CityRAG: Stepping Into a City via Spatially-Grounded Video Generation

Gene Chou[1,2] **Charles Herrmann**[1] **Kyle Genova**[1] **Boyang Deng**[3] **Songyou Peng**[1]
Bharath Hariharan[2] **Jason Y. Zhang**[1] **Noah Snavely**[1,2] **Philipp Henzler**[1]

[1]Google [2]Cornell University [3]Stanford University

Abstract

We address the problem of generating a 3D-consistent, navigable environment that is spatially grounded: a simulation of a real location. Existing video generative models can produce a plausible sequence that is consistent with a text (T2V) or image (I2V) prompt. However, the capability to reconstruct the real world under arbitrary weather conditions and dynamic object configurations is essential for downstream applications including autonomous driving and robotics simulation. To this end, we present CityRAG, a video generative model that leverages large corpora of geo-registered data as context to ground generation to the physical scene, while maintaining learned priors for complex motion and appearance changes. CityRAG relies on temporally unaligned training data, which teaches the model to semantically disentangle the underlying scene from its transient attributes. Our experiments demonstrate that CityRAG can generate coherent minutes-long, physically grounded video sequences, maintain weather and lighting conditions over thousands of frames, achieve loop closure, and navigate complex trajectories to reconstruct real-world geography. See our website for video playback.

1 Introduction

Imagine pulling up a photo of New York City, taken from the intersection of 42nd Street and 5th Avenue. Then, stepping into the image and walking toward the Empire State Building. Although the landmark is not visible in the input photo, as the virtual camera moves south the whole of the city—the roads, traffic lights, shops, fire hydrants, the Empire State Building itself—perfectly match the geographic layout of the real world. Furthermore, the environment preserves the specific weather conditions of the photo (a light drizzle around 2pm) and its elements come to life: a taxi completes its turn and a man in a blazer continues walking. In other words, the city is not a pure AI hallucination, but instead matches the real world; namely, the very world pictured in the input photo.

Such a capability would unlock applications in virtual tourism, gaming, and simulation for autonomous driving and robotics. For example, researchers could transform a snapshot of a snowstorm into a high-fidelity simulation to train self-driving cars, rather than driving thousands of miles in dangerous conditions [Waymo, 2026]. Specialized robots could be trained to adapt to a specific environment, such as a factory, and learn to avoid transient objects like people and boxes while navigating around the corners [Gao et al., 2026].

In this paper, we address the problem of generating a 3D-consistent, navigable environment that respects both the transient attributes of a first image condition, such as weather and pedestrians, and the static attributes derived from geospatial conditions, which take the form of pre-collected, geo-registered video frames, such as buildings and roads. Specifically, we focus on the domain of Street View for its dense coverage and semantic cues of the arrangement of static and dynamic elements. This allows us to ground generation in real-world environments.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

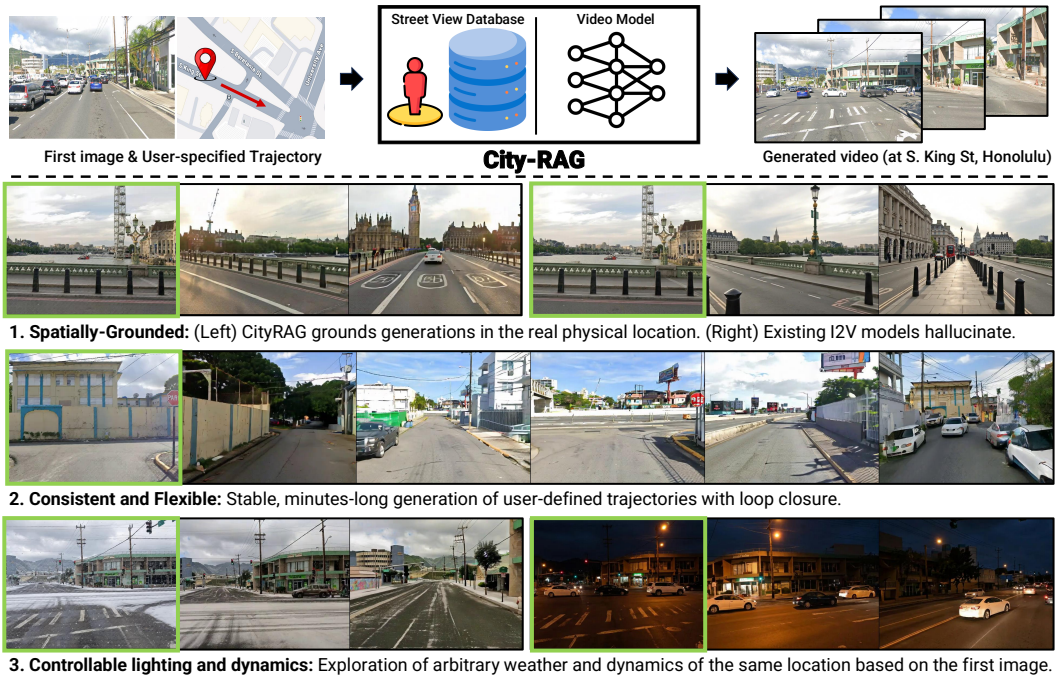

Figure 1: CityRAG generates minutes-long, spatially grounded video sequences that 1) Render real buildings, traffic lights, and roads of a city. From an image taken on the Westminster Bridge in London, CityRAG generates Big Ben and the Houses of Parliament as the viewpoint rotates left, while Veo [DeepMind, 2025] hallucinates. 2) Follow a user-defined path and perform loop closure after generating a thousand frames. 3) Are initialized from a first image and respects its weather conditions and dynamic objects. *Top:* The Westminster Bridge, London. *Middle:* Calle Quiñones St, San Juan. *Bottom:* S King St, Honolulu. Starting views are labeled with green bounding boxes.

Achieving this requires querying and incorporating external context on-the-fly, a task that is difficult for existing approaches. The dominant paradigm for generative models prioritizes scalability [Peebles and Xie, 2023, Blattmann et al., 2023] and thus relies on abundant and easily accessible data for conditioning, such as a text prompt or an image. But this approach cannot integrate external knowledge about the world during inference. On the other hand, non-generative 3D representations like NeRFs [Mildenhall et al., 2021] require dense captures of the exact moment and lack the capacity to produce realistic motion or complex appearance changes.

To this end, we propose CityRAG, a video generative model that leverages large corpora of geo-registered data as context to guarantee fidelity to the scene, while maintaining learned priors for complex motion and appearance changes.

Starting from an input image, CityRAG retrieves a multi-view "memory" of the location and injects it through a dedicated branch of attention layers. This architecture teaches the model to extract two distinct sets of information: transients from the image, such as lighting and dynamic objects, and statics from the "memory," such as buildings and roads. Through a carefully designed data-driven strategy, CityRAG learns to decouple and recombine these attributes, visualized in fig. 2 and fig. 3.

First, we curate a dataset of paired Street View videos that capture the same physical location at different times (e.g., morning vs. sunset) (section 3.1). This provides the data required for a model to semantically distinguish between static and transient attributes. Specifically, we collect a total of 5.5M Street View panoramas and their poses across 10 cities. These paired sequences allow a model to observe the same streets under diverse illumination and traffic conditions.

Second, we finetune a state-of-the-art I2V model, Wan 2.1 [Wan et al., 2025], on the paired data (section 3.2). While the pretrained model adheres to a first image condition, it lacks context beyond the immediate field of view. To address this, we introduce a training strategy that uses temporally unaligned frames (images of the same location captured at different times) as a structural anchor. By forcing the model to derive a static layout from morning frames to reconstruct a scene at night, we decouple permanent geometry from transient environmental conditions.

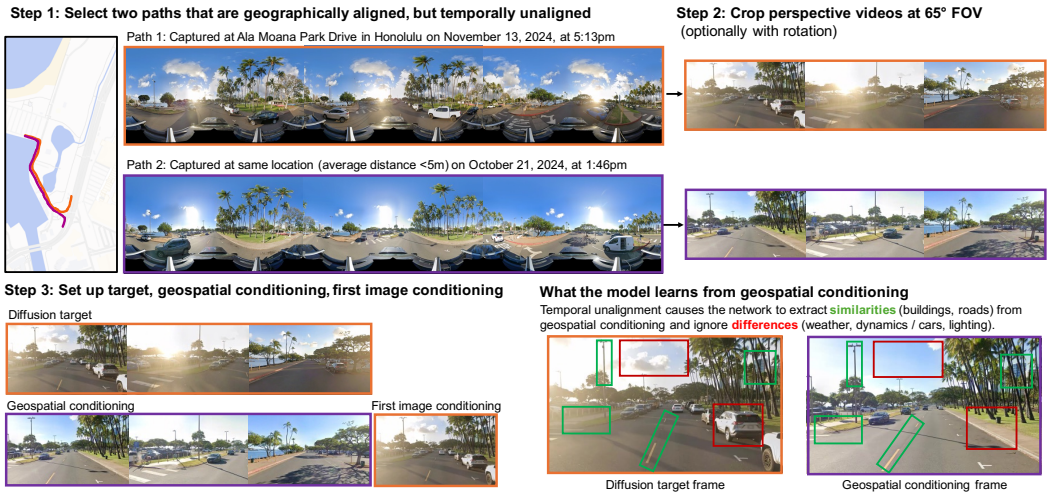

Figure 2: **Training data pipeline.** We use Street View data in the form of panoramas. We create a training pair if there is a continuous path where there exists 2 sets of captures at different times (e.g., morning vs. afternoon) but with an average distance < 5 meters, so the model learns to disentangle static and transient attributes, e.g., roads and buildings (green box) vs. weather and cars (red box).

During inference, given an input image and defined trajectory, CityRAG retrieves videos from the vicinity to serve as a reliable prior for the scene's identity (section 3.3). As the model learns to faithfully reconstruct the buildings and roads, generated videos remain consistent across independent and sequential samples, even without being trained for autoregression. The result is a model capable of generating minutes-long, 3D-consistent walkthroughs that simulate realistic motion of cars and pedestrians in a user's image while preserving the geography of the real physical location.

We evaluate our approach via a variety of metrics, testing scenes, and baselines. We show that our approach demonstrates strong 3D understanding of the underlying scene, disentangles dynamic and static elements without heuristics, and generates realistic sequences across diverse settings.

2 Related Works

2.1 Video Generative Models

Popular formulations for video generation include text-to-video (T2V) [Singer et al., 2022, Yang et al., 2024] and image-to-video (I2V) [Brooks et al., 2024, Blattmann et al., 2023, Bar-Tal et al., 2024] generation due to their scalability, and they can then be finetuned based on the requirements of downstream applications. Our application requires long-term consistency, pose control, and integration of external context.

Long-term consistency. Works in long-context or autoregressive generation [Chen et al., 2025a, Millon, 2025, Song et al., 2025, Zhang et al., 2025, Cai et al., 2026, Xiao et al., 2025, Huang et al., 2025] maintain consistency by balancing computational efficiency and storing past samples. Another line of work creates an explicit memory like point clouds [Wu et al., 2025, Gu et al., 2025, Ren et al., 2025, Yu et al., 2025]. However, these works rarely show the capacity to generate minutes-long videos without significant degradation, and have an orthogonal focus to our work. CityRAG retrieves external context for grounding, rather than past samples, to maintain consistency.

Pose-conditioning. Pose-conditioned models [Ren et al., 2025, Bahmani et al., 2025, Guo et al., 2023, Van Hoorick et al., 2024, Wang et al., 2023, Tung et al., 2024, Zhou et al., 2025] finetune a base generative model on camera poses, often in the form of camera parameters or warping and inpainting. They rely on generative priors of video models to remain temporally consistent and hallucinate plausible sequences while providing control. We similarly condition our model on camera extrinsics, but additionally adhere to large-scale real-world grounding.

Using additional context. Reference-to-video (R2V) [Chen et al., 2025b, Wei et al., 2024, Wang et al., 2024] and video-to-video (V2V) [Esser et al., 2023, Geyer et al., 2024, Wu et al., 2024, Liang

et al., 2024b, Ku et al., 2024, Zhou et al., 2025, Liang et al., 2024a, Fu et al., 2025] models are conceptually closer to our goal. But neither has shown 3D-awareness and both require strict adherence to the reference video. A few works experiment with conditioning without strict adherence. For example, LooseControl [Bhat et al., 2023] enables boundary control and scene editing with sparse depth maps. KFC-W [Chou et al., 2025] generates a 3D-consistent trajectory of a scene from random internet photos. However, none of them address a similar problem setting as ours.

2.2 Retrieval-Augmented Generation (RAG)

RAG has been shown to mitigate hallucination and ground model outputs in external knowledge [Lewis et al., 2020]. Recently, this framework has been applied to visual generative models to enhance fidelity and realism. RealRAG [Lyu et al., 2025] improves text-to-image synthesis by retrieving real-world reference images to fill in knowledge gaps during generation. MotionRAG [Zhu et al., 2025] retrieves video clips to provide demonstrations of motion. In a similar vein, our work retrieves geo-registered data to ground video generation in the real world.

3 Method

3.1 Data

With explicit permission from Google, we collect Street View data from Google Maps across 10 diverse cities scattered across the globe: Paris, Athens, Anchorage, Hyderabad, Philadelphia, San Francisco, San Juan, Honolulu, London, and Sao Paolo. Importantly, all sensitive information, such as license plates and faces, are blurred prior to collection.

The data is in the form of panoramas and their associated poses in the Earth-Centered, Earth-Fixed (ECEF) coordinate system. Thus, the poses are in metric scale (meters) and consistent across all cities. We sample with density roughly equivalent to 10 FPS. In addition, we sample captures of the same streets at different times, when available. We collected a total of 5.5M pano-pose pairs.

Then, we group all panoramas by their trajectories and time of capture. We create a training pair if we find a continuous path in a city of length N where there exists 2 sets of panoramas located along the same path with an average distance threshold smaller than ϵ meters, but captured at different times (e.g, different dates, or even morning vs. afternoon of the same day). We set N=73 the number of frames sampled for training, and $\epsilon = 5$. After filtering for these pairs, we obtain a total of 1.3M panoramas for training and a few thousand held out for testing. We show an example pair in fig. 2.

3.2 Architecture

We finetune from a state-of-the-art image-to-video (I2V) generative model, Wan 2.1 (14B) [Wan et al., 2025]. It consists of a spatio-temporal VAE and a diffusion transformer (DiT). We refer readers to the original paper for details.

Our goal is to condition the model on a first image that initializes the scene, a defined trajectory, and geo-registered data in the form of video frames along the defined trajectory. We visualize our architecture in fig. 3.

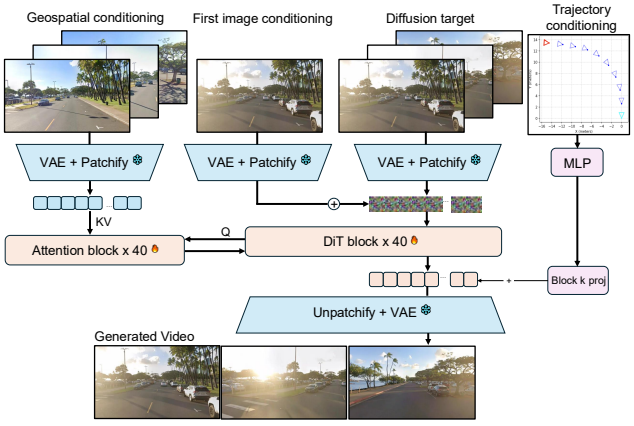

Figure 3: **Architecture.** The generator takes three conditions: the first image, a trajectory, and geo-registered videos (denoted geospatial conditioning) along the trajectory.

First image conditioning. We follow the same conditioning in the I2V base model. The first image of the target video is independently processed by the VAE, then padded and concatenated channel-wise to the the target latents. This image initializes the scene: the generated video is expected to follow its lighting conditions and animate the dynamic objects in the frame.

Trajectory conditioning. We specify a trajectory as a list of 4x4 extrinsic matrices. During training, we convert them to relative poses on-the-fly, where the first frame of the trajectory is at the origin. The poses are originally in ECEF coordinates, so the relative poses are in metric scale (meters) and consistent across cities.

We then flatten the matrices, downsample the temporal dimension by 4x with a Conv1D layer to match the temporal downsampling of the VAE, process them with a two-layer MLP, and then use a zero-initialized projection layer to match the dimensions of the Wan model, one for each attention block. The output of the k-th projection layer is added to the output of the k-th DiT block. This allows pose information to be weighted depending on which block is more important for handling video movement, without disrupting the video prior.

However, one limitation of Street View data is that the majority of trajectories move straight along driving paths. To improve generalization, we augment rotations by cropping the panoramas at random yaws. Specifically, we randomly select a yaw between 0 and 360 degrees as the starting viewing angle, and add a rotation uniformly sampled between 0 and 2 degrees between each frame. We show that our model generalizes to out-of-distribution rotations in fig. 7.

Geospatial conditioning. We sample from the set of paired panoramas during training. One serves as the target video to generate, the other as context for grounding. We crop both to a fixed 65° field of view (FOV), with the yaw following the rotation augmentation determined by the camera pose.

Because these captures are separate traversals, they exhibit spatial and temporal discrepancies. Spatial shifts occur due to variations in camera centers (e.g., lane changes), while temporal misalignments result from varying vehicle speeds. To ensure our model understands the relation between the two sets of captures and the underlying 3D scene, we vary the length of the condition panoramas during training. This forces the model to become robust to these discrepancies, rather than relying on a one-to-one mapping between frames. We show in fig. 6 that the generated sequences can contain accurate renderings of buildings that appear much later in the condition frames.

We pass the conditions to the Wan model via cross-attention. We duplicate the original self-attention blocks from the pretrained base model, but train them separately (denoted Attention block in fig. 3). During training, we first pass the condition video through the VAE, then use the latents as the keys and values for cross-attention. The target noisy latents serve as the query. This strategy allows each frame in the target sequence to attend to the entire context of the condition.

Since cross-attention sequences can be varying length, we randomly set the number of conditioning frames to be between 61 and 81, such that the number is mismatched to that of target frames, to force the model to extract global context rather than only pixel-aligned correspondences.

Classifier-free guidance [Ho and Salimans, 2022]**.** We set the unconditional probability to 10% for both the poses and panorama conditions, but sampled independently, so both can still be valid conditions in the absence of the other. Importantly, we find that without geospatial conditions, the model falls back to being a trajectory-conditioned I2V model, showing that it retains strong generative priors without only relying on external context. The left of fig. 7 shows such an example. The geospatial condition is completely mismatched to the trajectory due to traffic and therefore effectively ignored, yet the model follows the other conditions and generates a plausible video.

3.3 Inference via User Input and RAG

As shown in fig. 4, we describe the full pipeline of generating a consistent, minutes-long video given a user-defined trajectory and first image condition.

First, we randomly pick an image from the dataset, or a casual capture from the internet, or even a AI-modified image with snowy conditions (Honolulu scene in fig. 1). We identify the location on the map and ask the user for a trajectory (fig. 4, Steps 1 and 2). Next, we retrieve geo-registered Street View data along the defined path, and use them as conditioning to generate a video (Steps 2 and 3). Then, we can repeat this step autoregressively until we reach a desired location (Step 4).

The defined path may not exist as a single geo-registered video, so we stitch frames from multiple videos to navigate arbitrary trajectories. As shown in fig. 5, the initial retrieved path continues straight, so CityRAG retrieves a distinct video from cross traffic. By stitching the two videos, we create a proxy trajectory that turns right at the intersection (at a 90 degree angle). Though the model was always trained on continuous geospatial videos, the generated videos remain consistent with discontinuities

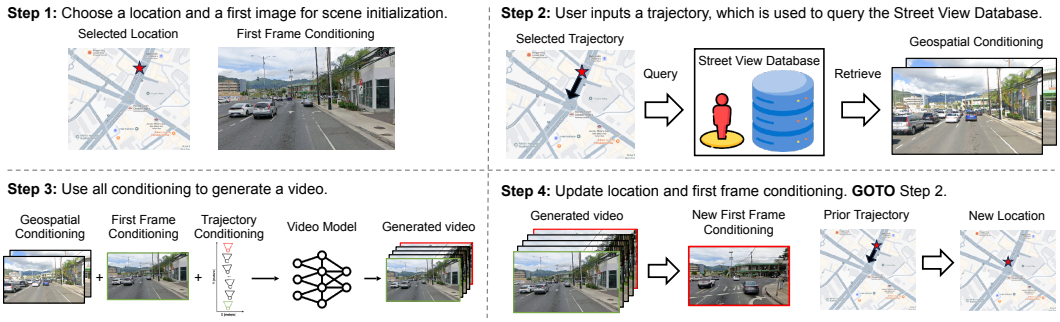

Figure 4: **RAG pipeline at inference-time.** The user first selects a location and image that they want to step into. Then with a user-specified trajectory we use the Street View Database to retrieve our geospatial conditioning. All conditions are passed to the video model which generates the output the user sees. We then automatically update the first frame and location and repeat the process.

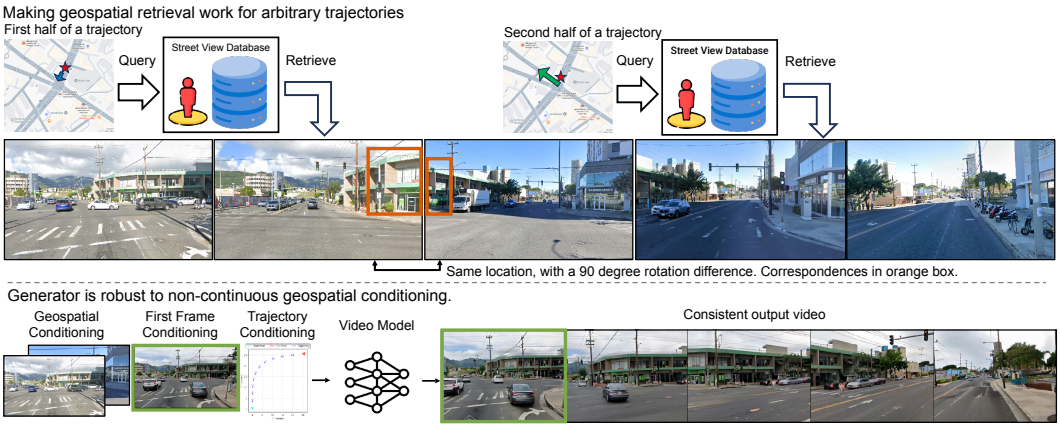

Figure 5: **Making geospatial retrieval work for arbitrary trajectories.** Navigating arbitrary trajectories may require stitching together distinct videos from the database. In this example, since the initial retrieved path continues straight, CityRAG retrieves a second, perpendicular path from the same intersection to construct a new trajectory that resembles turning right at the intersection. Despite the discontinuity in the geospatial condition frames, the generator produces a consistent video, indicating its robustness and its understanding of the static and transient elements in a scene.

in the conditions during testing. This indicates our model understands static and transient elements in a scene and is robust to appearance changes and pixel-mismatches in the conditions.

Our generated video of San Juan in fig. 1 is conditioned on multiple (4) stitched geospatial videos, yet remains consistent across a thousand frames. We show more examples on our website.

4 Experiments

Baselines. To the best of our understanding, there are no open-source video generation models that perform our task, which requires navigation control (trajectory conditioning) and adherence to both an initial condition (first image) for scene appearance and an external condition (geospatial video) for spatial grounding. We identify three closely related lines of work and run baselines from each:

1) I2V + pose control. We use Gen3C [Ren et al., 2025], a state-of-the-art video model with camera control. It shows driving simulations as one of its applications.

2) V2V + pose control. We use another variant of Gen3C and TrajectoryCrafter [Yu et al., 2025]. Both methods take a dynamic input video and re-render it given a different trajectory. For our setup, we provide the conditioning frames and re-render with the target camera trajectory.

3) V2V + style transfer. We use AnyV2V [Ku et al., 2024], a method that transforms a video to the style of an image. We provide the geospatial video as input, and the first image as the style reference.

Train-test split. From the 10 cities in our dataset, we use the first 8 for both training and testing. For testing, we hold out entire neighborhoods to ensure samples come from streets unseen during training. We reserve the remaining two cities, London and São Paulo, only for testing. In short, none of the quantitative and qualitative results shown in the paper are of streets seen during training. We did not observe any differences in generated visual quality for held out sets of training cities and for completely withheld cities. See videos in the supplement for a visual comparison and Section 4.2 for metrics.

For testing, we filter for trajectories that feature at least a 45° rotation to avoid models relying solely on context from the first image. Then, from each city, we randomly select 10 trajectories. For comparing with baselines, we do not perform autoregressive generation or provide a user-defined trajectory, but use preprocessed pairs of trajectories, as described in Section 3.1.

4.1 Qualitative Comparisons

In Figure 6, we highlight and analyze two challenging test samples, and show dozens of video results in the supplement. The video for geospatial conditioning (leftmost column), trajectory defined by the target video (rightmost column), and the first image of the target video are provided as conditions. CityRAG successfully handles these scenarios. In Scene A, the generated video follows both the weather conditions and the cars of the first image. As the video progresses, the black car in front continues to move realistically, and reappears even when it goes out of sight during the turn.

In Scene B, we show that CityRAG follows pose precisely even when there is a mismatch between it and the geospatial condition. Specifically, the geospatial condition stops at the intersection to yield to oncoming cars (see t=4s). However, the generated video follows the pose and accurately renders the structure at t=7s that only appears in the geo conditioning at t=10s. This shows our model can extract and render the structure of the scene, rather than relying on a pixel-aligned transfer as in V2V models. We also show a similar example in Figure 7. The capture representing the geospatial condition gets stuck in traffic, yet the model produces a plausible generation.

All baselines fail. AnyV2V copies over the first image but fails to reconcile the differences between the source video (geospatial condition) and the first image. In both scenes, the virtual camera never moves. Gen3C (I2V) shows stability and relatively high visual quality in the first few seconds (t=1s to 4s) when the car is only moving forward, but its generation breaks down when the car turns. Gen3C (V2V) struggles even more with complex poses. Through testing, we find that it can only re-render videos with very limited camera movement (e.g., a small wobble).

Flexibility of trajectory conditioning. We further demonstrate CityRAG's robustness and flexibility. In Figure 7, the geospatial video is stuck in traffic. However, CityRAG's generation follows the defined trajectory to move forward and take a left turn, showing its strong generative priors even without external context. Our model can also follow extreme rotations, such as 360° within a single sequence, which is double the maximum rotation present in the training set.

4.2 Quantitative Comparisons

Reconstruction and fidelity. A major objective of our task is fidelity to the ground truth scene. Thus, we first evaluate with metrics including PSNR, LPIPS [Zhang et al., 2018], and SSIM [Wang et al., 2004] (Table 1). Since we are focused on static structures, we also evaluate on a static-variant of these metrics (denoted -S). Specifically, we use Mask2Former [Cheng et al., 2022] to segment dynamic classes (i.e., vehicles and people), and mask their pixels during calculation. We also include FID [Heusel et al., 2017] that assesses the quality of generated images by comparing their feature distributions to those of real images. Lower indicates generations are more similar to real images.

Compared to dedicated view synthesis or reconstruction techniques, all methods we test obtain relatively low scores, including ours. This is because generative models are inherently stochastic and do not aim for the exact pixel-level reconstruction or overfitting that traditional NVS methods prioritize. Minor shifts in camera movement or hallucination of plausible geometry can lead to high pixel-wise error. Despite this, CityRAG outperforms all baselines and maintains the best fidelity to the ground truth scenes, which matches our qualitative observations. Furthermore, our method significantly leads in metrics that measure perceptual similarity, such as LPIPS and FID.

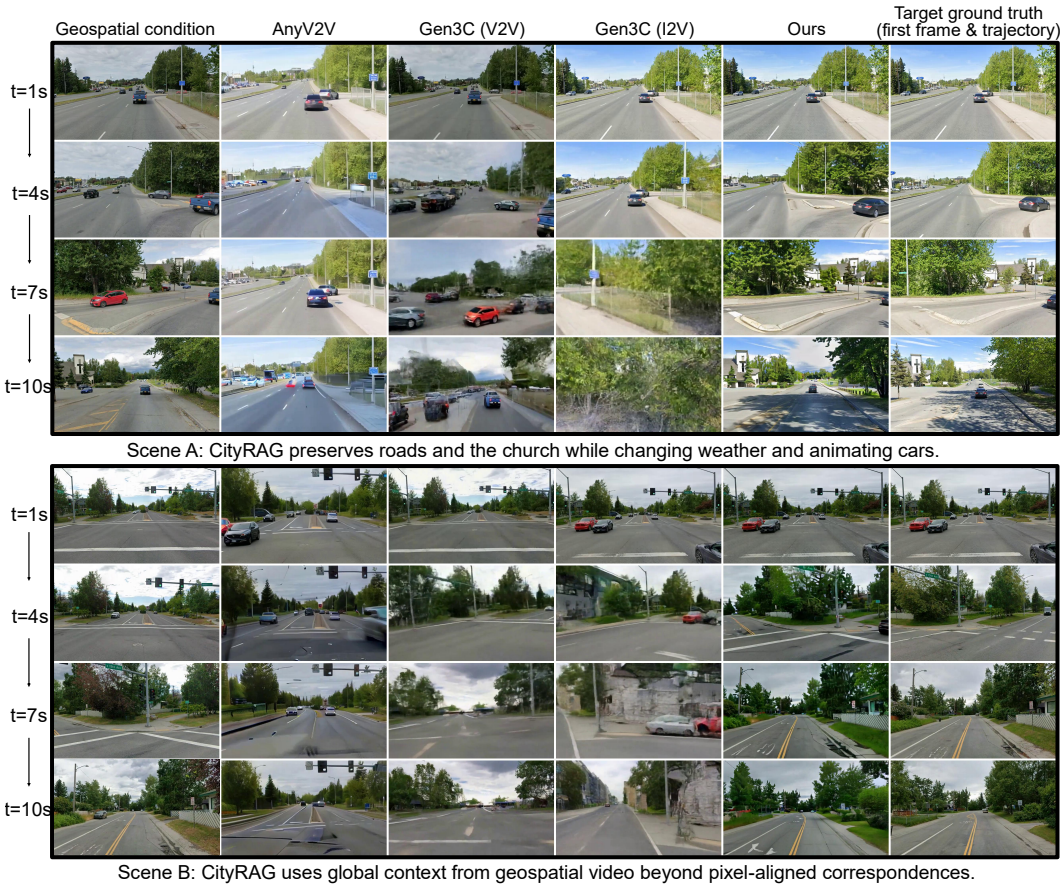

Figure 6: **Qualitative comparisons.** We show two challenging test samples. Input conditions include the video for geospatial conditioning (leftmost column) and the first image and the trajectory of the ground truth video (rightmost column). *Scene A:* CityRAG follows the weather and animates the black car in the first image, and faithfully renders the church and roads. *Scene B:* CityRAG renders the building and fence (t=7s) that appear later in the geospatial conditioning (t=10s, lags due to traffic), showing its ability to extract global context rather than only pixel-aligned details.



Figure 7: **Flexibility of trajectory conditioning.** Our trajectory conditioning does not have to be precisely aligned with the geospatial conditioning. *Left:* Even though there is a mismatch between the geospatial condition (car stuck in traffic) and trajectory (left turn), our model generates a plausible sequence following the trajectory. *Right:* Our model can rotate 360° in a single sequence. The low visual quality is an artifact of the temporal VAE.

We also note that we did not observe any meaningful performance gap between the held out scenes of the eight trained cities and the two untrained cities, suggesting that our method is generalizable to a variety of diverse scenes and conditions. Specifically, the PSNR, SSIM, LPIPS, and FID scores of the generations of the untrained cities were 15.11, 0.461, 0.517, 16.90, respectively, comparable to that of the full test set.

Table 1: **Reconstruction and fidelity.** We calculate view synthesis metrics to measure the fidelity of generated videos to real-world scenes, and FID to measure visual quality.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR-S ↑	SSIM-S ↑	LPIPS-S ↓	FID ↓
TrajCrafter	11.90	0.403	0.705	11.92	0.536	0.548	55.45
AnyV2V	11.82	0.385	0.698	11.83	0.521	0.551	47.56
Gen3C V2V	12.34	0.432	0.677	12.36	0.538	0.558	57.13
Gen3C I2V	13.28	0.453	0.654	12.86	0.545	0.543	61.07
Ours	**15.03**	**0.466**	**0.504**	**15.86**	**0.560**	**0.432**	**16.55**

Table 2: **Trajectory adherence and static-layout consistency.** Results are averaged over all 100 test scenes. Translation errors are normalized by reference path length; rotation errors are in degrees. RPE uses one-second intervals. Static DINO measures cosine similarity of reprojected DINOv2 features in static regions.

	Trajectory adherence				Static layout
Method	ATE RMSE ↓ (% path)	Rot. RMSE ↓ (°)	RPE trans. ↓ (% path)	RPE rot. ↓ (°)	Static DINO ↑
TrajCrafter	13.69	102.60	13.14	10.21	0.339
AnyV2V	22.44	128.64	17.29	11.22	0.240
Gen3C V2V	7.78	57.88	7.92	4.55	0.370
Gen3C I2V	9.45	82.89	10.92	6.55	0.299
CityRAG	**2.49**	**6.32**	**1.95**	**3.84**	**0.611**

Trajectory adherence. We evaluate how closely generated videos follow the reference camera trajectory. We recover camera poses from both the generated and ground truth videos using MegaSaM [Li et al., 2025], and align each generated trajectory to its reference with a single $\mathrm{Sim}(3)$ transformation, accounting for rotation, translation, and scale. We report absolute trajectory error (ATE) as the root mean squared error (RMSE) of camera positions, and rotation RMSE as the error in camera orientations at corresponding frames. We also report relative pose error (RPE) in translation and rotation over one-second intervals, measuring discrepancies in local camera motion. Translation errors are expressed as percentages of the reference trajectory's total path length, and rotation errors are measured in degrees. We average the results over all 100 test scenes; lower is better for all four measures. As shown in Table 2, CityRAG achieves the lowest errors throughout, with an ATE of 2.49% and rotation RMSE of 6.32°, compared to 7.78% and 57.88° for Gen3C V2V. These results agree with the qualitative observations that CityRAG follows trajectories with substantial camera rotation.

Static-layout consistency. We measure the similarity of static scene content after accounting for viewpoint differences using reprojected DINOv2 [Oquab et al., 2023] features. We segment generated and ground truth images with OneFormer Swin-L [Jain et al., 2023], retaining patches containing only road, sidewalk, building, wall, fence, pole, traffic light/sign, or terrain classes, and excluding vegetation, sky, people, and vehicles. Using MegaSaM depth and camera poses, we align each generated sequence to the ground truth with a single $\mathrm{Sim}(3)$ transformation, lift generated-image DINOv2 patch centers into 3D, and project them into the corresponding ground truth camera. We retain correspondences only when both patches are static, rejecting out-of-bounds projections, semantic boundaries, and depth discontinuities. We compute the cosine similarity between matched patch features and report the mean static DINO score across all 100 test scenes; higher is better. CityRAG achieves 0.611, compared to 0.370 for the strongest baseline, Gen3C V2V, indicating greater consistency with the static layout of the reference scene.

4.3 User Study

We conduct a user study consisting of three questions, asking users to evaluate 1) visual quality, 2) whether videos are smooth continuations of the first image, and 3) fidelity to the physical location (using the last frame of the geospatial conditioning video as the reference destination). Users rate each sample 1 (lowest), 2, or 3 (highest). We provide details in the supplement.

The responses are plotted in Figure 8. The x- and y-axes record the average scores for the second and third questions, respectively, and the radius indicates visual quality; larger is higher. In addition to our baselines, we include the Wan I2V base model and the retrieved geo-registered data as a reference for the specific axis each specializes in. Again, we observe that our task is a novel yet practical problem setting that existing methods cannot perform.

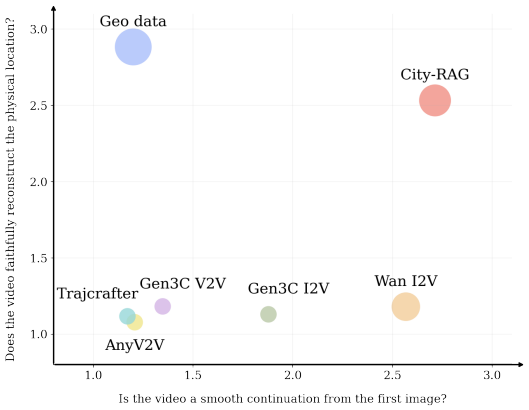

Figure 8: **User study results.** Users rate each video on a scale of 1 (lowest) to 3 (highest). Only CityRAG generates videos that are both smooth continuations from first images and faithful renders of the real physical location.

4.4 Discussion

We acknowledge that the baselines are not trained for our task, and therefore expectedly fail. The closest approach, conceptually, would be methods like AnyV2V that perform style transfer on a video; i.e., transform an input video (geospatial condition) into a desired appearance (first image). But this transfer is extremely non-trivial. The model would need to understand that the cars, pedestrians, and weather are all part of the "style." It would then need to have the flexibility to animate the cars and pedestrians realistically. And fundamentally, the trajectory of the generated video could only follow that of the input video, rather than taking user-defined trajectories.

Thus, we highlight the robustness and flexibility of our approach: CityRAG creates a simulated environment of a scene by understanding the global structure even when geospatial conditions are imperfect, realistically preserves and animates the scene initialization, and allows users to freely navigate arbitrary trajectories.

5 Conclusion

To the best of our knowledge, CityRAG is the first video generative model that emphasizes adherence to our real world, and therefore it can help unlock a variety of applications that rely on specific environment layouts. Furthermore, we introduce a fully data-driven strategy that teaches the model to be aware of semantics, generalizable to diverse conditions, and to disentangle static and dynamic attributes via temporally unaligned training data.

In the appendix, we provide additional details including (A) Limitations and future work, (B) Latency and inference costs, (C) Training optimization, (D) Architecture ablations, (E) User study details, (F) Extended related works, (G) Ethics and privacy. We also encourage readers to view video results hosted on our website.

Acknowledgements

Gene Chou was supported by an NSF graduate fellowship (2139899). We thank Gordon Wetzstein, Aleksander Holynski, Jon Barron, Dor Verbin, Pratul Srinivasan, Rundi Wu, Ruiqi Gao, and Haofei Xu for discussions and support.

References

Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 72–79, 2009. doi: 10.1109/ICCV.2009.5459148.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *ACM SIGGRAPH Asia 2024 Conference Papers*, 2024.

Shariq Farooq Bhat, Niloy J. Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. *arXiv preprint arXiv:2312.03079*, 2023. doi: 10.48550/arxiv.2312.03079.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. In *ICLR*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Yuwei Fang, Kwot Sin Lee, Ivan Skorokhodov, Kfir Aberman, Jun-Yan Zhu, Ming-Hsuan Yang, and Sergey Tulyakov. Video alchemist: Multi-subject open-set personalization in video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Xuyang Chen, Conglang Zhang, Chuanheng Fu, Zihao Yang, Kaixuan Zhou, Yizhi Zhang, Jianan He, Yanfeng Zhang, Mingwei Sun, Zengmao Wang, Zhen Dong, Xiaoxiao Long, and Liqiu Meng. Driving with dino: Vision foundation features as a unified bridge for sim-to-real generation in autonomous driving. *arXiv preprint arXiv:2602.06159*, 2026.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

Estelle Chigot, Dennis G. Wilson, Meriem Ghrib, and Thomas Oberlin. Style transfer with diffusion models for synthetic-to-real domain adaptation. *Computer Vision and Image Understanding*, 259: 104445, 2025. ISSN 1077-3142. doi: 10.1016/j.cviu.2025.104445.

Gene Chou, Kai Zhang, Sai Bi, Hao Tan, Zexiang Xu, Fujun Luan, Bharath Hariharan, and Noah Snavely. Generating 3d-consistent videos from unposed internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL `https://arxiv.org/abs/2411.13549`.

Google DeepMind. Veo: a text-to-video generation system, 2025. URL `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`.

Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasios Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Xiao Fu, Shitao Tang, Min Shi, Xian Liu, Jinwei Gu, Ming-Yu Liu, Dahua Lin, and Chen-Hsuan Lin. Plenoptic video generation. *arXiv preprint arXiv:2601.05239*, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K.R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, 2024.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction, 2026. URL `https://arxiv.org/abs/2601.04090`.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2989–2998, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10486–10496, 2025.

Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. *arXiv preprint arXiv:2405.15757*, 2024a.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. URL https://arxiv.org/abs/2502.16982.

William Ljungbergh, Bernardo Taveira, Wenzhao Zheng, Adam Tonderski, Chensheng Peng, Fredrik Kahl, Christoffer Petersson, Michael Felsberg, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. R3d2: Realistic 3d asset insertion via diffusion for autonomous driving simulation. *arXiv*, 2025. doi: 10.48550/arxiv.2506.07826.

Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.00848.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *ICCV*, 2021.

Erwann Millon. Krea realtime 14b: Real-time video generation, 2025. URL https://github.com/krea-ai/realtime-video.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=mFppY38Z36C.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Uriel Singer, Adam Polyak, Thomas Hayes, Dafna Shaham, Chitwan Saharia, William Chan, and Mohammad Norouzi. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. URL https://arxiv.org/abs/2209.14792.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2006. URL https://api.semanticscholar.org/CorpusID:13385757.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. URL https://arxiv.org/abs/2502.06764.

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022.

Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12931, 2022.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision (ECCV)*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. URL https://arxiv.org/abs/2401.09962.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2023.

Waymo. The waymo world model: A new frontier for autonomous driving simulation. Waymo Blog, February 2026. URL https://waymo.com/blog/2026/02/the-waymo-world-model-a-new-frontier-for-autonomous-driving-simulation.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024.

Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory, 2025. URL https://arxiv.org/abs/2506.05284.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory, 2025. URL https://arxiv.org/abs/2504.12369.

Zhuoran Yang, Xi Guo, Chenjing Ding, Chiyu Wang, Wei Wu, and Yanyong Zhang. Instadrive: Instance-aware driving world models for realistic and consistent video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 25410–25420, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 100–111, October 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Hexu Zhao, Haoyang Weng, Daohan Lu, Ang Li, Jinyang Li, Aurojit Panda, and Saining Xie. On scaling up 3d gaussian splatting training, 2024. URL https://arxiv.org/abs/2406.18533.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023.

Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint*, 2025.

Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. Simgen: Simulator-conditioned driving scene generation. *arXiv preprint arXiv:2406.09386*, 2024.

Chenhui Zhu, Yilu Wu, Shuai Wang, Gangshan Wu, and Limin Wang. Motionrag: Motion retrieval-augmented image-to-video generation. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, 2025.

A Limitations and Future Work

There are a few limitations of CityRAG. First, we perform autoregression by only providing the generated last frame as the first frame of the subsequent sample. Existing methods for autoregression could be incorporated to improve long-term consistency. However, we do want to highlight the stability that geospatial grounding brings. Using the last frame as the first image of the next generation leads to significant drift and degradation after just 1 or 2 iterations in typical I2V models, yet our generated scenes remain stable (especially the static structures) even after dozens of iterations. We believe that this technique would complement existing autoregression methods.

Second, we provide no heuristics to the model regarding static vs. transient objects — the disentanglement is completely data-driven. Future work could include fine-grained control and annotations over individual elements in the scene to improve controllability and customization.

Third, though we mention potential downstream applications such as virtual tourism and simulations, our method is not real-time. A truly interactive video world model requires multiple dimensions of progress, including faster inference, long-horizon temporal consistency, realism, controllability, and more. CityRAG specifically focuses on spatial grounding to real physical locations for realism and long-horizon temporal consistency, and is orthogonal to other dimensions. We describe our inference costs and latency in the next section.

Fourth, because we do not have captions for our data, we freeze the original text cross-attention blocks and use a fixed prompt: "A photorealistic, cinematic video of a city street. The camera performs a smooth, steady tracking shot moving along the asphalt road, maintaining a consistent level angle that offers an immersive street-level perspective." After finetuning, we observe the model no longer responds to new text captions; we leave captioning Street View and text conditioning for future work.

Finally, there exist data biases. For instance, the data does not include snowy, rainy, or nighttime conditions due to hardware and sensor limitations. Augmenting the data, and perhaps introducing modalities like text, is another important future work.

B Latency and Inference Costs

During training and inference, the target video consists of 73 frames at 480p resolution (832×480). The VAE downsamples the temporal dimension by $4\times$ and spatial dimensions by $8\times$. With a DiT patch size of 2, the resulting latent size is $18 \times 30 \times 52$ ($T \times H \times W$).

The inference costs can be divided into two parts: retrieving the geospatial conditions from the Street View database, and the video generation. We first built the entire map of a city using Scipy's cKDTree, such that given any image with GPS information, the lookup time for the nearest neighbor is O(log N). This is a one time cost and less than a minute in wall time. All sequential frames following the desired relative poses can be retrieved in O(1) constant time.

The video generation cost is comparable to the Wan 2.1 base model. We used multiple inference tricks such as TeaCache [Liu et al., 2024] and DeepSpeed-Ulysses [Jacobs et al., 2023]. Inference is 40 steps using the UniPC [Zhao et al., 2023] solver.

On a node of 8 A100 GPUs, the inference wall time measured 90 seconds for 73 frames, with an amortized cost of roughly 0.8 FPS. On a node of 8 B200s, the time was reduced to 30.5 seconds, or 2.4 FPS.

As mentioned in the previous section, faster inference was not the focus of this project, but we believe ongoing works in distillation, few-step generation, caching...etc will continue to improve latency. For instance, a model distilled to 4 steps would roughly reduce the inference time 10 fold.

C Training Optimization

We adopt the v-prediction [Salimans and Ho, 2022] objective with a shifted noise schedule toward higher timesteps (a factor of 3.0, following SD3 [Esser et al., 2024] and Flux [Labs et al., 2025]). We use the Muon [Liu et al., 2025] optimizer with a fixed learning rate of 1e-5 with warmup. We train our model on 32 A100 GPUs for a week, for 20k iterations. Empirically, the AdamW [Kingma and Ba, 2015] optimizer required significant noise schedule shift toward higher timesteps ($t > 900$), which

Table 3: Ablations for trajectory conditioning (left) and geospatial conditioning (right).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Plucker Rays	13.39	0.432	0.503	15.51
Concat MLP Output	13.31	0.422	0.516	14.82
Residual Add (Ours)	**16.40**	**0.486**	**0.485**	**14.50**

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
No Geospatial	14.13	0.479	0.505	**14.04**
VGGT + CrossAttn	14.11	0.479	0.502	14.07
RGB + ControlNet	16.01	0.484	0.499	15.28
RGB + CrossAttn (Ours)	**16.40**	**0.486**	**0.485**	14.50

led to a degradation in output visual quality. We also implement FSDP and gradient checkpointing to reduce vram usage.

D Architecture Ablations

Due to computation constraints, we did not run all design ablations on the entire training set. Here, we provide quantitative results when training and testing only on one city (Honolulu), with the testing set streets of a few unseen neighborhoods, same as the setup described in the main paper.

Trajectory conditioning. We also experimented with other conditioning methods: 1) concatenating plucker rays to the noisy target latents, which is common in novel view synthesis methods; 2) concatenating the MLP output (same process as residual addition described in the paper) to the noisy target latents by repeating the values across the spatial dimensions. We empirically found that our method of residual addition worked best. Metrics on Honolulu are provided in table 3. The first image and geospatial conditions are the same as those described in the main paper.

Geospatial conditioning. We explored alternative methods for formatting the conditioning video, such as directly using the raw panoramas. While this alternative should provide a complete 360-degree view of the scene and thus more context for conditioning, we found that the model struggled to follow the context, likely because this data was rare or absent in the Wan base training. Even with extensive training, the model showed no signs of rendering buildings and roads outside the first image's FOV.

Before using RGB images as conditions, we also tried VGGT [Wang et al., 2025] features, which has, in many applications, demonstrated 3D awareness and semantic understanding of static and dynamic elements. However, our model showed no signs of integrating this information, despite various adapter layers, prolonged training, or even during small-scale, overfitting experiments. This observation is supported by concurrent work such as Gen3R [Huang et al., 2026], which only successfully injected VGGT features into Wan after training an adapter to recast the features into the distribution of the VAE latents via KL and reconstruction losses.

In short, we found that the Wan base model more effectively integrates information that originate from its own temporal VAE. Shifts in the latent distribution hinder the learning process. For instance, in an ablation where condition frames were processed individually through the VAE (treating them as independent image latents rather than a video latent) the model required three times as many iterations to converge.

Finally, we conducted experiments with other conditioning mechanisms such as ControlNet [Zhang et al., 2023], and found that our approach yielded the best adherence to the conditions. Metrics on Honolulu are provided in table 3. The first image and trajectory conditions are the same as those described in the main paper. Using VGGT + CrossAttn effectively kept the model a trajectory-conditioned I2V model, with no additional context learned by the base model. Note that with ControlNet, we set the first 10 layers as the encoder from which we cloned weights, and needed to continue to finetune the remaining 30 (decoder) layers. The main potential benefit of this architecture was the skip connections, though our final proposed method of cross attention worked best.

E User Study Details

As mentioned in the main text, we conduct a user study to evaluate the capabilities and limitations of each method. We set up three questions with 10 samples each, randomly sampled from our evaluation set. In total, we collected responses from 20 users. We provide screenshots of the interface in fig. 9 The questions are as follows:

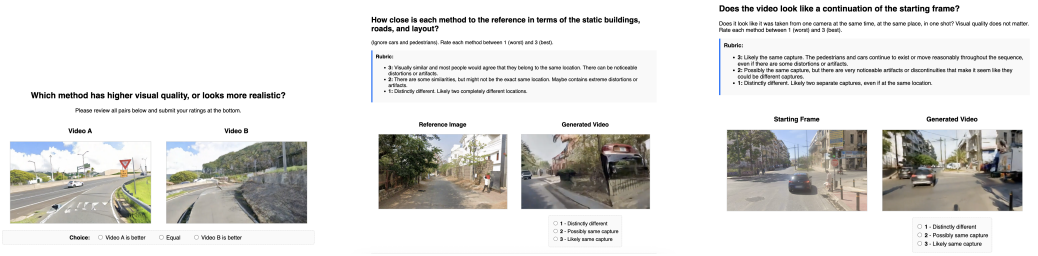

Figure 9: Three types of questions in our user study.

Q1: "Which method has higher visual quality, or looks more realistic?" We provide two videos, A and B, and three choices: "Video A is better," "Equal," "Video B is better," and we conduct a head-to-head between our method and all baselines.

Q2: "Does the video look like a continuation of the starting frame? Does it look like it was taken from one camera at the same time, at one place, in one shot? Visual quality does not matter. Rate each method between 1 (worst) and 3 (best)." We provide a starting frame, which is the first image condition explained in the main text. Then, we provide a generated video from a random method, and ask users to rate based on the following rubric.

3: Likely the same capture. The pedestrians and cars continue to exist or move reasonably throughout the sequence, even if there are some distortions or artifacts.

2: Possibly the same capture, but there are very noticeable artifacts or discontinuities that make it seem like they could be different captures.

1: Distinctly different. Likely two separate captures, even if at the same location.

Q3: "How close is each method to the reference in terms of the static buildings, roads, and layout? Ignore cars and pedestrians. Rate each method between 1 (worst) and 3 (best)." We provide a reference image, which is the last image of the geospatial conditions and meant to signify the desired destination of the generated video. Then, we provide a generated video from a random method, and ask users to rate based on the following rubric.

3: Visually similar and most people would agree that they belong to the same location. There can be noticeable distortions or artifacts.

2: There are some similarities, but might not be the same location. Maybe contains distortions or artifacts.

1: Distinctly different. Likely two completely different locations.

F Extended Related Works

We further provide additional related work that is relevant but not central to the main paper.

F.1 Driving simulations.

Although our work is not aimed to specifically address driving simulations, our training and evaluation domain is closely related. However, to the best of our knowledge, existing works have a different focus from ours. The simulations either look synthetic [Chen et al., 2026, Chigot et al., 2025, Zhou et al., 2024], or cannot handle transfer of style, weather, and dynamic objects at once [Yang et al., 2025, Ljungbergh et al., 2025, Deng et al., 2024].

F.2 Large-Scale Novel View Synthesis and Reconstruction

Novel view synthesis and reconstruction at city-scale are relevant to our task. Early city-scale reconstruction were based on structure-from-motion (SfM) [Schönberger and Frahm, 2016, Snavely et al., 2006]. For instance, Building Rome in a Day [Agarwal et al., 2009] handled 100K+ images via

scalable SfM pipelines and cluster computing. These works established the foundation of large-scale reconstruction but focused on geometry rather than rendering.

Block-NeRF [Tancik et al., 2022] scaled NeRFs [Mildenhall et al., 2021] to the city level by spatially decomposing scenes into many per-block NeRFs with appearance embeddings, pose refinement, and exposure alignment. Mega-NeRF [Turki et al., 2022] similarly partitions large outdoor regions into spatial cells and trains sub-modules with geometry-aware sampling to enable interactive flythroughs over areas orders of magnitude larger than single-scene NeRFs. Grendel-GS [Zhao et al., 2024] distributes tens of millions of Gaussians [Kerbl et al., 2023] across GPUs to represent large scenes. However, these methods do not support dynamic motion and also require dense data with the same appearance because they use a deterministic rendering loss, which is difficult to obtain at scale.

G Ethics and Privacy

As CityRAG aims to generate realistic videos of our world and is trained on a large corpora of Street View data, which in itself presents significant privacy and ethical challenges, it introduces unique challenges.

G.1 Privacy and Anonymization

All of our data, prior to collection from the Street View database, were rigorously cleaned for identifiable information. All license plates and faces were blurred. Buildings and streets were blurred on request. No authors of this paper had access to the raw imagery.

Additionally, we heavily mitigated the appearance of people in the presentation of our results. We used tools such as Nano Banana to replace people in the condition images (both the first image and geospatial conditions) for synthetic ones, where applicable. We will also mask all people via a segmentation model when we show geospatial videos for the public release.

We acknowledge these steps still cannot remove sensitive information 100%, so we will closely monitor any request to remove videos and results after release.

G.2 Bias in Data Distribution

Although we collected data from 10 cities, across 4 continents, the majority of the data is located in Western countries. This could introduce representation bias. Though CityRAG is a research paper without direct use in products or applications, in the future, any follow up work should attempt to mitigate this bias via more diverse data collection or algorithmic corrections.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a new task, design our method and data to address this task, and evaluate it thoroughly in the experimental sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We propose avenues for future work in the paper, and address multiple limitations in Appendix A.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not provide theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: We believe the core architecture and method is relatively straightforward, and we attempt to provide details for reproduction in the method sections of the paper. We also provide ablations for our architectural designs in Appendix D. However, due to the

sensitivity of the data, we cannot release the data. We believe that one would be able to reproduce our paper if given access to similar forms of street-view imagery.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to the sensitivity of the data, we cannot release it. However, we are happy to release the training code upon acceptance of the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain our data splits in the experiment sections and our training hyperparameters and optimizer in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We do not provide error bars, but we show evaluation on a large set of scenes and images, provide appropriate metrics, and conduct a user study.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: In Appendix B and C, we describe the latency, inference costs, and training costs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper respects the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential impacts and the authors' attempts to mitigate them in Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The data will not be released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are credited and the terms of use are respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We are releasing many videos to support our claims and experiments in the paper. They are hosted on an anonymous website.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the complete rubric and samples interfaces given to human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: Although we conducted a user study, we closely adhered to the NeurIPS Code of Ethics and believe that there were no potential risks incurred by participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used in any meaningful way to shape this paper or research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.