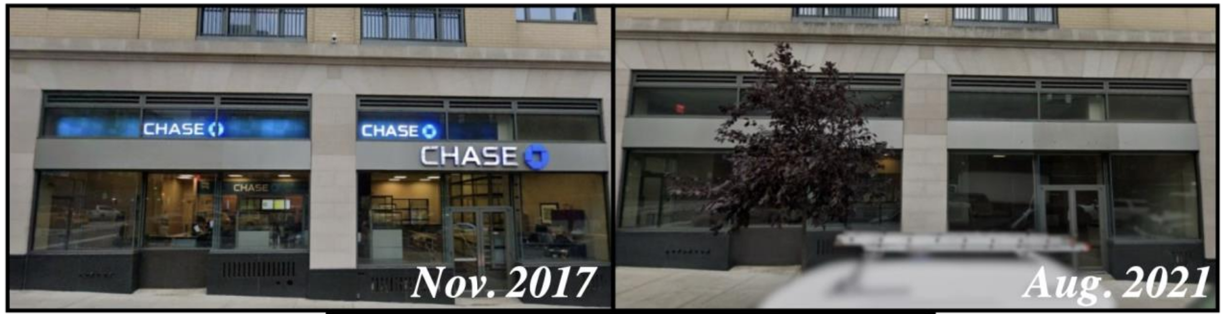


Juice Shops (318 opened)

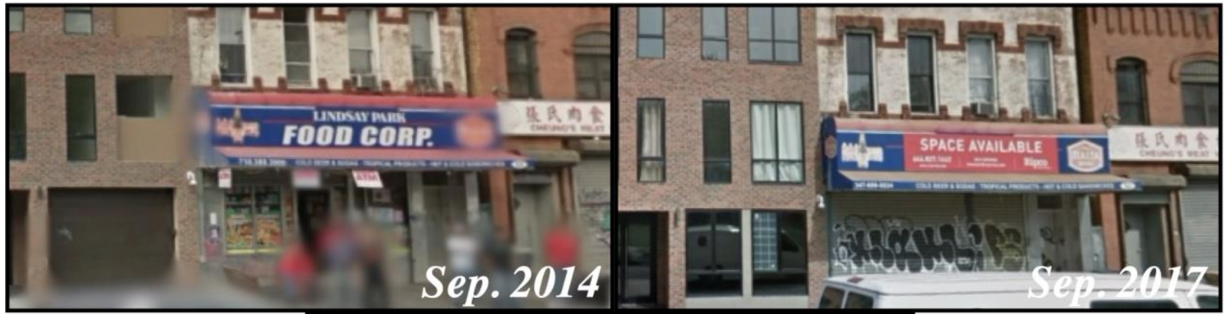


Bakeries (512 opened)

Some retail stores closed in NYC, 2011 - 2023.



Banks (1614 closed)

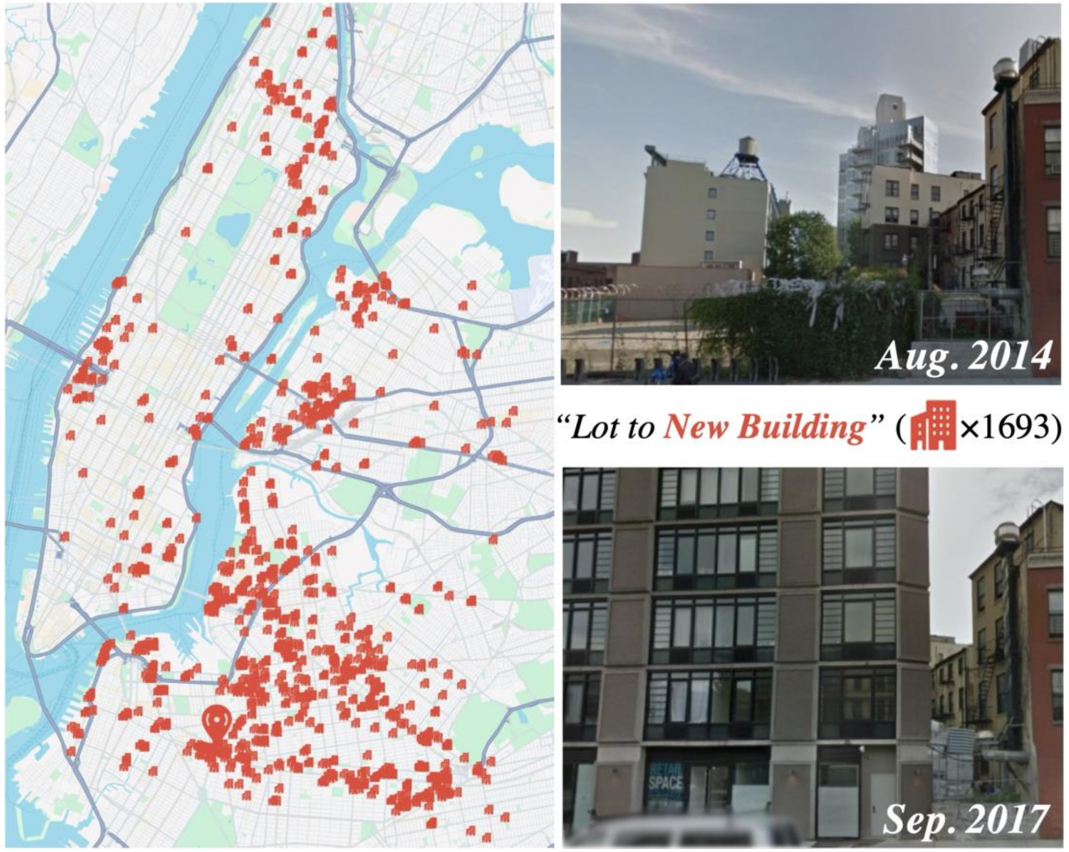


Groceries (741 closed)

Results

Other interesting applications

Where are new buildings built in NYC? **(A Spatial Insight)**

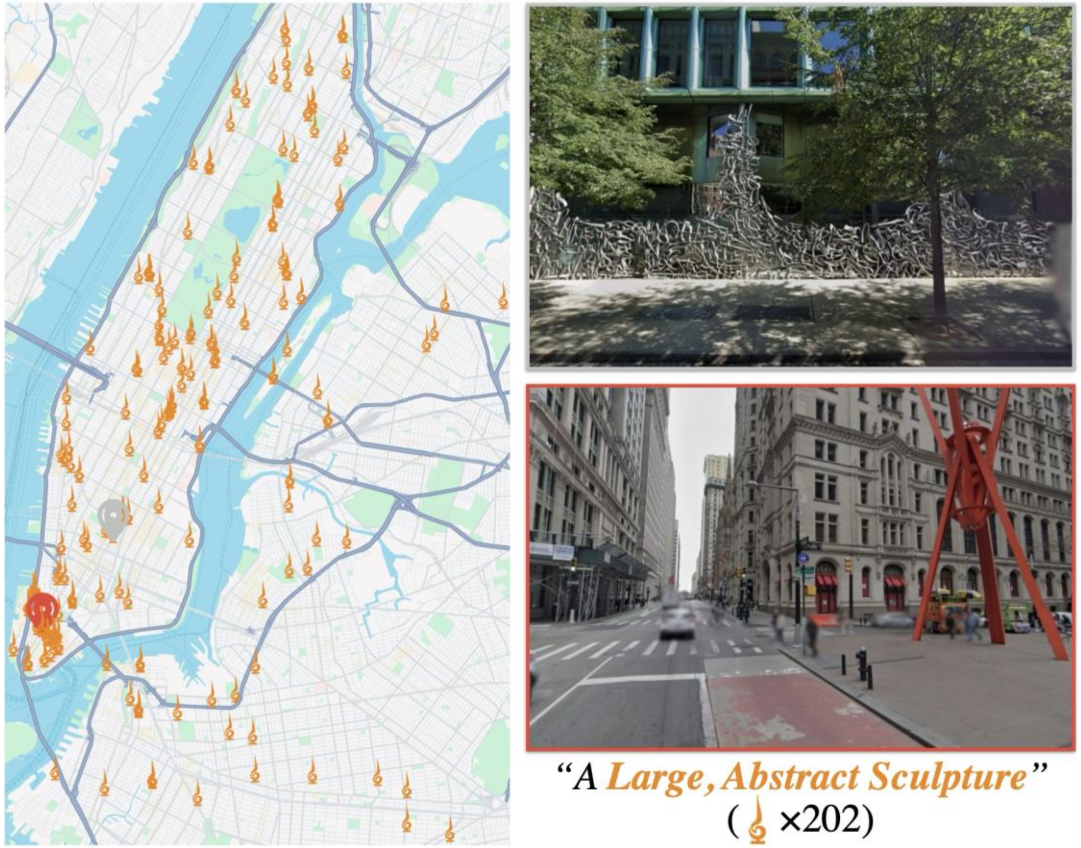


Aug. 2014

"Lot to *New Building*" (🏢×1693)

Sep. 2017

What are the unusual things in NYC? **(A Non-Temporal Query)**

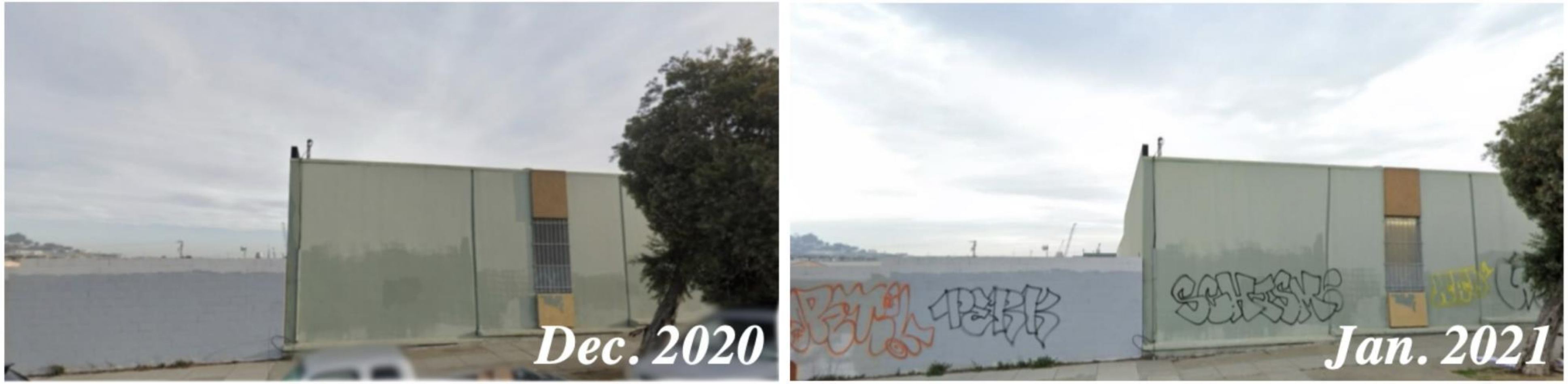


"A *Large, Abstract Sculpture*"
(🗿 ×202)

Results

Another Case Study

"Added Graffiti" were spotted ~3x more post-2020 (3152 times) than pre-2020 (1150 times).



Dec. 2020

Jan. 2021

> *"San Francisco deals with increasing graffiti …*
> *Especially after COVID …"*

abc7

> *"As part of the unprecedented COVID pandemic, the Board of*
> *Supervisors temporarily suspended Public Works' enforcement of*
> *the San Francisco Graffiti Ordinance …"*

SAN FRANCISCO PUBLIC WORKS

We **must be careful** when drawing socioeconomic conclusion.

Take-home Messages

- We study the open-ended analysis of massive image collection
- MLLMs as a critical tool to this problem
- Design a practical and effective system
- Find interesting insights about SF and NYC

Foundation Model for Visual Intelligence

From 2 Views to 10 Million

NoPoSplat

ICLR 2025 (**Oral**)

Visual Chronicles

ICCV 2025 (**Highlight**)

Building Visual Intelligence

Grounding

Reconstruct and understand 3D

Reasoning

Solve complicated tasks

Scaling

Foundation Model for Generalization

Action

Agent and tool use

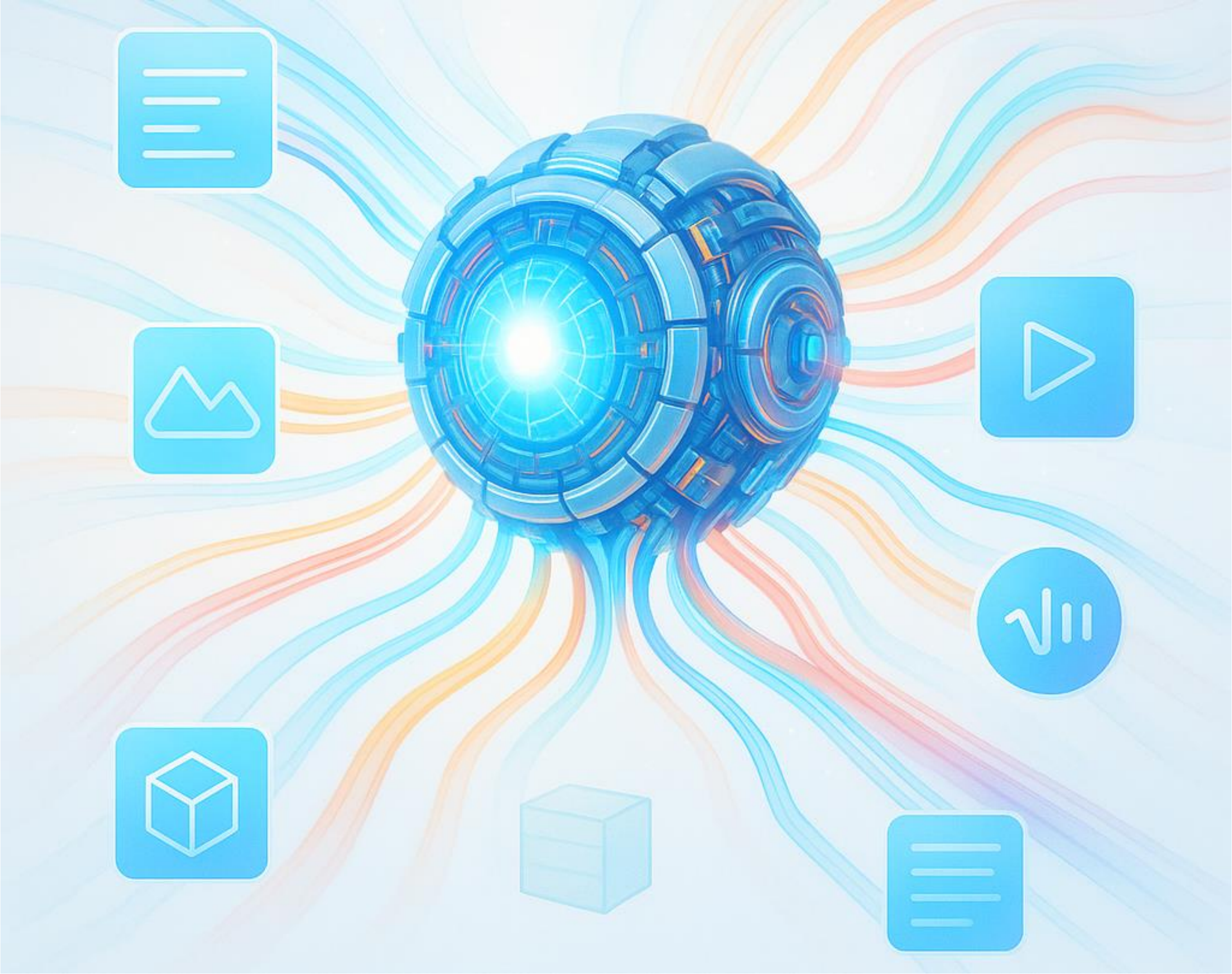
My Vision

Agentic perception is the future

Perception → Reasoning → Tools→ New Observation → Refinement

What is next?

Omni Model



Building Visual Intelligence

Songyou Peng

pengsongyou.github.io songyou@google.com @songyoupeng