CVSP 겨울 세미나

Prompt-controllable Camera Parameters in Image manipulation

Pusan National University
Dept. of Electronics Engineering
Computer Vision & Signal Processing Lab.

Presented By
박사과정 정성훈

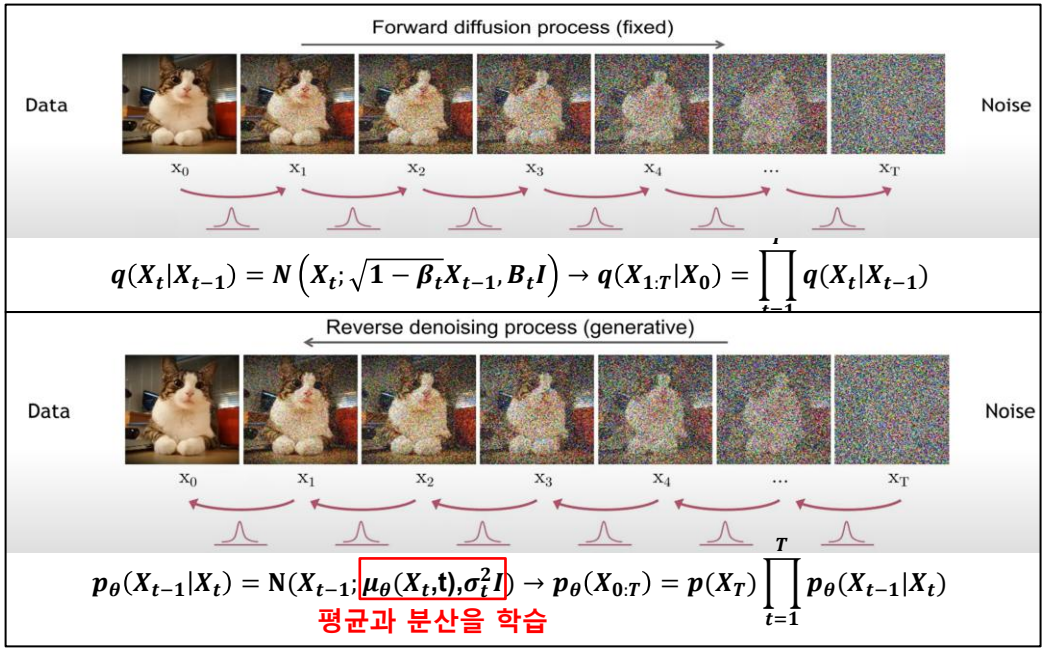
목차

- Preliminary
- 개요
  - Prompt-controllable Camera Parameters in Image manipulation
- Generative Photography: Scene-Consistent Camera Control for Realistic Text-to-Image Synthesis **(CVPR 2025 Highlight)**
- Thinking with Camera: A Unified Multimodal Model for Camera-Centric Understanding and Generation **(ICLR submitted, Rate: 6,8,8,6)**

Preliminary

* Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

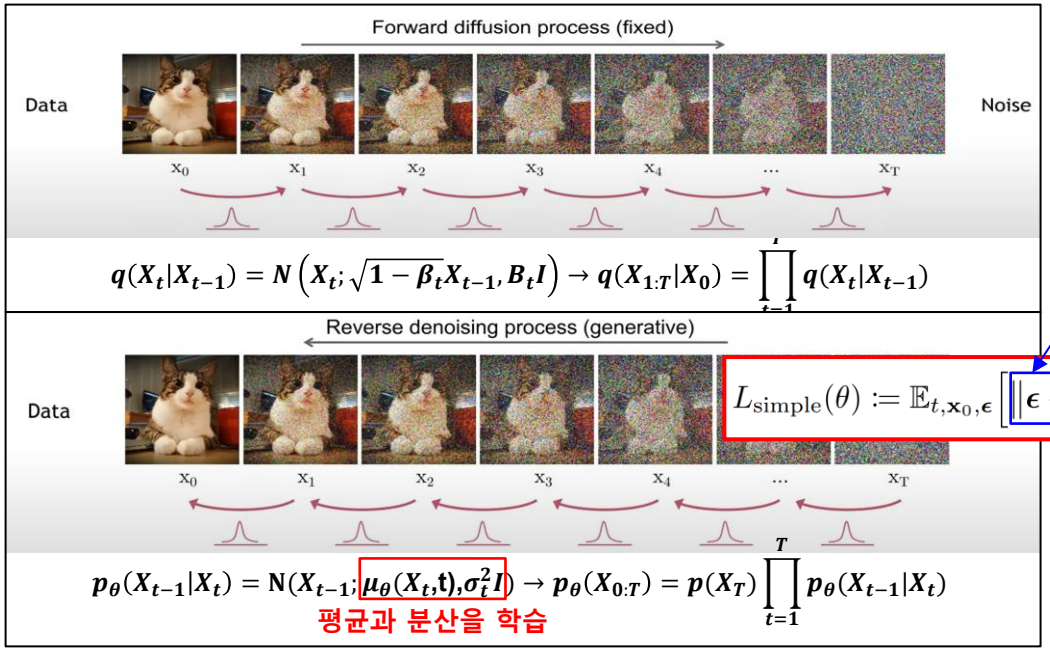
- Diffusion Models*
  - 고품질 이미지를 생성할 수 있는 모델
  - Gaussian noise를 더하는 과정과 역으로 추정하여 제거하는 과정을 학습함으로써 이미지 생성할 수 있음.
  - 생성되는 이미지 계속적으로 변경
    - 실제 noise 생성 시 Gaussian distribution에서 항상 다른 값이 샘플링 됨



Preliminary

* Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

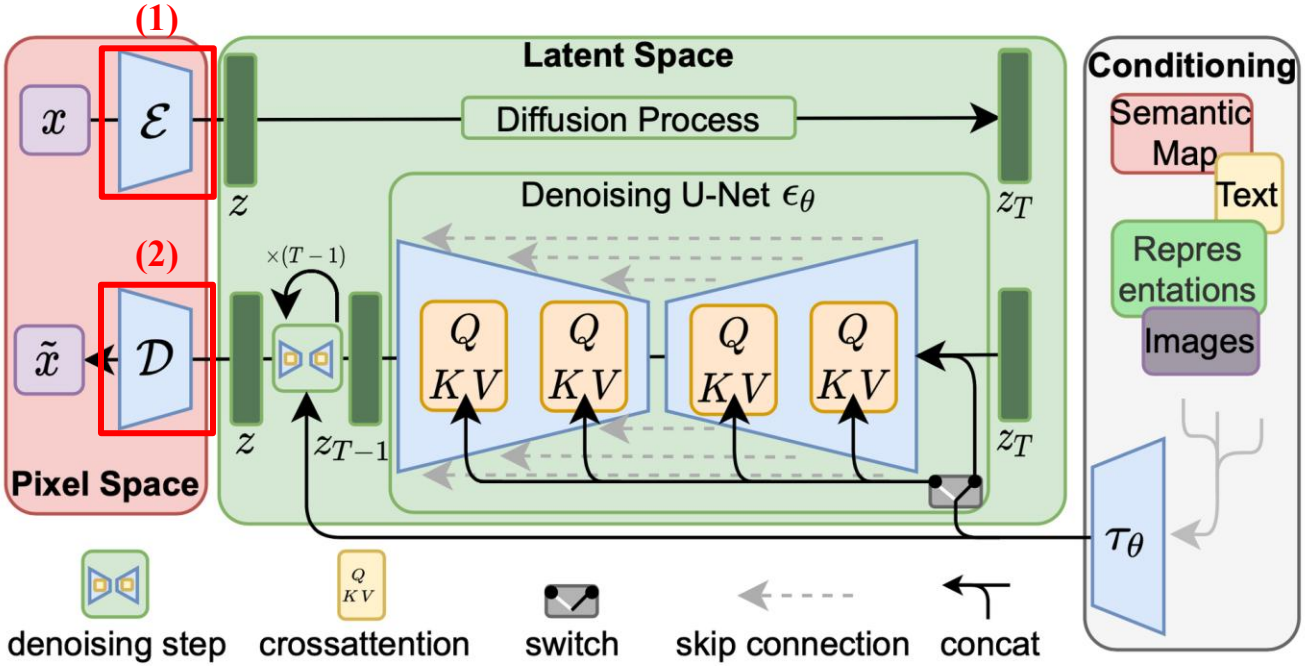
- Diffusion Models*
  - 고품질 이미지를 생성할 수 있는 모델
  - Gaussian noise를 더하는 과정과 역으로 추정하여 제거하는 과정을 학습함으로써 이미지 생성할 수 있음.
  - 생성되는 이미지 계속적으로 변경
    - 실제 noise 생성 시 Gaussian distribution에서 항상 다른 값이 샘플링 됨



Preliminary

* Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- Latent Diffusion Models* (Stable Diffusion)
  - 기존 Pixel 단위 Diffusion에서 AutoEncoder(1)와 Decoder(2)를 통해 Latent 단위로 변환하여 Diffusion을 진행
    - Latent space에서 feature level로 수행하여 pixel에서 보다 더 좋은 scaling property를 보이고 계산이 훨씬 덜 복잡해 이미지 생성에 더 효과적임.
  - Cross-Attention을 통해 추가적인 Conditioning 가능
    - Text, Images, Semantic Map 등

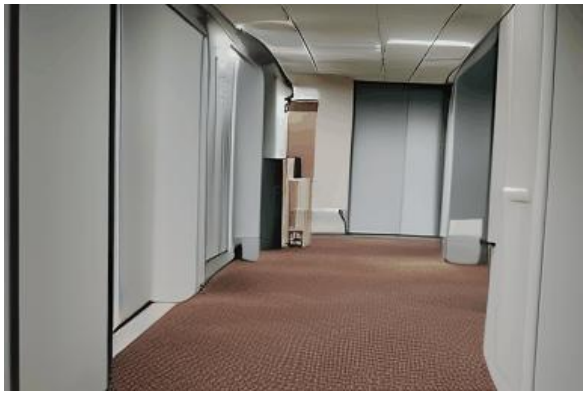


개요

- Prompt-controllable Camera Parameters in Image manipulation
  - 분야 정의
    - Text prompt를 통해 이미지 생성·편집 과정에서 카메라 파라미터를 제어하는 연구 분야



Bokeh* Rendering



Focal Length



Shutter Speed



Color Temperature

*Bokeh: 사진에서 초점이 맞지 않은 부분의 흐려짐 또는 그 흐려짐의 미학적인 양상

Generative Photography:
Scene-Consistent Camera Control
for Realistic Text-to-Image Synthesis

(CVPR 2025 Highlight)

Introduction

* Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. arXiv preprint arXiv: 2411.02385, 2024.

- Image Generation and Editing without Camera Awareness
  - 기존 방법들의 한계
    - 생성 및 편집의 quality는 상승했지만, 현실적인 카메라 효과는 재현하지 못함
      - 데이터 기반 모델과 물리 세계의 원리 사이에 존재하는 간극
      - 학습 데이터와 모델 파라미터를 증가시키는 것만으로는 생성 모델이 핵심적인 물리 원리를 포착하기에 충분하지 않음*



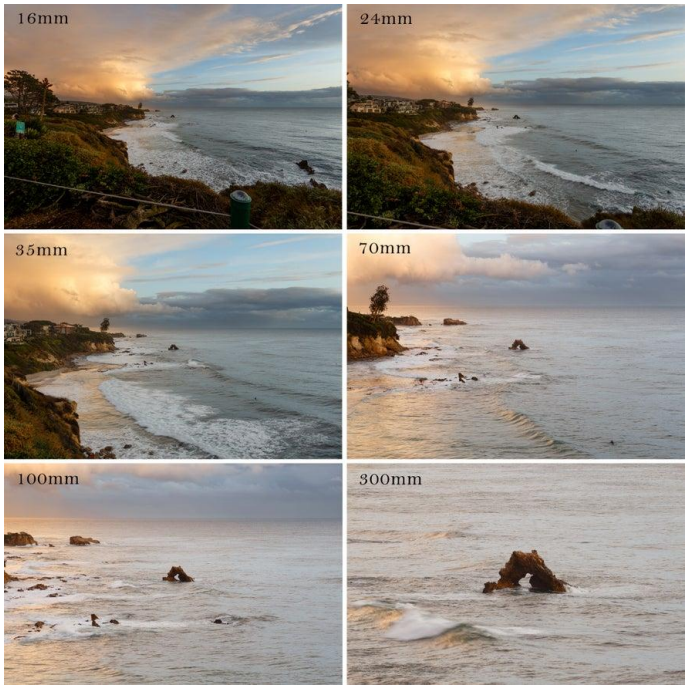
"... desserts;
with bokeh blur parameter [2, 6, 10, 14, 18]."

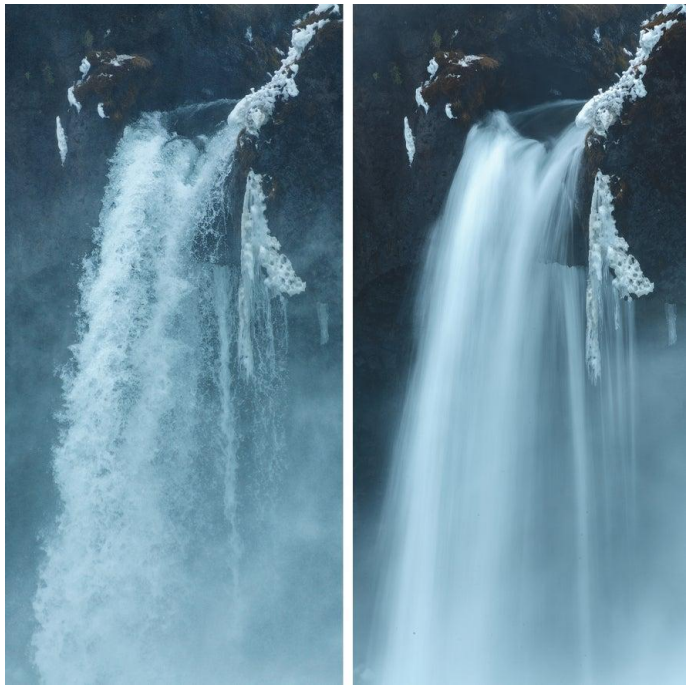


"... living room;
with temperature [3200, 3050, 3700, 4500, 9500] kelvin."

Introduction

- Why do Existing Generative Models Fail to Understand Camera Physics?
  - 1) 사용 가능한 학습 데이터가 매우 제한적임
    - 카메라 정보를 포함하여 학습 시키기 위해서는 **‘서로 다른 카메라 설정’**을 가진 **‘동일한 scene 이미지’** set 필요
      - ⁛ 현실적으로 데이터 셋을 구축하기 어려움





Introduction

- Why do Existing Generative Models Fail to Understand Camera Physics?
  - 2) Scene embedding과 camera embedding의 disentangle의 어려움
    - 하나의 이미지에서 어떻게 **‘scene에 대한 정보’**와 **‘camera에 대한 정보’**를 분리해낼 수 있는지에 대한 어려움



Scene Info
- 강 또는 호수 주변의 자연 풍경, 물 위에 반사된 하늘색
- 따뜻한 오렌지·노랑 계열 하늘, 차가운 어두운 실루엣 대비, 고요하고 정적인 분위기

Camera Info
- 카메라 바디: Sony A7R III
- 렌즈: Sony FE 24–105mm f/4.0 G OSS
- 조리개 (Aperture): f/4.0
- 셔터 스피드 (Shutter Speed): 30 sec
- ISO: 1600
- 초점거리 (Focal Length): 31 mm
- 노출 보정 (Exposure Compensation): −0.3 EV
- …

Introduction

- What is Generative Photography?
  - 정의
    - 물리적으로 사진을 촬영하는 대신,
      조리개·셔터 속도·초점거리 등 **카메라 내부 파라미터를 인식하고 제어함으로써**,
      장면의 내용 일관성을 유지한 채 사실적인 사진 효과를 생성하는
      카메라 인식 기반 이미지 생성 패러다임

  - 목표

    1) **카메라 효과의 물리적 사실성**
       - 조리개, 셔터 속도, 초점거리, 색온도 등 카메라 설정에 따른 효과가 실제 촬영과 일관되게 **현실적으로 렌더링**되어야함

    2) **카메라 설정 변화에도 장면 내용 일관성 유지**
       - 카메라 내부 파라미터가 변경되더라도 **장면의 구조와 정체성**은 변하지 않아야 함

    3) **기준 모델 대비 이미지 품질 유지**
       - 카메라 인식 기능을 추가하더라도, 이를 고려하지 않는 기존 **camera-agnostic 생성 모델과 비교해 이미지 품질이 저하되지 않아야 함**

Introduction

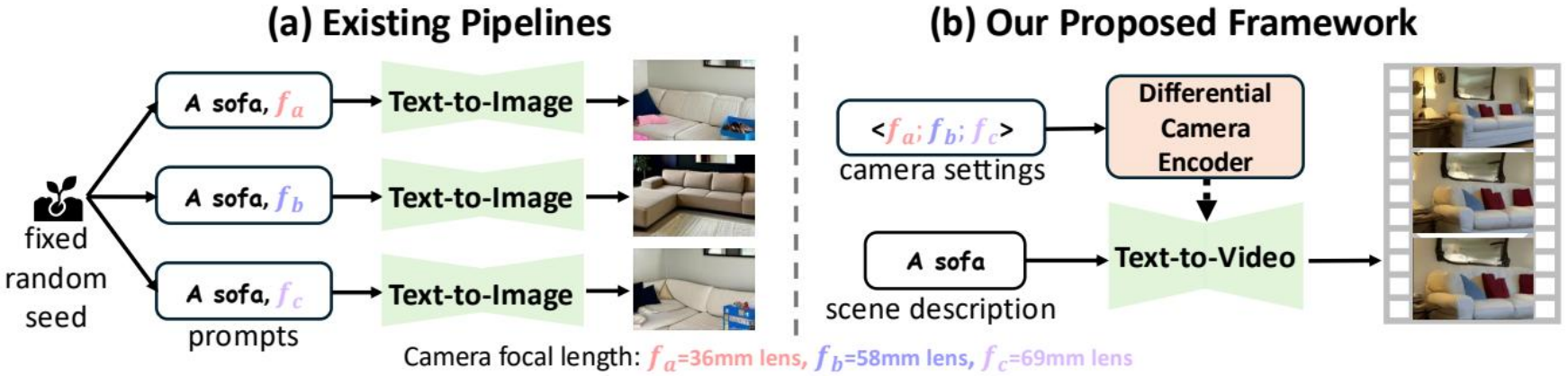
- Proposed Method
  - Dimensionality Lifting
    - 카메라 파라미터를 장면과 분리된 독립 차원으로 모델링
    - Text-to-Image

$$I = f(\text{text}, \mathbf{z})$$

    - Dimensionality Lifting (Camera parameter)

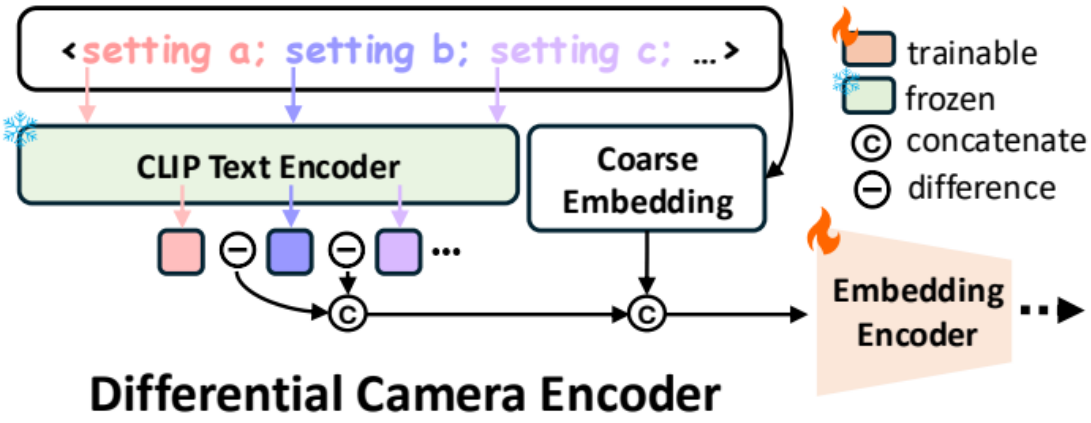
$$I = f(\text{text}, \mathbf{z}, \mathbf{c})$$

      - $\mathbf{c}$: camera intrinsics (focal length, aperture, shutter, …)



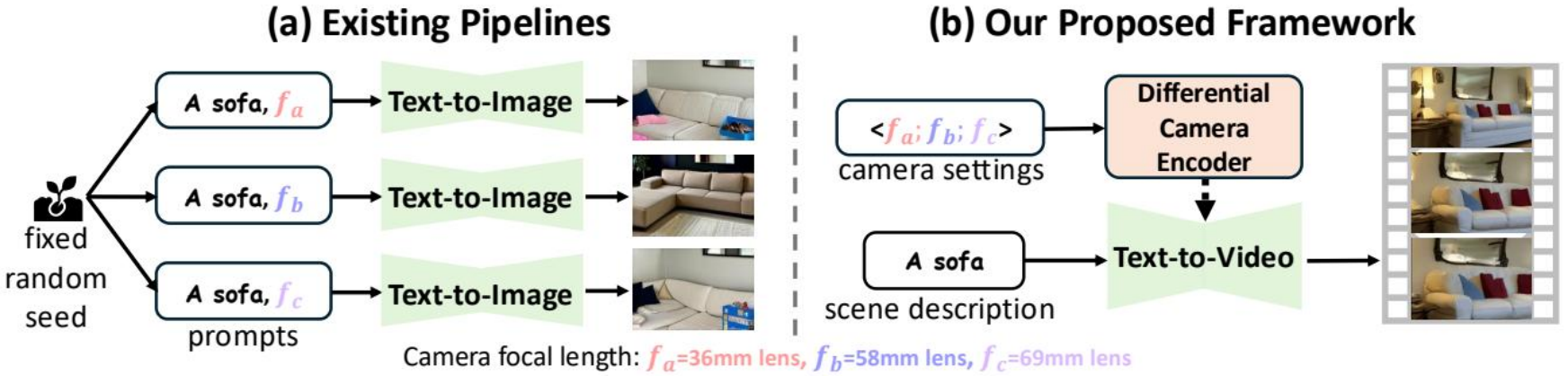
Introduction

- Proposed Method
  - Differential Camera Intrinsics Learning
    - 동일한 장면에서 서로 다른 카메라 설정 간의 차이를 명시적으로 학습하도록 differential data, differential camera encode를 활용
      - Temporal consistency를 유지하면서 camera parameter를 control 할 수 있음



Method

- How to disentangle the camera embedding from the scene embedding
  - Dimensionality Lifting
    - 카메라 정보가 scene embedding에 섞이면 카메라를 바꿀 때마다 장면 자체가 흔들림
      - Scene consistency 붕괴
    - 기존 T2I는 camera parameter를 **prompt의 일부**로만 취급
      - 결과적으로 **일관성 없는 이미지 생성**



Method

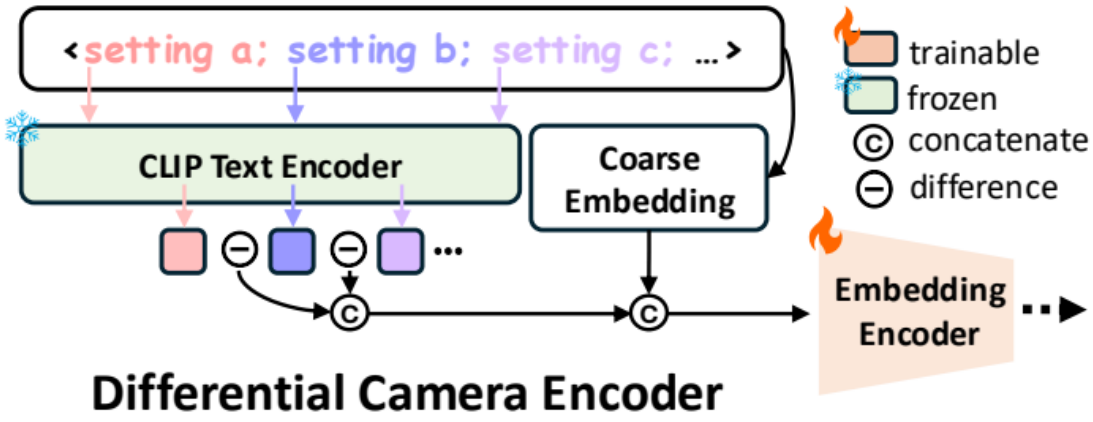
- How to disentangle the camera embedding from the scene embedding
  - Dimensionality Lifting
    - 문제 정의를 2D spatial domain에서 special + camera dimension으로 확장
    - 다중 카메라 설정을 갖는 이미지 생성을 위해 비디오 생성 모델 활용
      - 비디오 생성 모델은 time이라는 축을 가지고 있기 때문에 이러한 축을 camera dimension으로 대체
      - Spatial-temporal attention을 통해 frame consistency 유지에 유리
    - in T2V model
      - Scene description: 변하지 않음 (모든 프레임 공유)
      - Camera parameters: 프레임별로 변화
      - **"같은 장면을 서로 다른 카메라로 본 프레임들의 집합" 생성**



Prompt:
A display of frozen desserts, including cupcakes and donuts, is arranged in a row on a counter.
The desserts are placed in plastic containers, and there are several of them in various sizes and flavors;
with bokeh blur parameter **

Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Intrinsics Learning
    - Generative photography의 또 다른 목표는 모델이 카메라 내부 파라미터의 '물리적 의미'를 실제로 이해하도록 만드는 것
      - ⁂ 단순히 카메라 값을 condition으로 주는 것만으로는 물리적으로 타당한 효과 학습이 어려움
    - 카메라 설정을 절대값이 아닌 '차이(difference)' 중심으로 학습



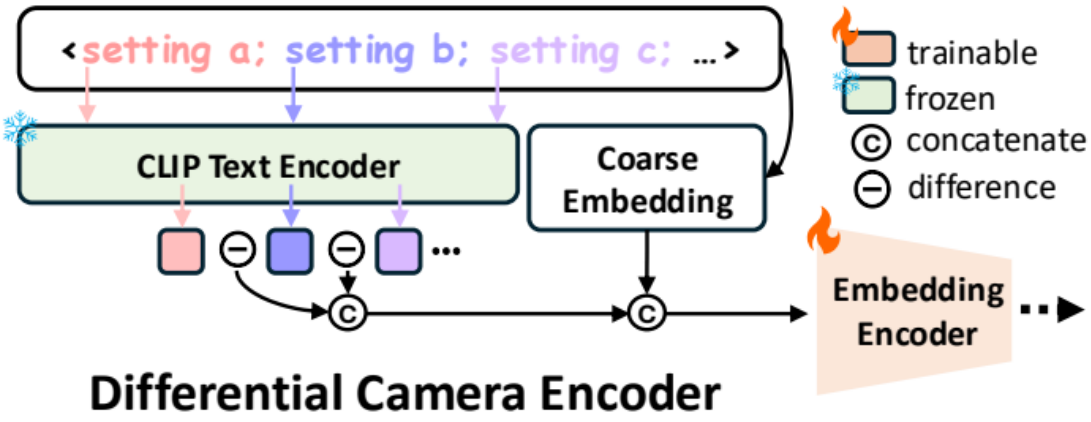
Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Intrinsics Learning
    - Differential Data

      → 무엇을 학습시킬 것인가

    - Differential Camera Encoder

      → 어떻게 학습시킬 것인가



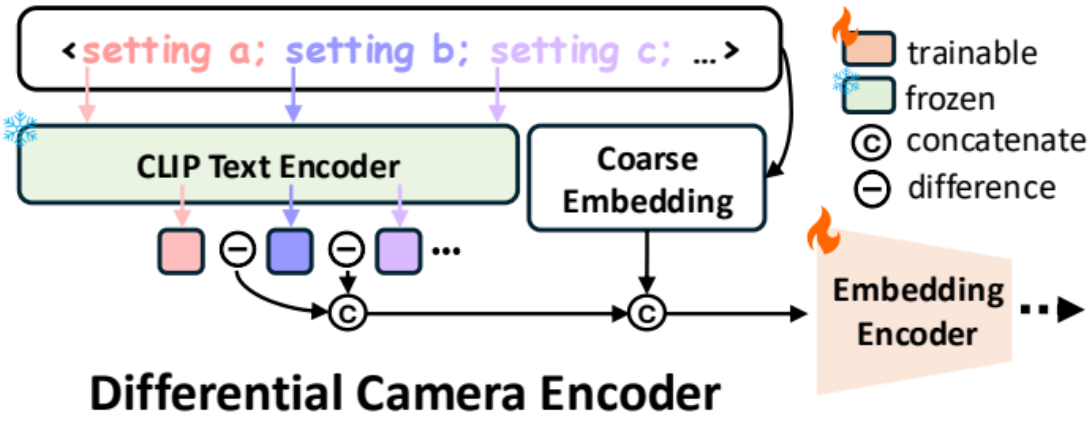
Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Intrinsics Learning
    - Differential Data

      → 무엇을 학습시킬 것인가

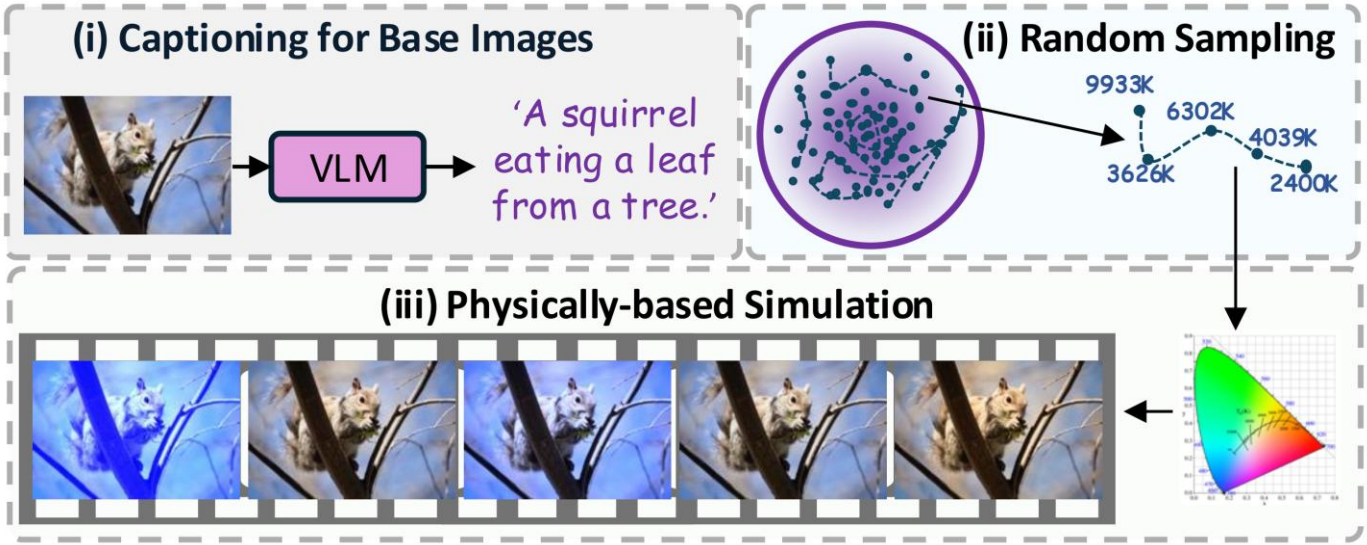
    - Differential Camera Encoder

      → 어떻게 학습시킬 것인가



Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Data
    - 장면은 고정, **카메라 설정만 변화**시키는 이미지 데이터 셋
      - 실제 데이터의 방대한 다양성 대신 카메라 변화로 인한 시각적 차이에만 집중시키도록 모델을 학습시키기 위해 구축
    - 차분 데이터셋은 **3단계 파이프라인**으로 구성



Method

** Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.

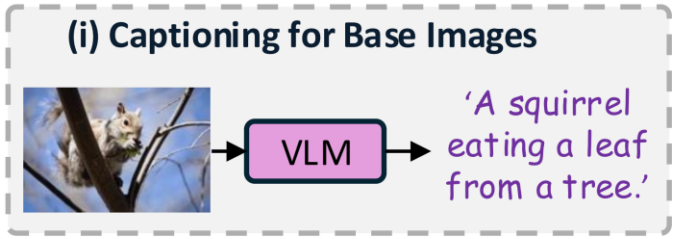
- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 1) Vision Language Model Captioning for Base Images
    - 기본 이미지 조건
      - 고품질 이미지
      - 카메라 효과별 요구사항 충족 필요 ⇒
        - Shutter speed → Wide Dynamic Range*, appropriate exposure
          * 한 이미지 안에서 아주 어두운 영역부터 아주 밝은 영역까지 동시에 표현할 수 있는 능력
        - Color temperature → neutral color temperature
        - Bokeh → all-in-focus
        - Focal length → sufficiently high resolution.
    - VLM LLaVA**를 사용해 기본 이미지 caption 생성
      - 해당 세트의 **고정 scene description**으로 사용



for Shutter speed for Color temperature

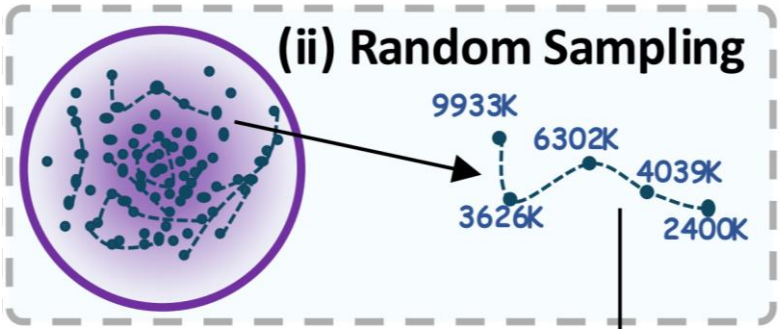
for Bokeh for Focal length

Base Images



Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 2) Random Sampling on Continuous Camera Space
    - 카메라 파라미터는 연속 변수(continuous variables)
      - 몇몇 개의 이산 값이 아니라 **물리적 스케일 전체**를 학습해야 함
    - 연속 카메라 스케일에서 **무작위 샘플링**
      - Ex) 색온도 = {2400K, 4039K, 3626K, 6302K, 9933K}
      - 순서가 아닌 상대적 차이를 학습 시키기 위해 sort 하지 않음



Method

* Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
*** Peng, Juewen, et al. "Bokehme: When neural rendering meets classical rendering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 3) Physically-based Simulation
    - 무작위로 샘플링한 카메라 설정을 실제 카메라 물리에 근거한 방식으로 base image에 적용

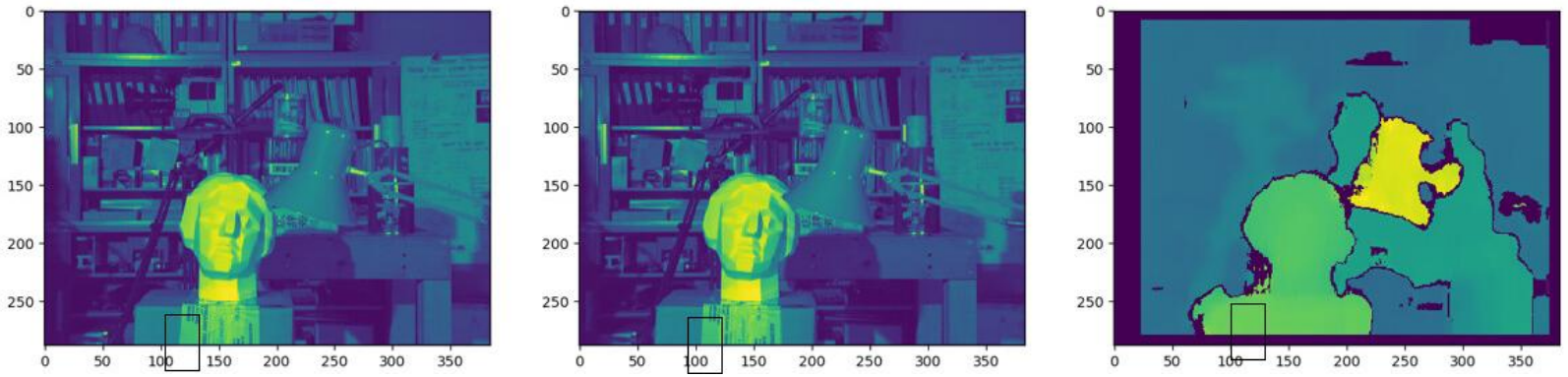
1) Bokeh rendering

Depth Anything*을 사용하여 base image의 depth map을 추출, 이를 disparity map**으로 변환

이미지의 각 픽셀이 **카메라에 얼마나 가까운지를 상대적으로 표현, 일반적으로 **depth의 역수**로 정의

Disparity map은 base image 및 샘플링된 blur parameter와 함께 bokeh rendering module***의 입력으로 활용

✓ 이 과정에서 refocused 시차는 전경 깊이에 고정되어, 전경 객체는 선명하게 유지되고 배경 blur 만 조정



Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 3) Physically-based Simulation

    2) Focal length

    ⁘ 고해상도의 base image가 주어졌을 때, 원래의 초점거리에 비해 샘플링된 초점거리에 해당하는 시야각 비율 계산

    ✓ 이 비율에 따라 기본 이미지에서 해당 영역을 center crop

    ✓ 서로 다른 초점거리에서 획득된 모든 이미지를 동일한 해상도로 resize



Method

** Li, Zhihao, et al. "Efficient visual computing with camera raw snapshots." IEEE Transactions on Pattern Analysis and Machine Intelligence 46.7 (2024): 4684-4701.

- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 3) Physically-based Simulation

3) Shutter speed

- 촬영된 이미지의 조사량*(irradiance)을 여러 shutter speed로 변경

  * 일정 시간 동안 가해지는 빛

- 이를 simulated image signal processor (ISP)**를 통해 shutter speed가 변경된 RGB 이미지로 변환함으로써, 서로 다른 노출 수준의 이미지를 얻음



Base Images

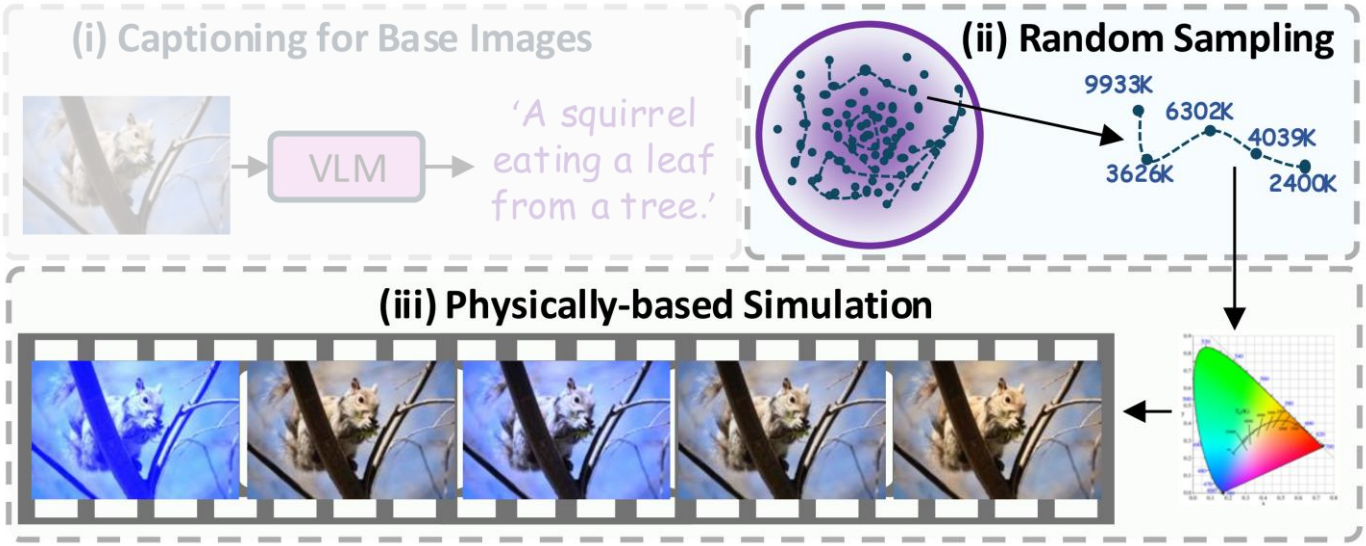
Method

* Fairchild, Mark D. Color appearance models. John Wiley & Sons, 2013.

- How to disentangle the camera embedding from the scene embedding
  - Differential Data – 3) Physically-based Simulation

    4) Color temperature

    - Black body radiation과 color temperature 간의 관계를 empirical approximation*
    - 다양한 색온도에 맞게 RGB 채널 비율을 조정



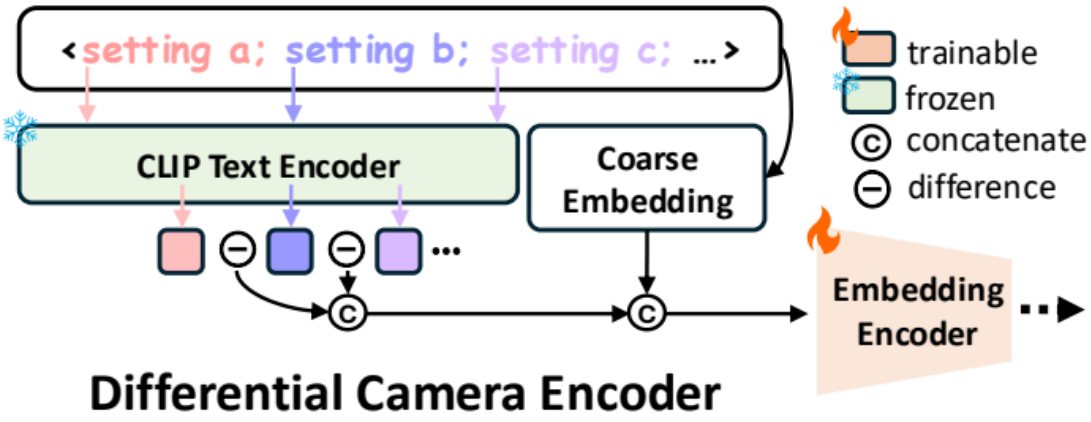
Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Intrinsics Learning
    - Differential Data

      → 무엇을 학습시킬 것인가

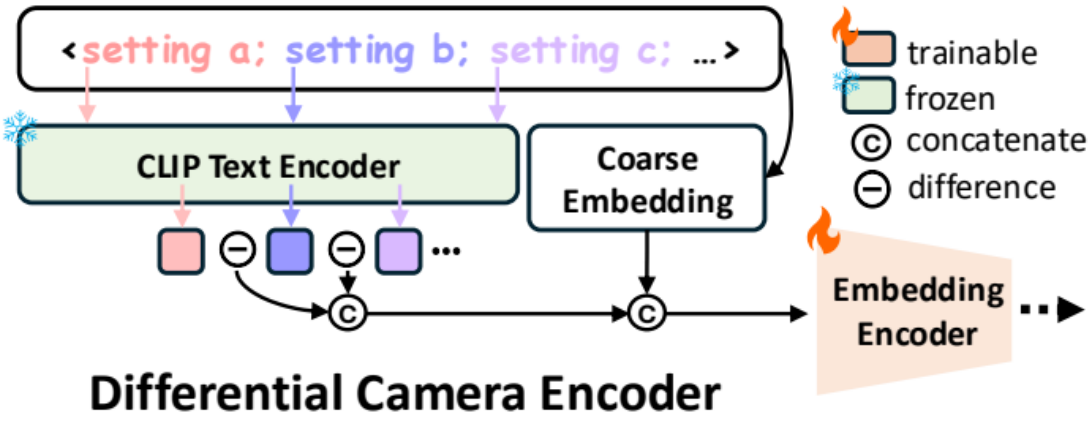
    - Differential Camera Encoder

      → 어떻게 학습시킬 것인가



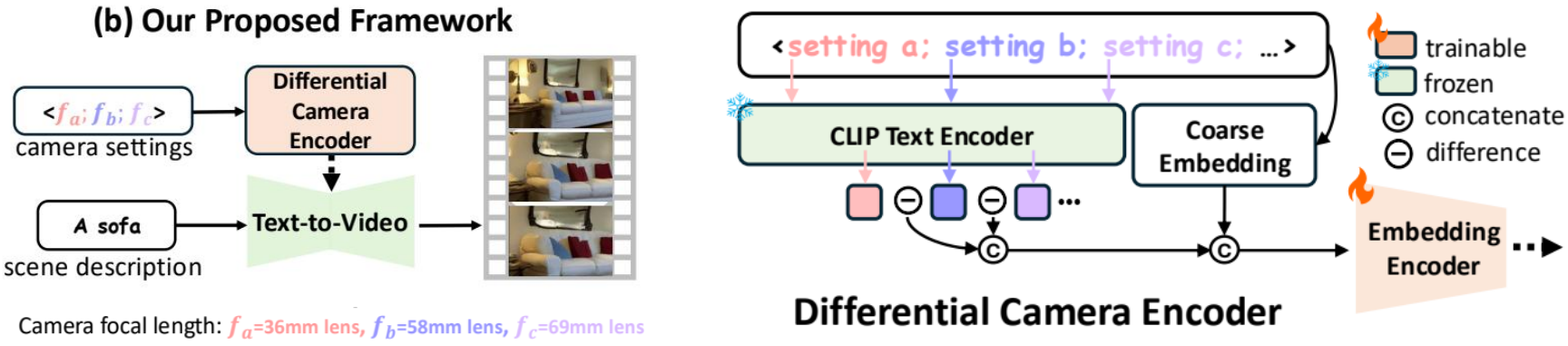
Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Intrinsics Learning
    - Differential Data
      - → 무엇을 학습시킬 것인가
    - Differential Camera Encoder
      - → 어떻게 학습시킬 것인가



Method

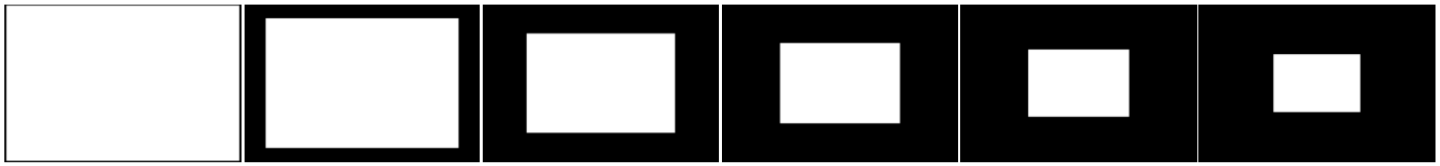
- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Encoder
    - Camera settings을 embedding하고 추가적인 differential 정보를 제공함으로써, 모델이 camera setting을 이해하고 생성하는 능력을 강화
      - 카메라 설정을 "값 자체"가 아니라, "프레임 간 차이와 비교 관계"로 이해하도록 모델을 유도
    - 두 종류의 embedding을 결합하여 입력으로 사용
      - Coarse embedding
      - Differential camera embedding



Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Encoder
    - Coarse embedding
      - '물리적 방향'을 네트워크에 알려주는 공간적 힌트 $F \times C \times H \times W$
        - ✓ F (Frames): 각 카메라 설정에 대응하는 프레임
        - ✓ C (Channels): RGB 또는 다중 물리 채널
        - ✓ H×W: 공간적으로 정렬된 픽셀 맵

Camera Parameter	Coarse Embedding	Meaning
Bokeh	Blur Parameter	• 파라미터 ↑ → 가우시안 커널 σ ↑ → 각 픽셀의 가중치 ↓
Focal Length	Mask	• FoV 비율 계산 후 포함되지 않는 영역을 마스킹
Shutter Speed	Shutter Time (s)	• 기준 셔터(≈0.2s) 대비 비율 계산 • 전체 embedding에 곱해지는 전역 계수
Color Temperature	Color Temperature (k)	• Black body radiation approximation 단순화

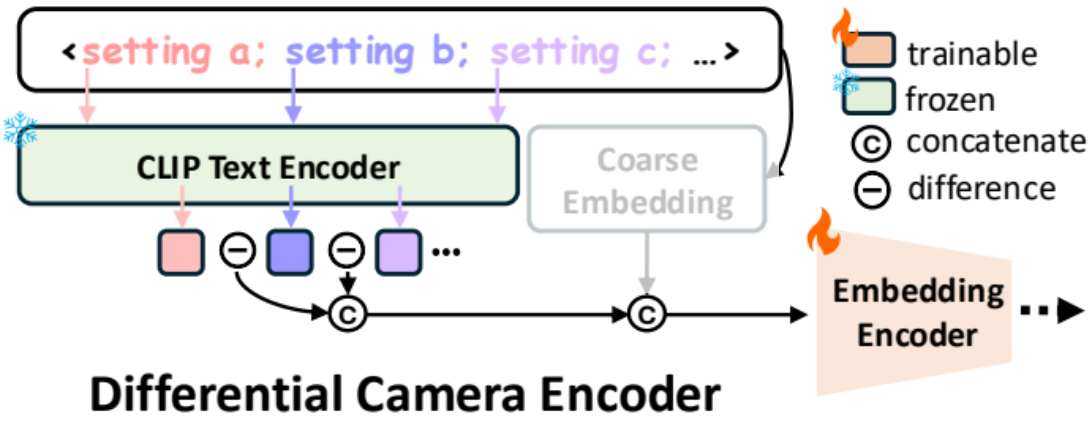


Focal length에 적용되는 coarse embedding

부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

Method

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Encoder
    - Differential camera embedding
      - 프레임 간 카메라 설정의 **의미적 차이**
      - 고정된 CLIP text encoder 사용
      - 각 프레임의 카메라 설정 → text embedding
      - 인접 프레임 간 embedding 차이(difference) 계산

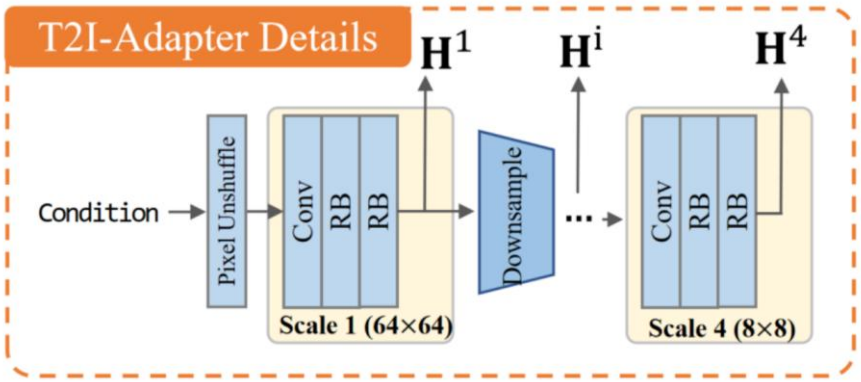


Method

* Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models." Proceedings of the AAAI conference on artificial intelligence. Vol. 38. No. 5. 2024.

- How to disentangle the camera embedding from the scene embedding
  - Differential Camera Encoder
    - T2I-Adapter*의 encoder 구조 차용
      - 텍스트만으로는 제공할 수 없는 공간적·구조적 gauidance를 외부 모듈로 추출
      - Pre-trained T2I diffusion 모델에 안정적으로 주입하는 lightweight adapter
    - 본 연구에서는 공간 구조 대신 camera parameter를 pre-trained T2V model에 입력으로 제공하기 위해 활용
    - 입력
      - Camera embedding: (B, C_{in}, F, H, W)
    - 출력
      - features = $[\mathbf{H}^1, \mathbf{H}^2, \mathbf{H}^3, \mathbf{H}^4]$

입력 shape
- B : batch size
- C_{in} : 카메라 임베딩 채널 수 (예: coarse + differential)
- F : 프레임 수 (T2V의 time / camera dimension)
- H, W : 공간 해상도



출력 shape
- H¹: (B, C₁, F, H₁, W₁)
- H²: (B, C₂, F, H₂, W₂)
- H³: (B, C₃, F, H₃, W₃)
- H⁴: (B, C₄, F, H₄, W₄)

Experiments

- 데이터셋
  - 전체 정보
    - 다양한 장면과 촬영 조건을 포함한 1,000장의 이미지로 구성
  - Bokeh
    - 전경 객체가 명확하고 배경과 구분되는 all-in-focus base image
      - Ershov, Egor I., A. V. Belokopytov, and A. V. Savchik. "Problems of dataset creation for light source estimation." arXiv preprint arXiv:2006.02692 (2020).
      - Hasinoff, Samuel W., et al. "Burst photography for high dynamic range and low-light imaging on mobile cameras." ACM Transactions on Graphics (ToG) 35.6 (2016): 1-12.
      - Zhang, Xuaner, et al. “Synthetic Defocus and Look-Ahead Autofocus for Casual Videography.” ACM Transactions on Graphics, vol. 38, no. 4, 2019, article 30, pp. 1–16. Association for Computing Machinery, https://doi.org/10.1145/3306346.3323015
  - Focal Length
    - 24mm 렌즈가 장착된 풀프레임 카메라로 촬영된 이미지 선정
      - Ershov, Egor I., A. V. Belokopytov, and A. V. Savchik. "Problems of dataset creation for light source estimation." arXiv preprint arXiv:2006.02692 (2020).
      - Zhang, Xuaner, et al. "Zoom to learn, learn to zoom." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
  - Exposure and white balance conditioning tasks (Shutter speed, Color temperature)
    - Ouyang, Hao, et al. "Neural camera simulators." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
    - Zhang, Xuaner, et al. "Zoom to learn, learn to zoom." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Experiments

* Guo, Yuwei, et al. “AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning.” Proceedings of the International Conference on Learning Representations (ICLR), 2024, Spotlight Presentation.

- Train
  - Baseline model
    - AnimateDiff*: T2V model
  - Number of Frame
    - 5
  - 학습 시 video resolution
    - 256 x 384
  - Batch
    - 8
  - Epoch
    - 25,000
  - GPU
    - NVIDIA A100 80 GB x 1, 10 h
- Inference
  - Setting
    - Input: 하나의 프롬프트와 함께 카메라 설정들의 집합
    - Output: 여러 프레임을 생성

Experiments

* Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.
** https://docs.opencv.org/4.x/dc/dc3/tutorial_py_matcher.html

- Metric
  - CorrCoef
    - 생성 이미지와 base image 간의 추세 변화에 대한 Pearson correlation coefficient
      - 생성된 이미지가 카메라 물리 설정을 얼마나 정확히 반영하는지 평가
      - Bokeh: Laplacian 연산자를 사용하여 프레임 별 평균 블러 수준을 계산
      - Focal Length: SIFT*를 사용하여 특징점을 검출, Brute-Force Matcher**를 이용해 인접 프레임 간 특징점 매칭을 수행
        - ✓ 매칭된 점들로부터 유사 변환 행렬을 계산하고, 해당 변환 행렬에서 스케일링 계수를 추출
      - Shutter Speed: 프레임 별 평균 밝기를 계산
      - Color Temperature: 프레임 별 평균 색상을 계산
  - LPIPS
    - 장면 일관성 평가
      - 서로 다른 카메라 설정으로 생성된 프레임들 간의 지각적 특징 거리를 프레임 단위 계산
  - CLIP Score
    - 프롬프트 충실도 평가
      - 생성 이미지와 텍스트 프롬프트 간의 유사도 점수를 계산

Experiments

* https://stability.ai/news/stable-diffusion-3
** https://github.com/black-forest-labs/flux
*** Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.
**** He, Hao, et al. "Cameractrl: Enabling camera control for text-to-video generation." arXiv preprint arXiv:2404.02101 (2024).

- Comparisons with Other Methods
  - 비교 모델
    - Stable Diffusion 3(SD3), FLUX**
      - The state-of-the-art text- to-image generation models
    - AnimateDiff***
      - Purely data-driven approach의 성능을 평가하기 위해, differential dataset으로 fine-tuned
      - 즉, 구축한 데이터 셋으로만 학습된 AnimateDiff와 제안하는 방법과 비교
    - CameraCtrl****
      - External camera settings 대신 intrinsic camera settings를 제어하도록 모델을 변형하고 구축한 데이터셋으로 학습

Experiments

- Comparisons with Other Methods
  - Qualitative Results



(a) Bokeh rendering: A variety of potted plants are displayed on a window sill; with bokeh blur parameter **



(b) Focal length: A large office space with a carpeted floor and white walls; with ** lens

Experiments

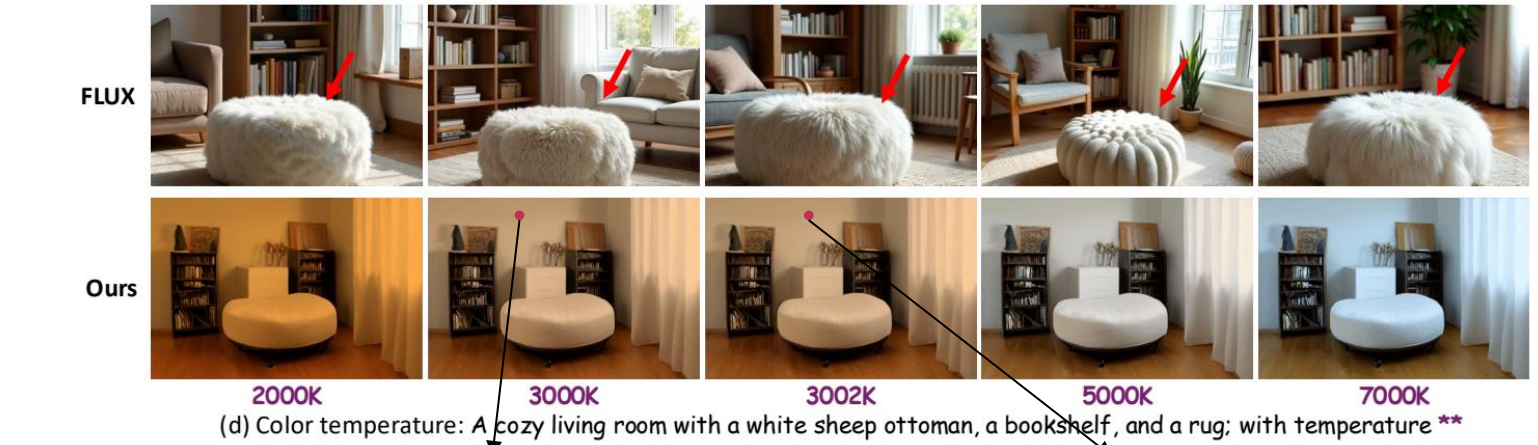
- Comparisons with Other Methods

  https://yuyuanspace.com/GenerativePhotography/

  - Qualitative Results



(c) Shutter speed: A modern bathroom with a mirror and soft lighting; with shutter speed **



(d) Color temperature: A cozy living room with a white sheep ottoman, a bookshelf, and a rug; with temperature **

Experiments

- Comparisons with Other Methods
  - Quantitative Results

Methods	Bokeh Rendering			Focal Length			Shuttle Speed			Color Temperature		
	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑
Reference	1.0000	0.0527	0.3974	1.0000	0.4709	0.3853	1.0000	0.0511	0.3783	1.0000	0.0398	0.4053
SD3 [4]	0.2492	0.7253	—	0.2356	0.7108	—	0.2731	0.6937	—	0.2312	0.6891	—
FLUX [1]	0.2006	0.6770	—	0.2003	0.6461	—	0.2398	0.6403	—	0.2363	0.6155	—
AnimateDiff [21] (w/o *FT*)	0.2960	0.1005	0.2753	0.2613	0.1208	0.2532	0.1843	0.1002	0.2631	0.1834	0.0805	0.2659
AnimateDiff [21] (w/ *FT*)	0.3714	0.0255	0.2984	0.2597	0.2288	0.2739	0.2198	0.0948	0.2936	0.2897	0.0205	0.2839
CameraCtrl [23] (w/o *FT*)	0.3303	0.1447	0.2804	0.2913	0.1144	0.2644	0.1896	0.0986	0.2912	0.1773	0.0935	0.2753
CameraCtrl [23] (w/ *FT*)	0.6025	0.1158	0.3017	0.8671	0.4606	0.2865	0.7526	0.0775	0.2981	0.5812	0.0651	0.2885
Ours	0.8626	0.0788	0.3007	0.9695	0.4647	0.2871	0.9264	0.0695	0.3015	0.8970	0.0499	0.2910

  - Ablation Study

Methods	Bokeh Rendering			Focal Length			Shuttle Speed			Color Temperature		
	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑	Accuracy CorrCoef ↑	Consistency LPIPS	Following CLIP ↑
Reference	1.0000	0.0527	0.3974	1.0000	0.4709	0.3853	1.0000	0.0511	0.3783	1.0000	0.0398	0.4053
w/o differential	0.7631	0.1239	0.2964	0.8594	0.4943	0.2885	0.8864	0.0698	0.3007	0.8586	0.0699	0.2904
Discrete	0.6774	0.0854	0.2948	0.7254	0.4173	0.2833	0.8384	0.0584	0.2976	0.8140	0.0524	0.2946
Ours (3 frames)	0.8355	0.0814	0.2895	0.9273	0.3921	0.2857	0.9073	0.0685	0.2873	0.8848	0.0835	0.2955
Ours (5 frames)	0.8626	0.0788	0.3007	0.9695	0.4647	0.2871	0.9264	0.0695	0.3015	0.8970	0.0499	0.2980
Ours (7 frames)	0.8835	0.0747	0.2988	0.9783	0.4813	0.2909	0.9294	0.0535	0.3034	0.9095	0.0581	0.2985

Thinking with Camera:
A Unified Multimodal Model for Camera-Centric Understanding and Generation

(ICLR submitted, Rate: 6,8,8,6)

Introduction

- Motivation
  - 기존 연구의 한계
    - 카메라는 공간 인식의 핵심 모달리티임에도 불구하고, 기존 모델에서는 단순한 condition 혹은 보조 입력으로 취급
      - 공간적으로 일관된 이해와 정밀한 카메라 제어 생성이 동시에 불가능

- Main Idea
  - Puffin
    - Large Language Model (LLM)의 추론 능력을 카메라 geometry로 확장
      - 카메라를 '사고(Thinking)의 대상'으로 다루는 멀티모달 프레임워크를 구축
    - Autoregressive and diffusion modeling을 통합하여 '카메라 중심 이해'와 '생성'을 수행
      - 시각적 단서 – 전문적인 사진학 용어 정렬

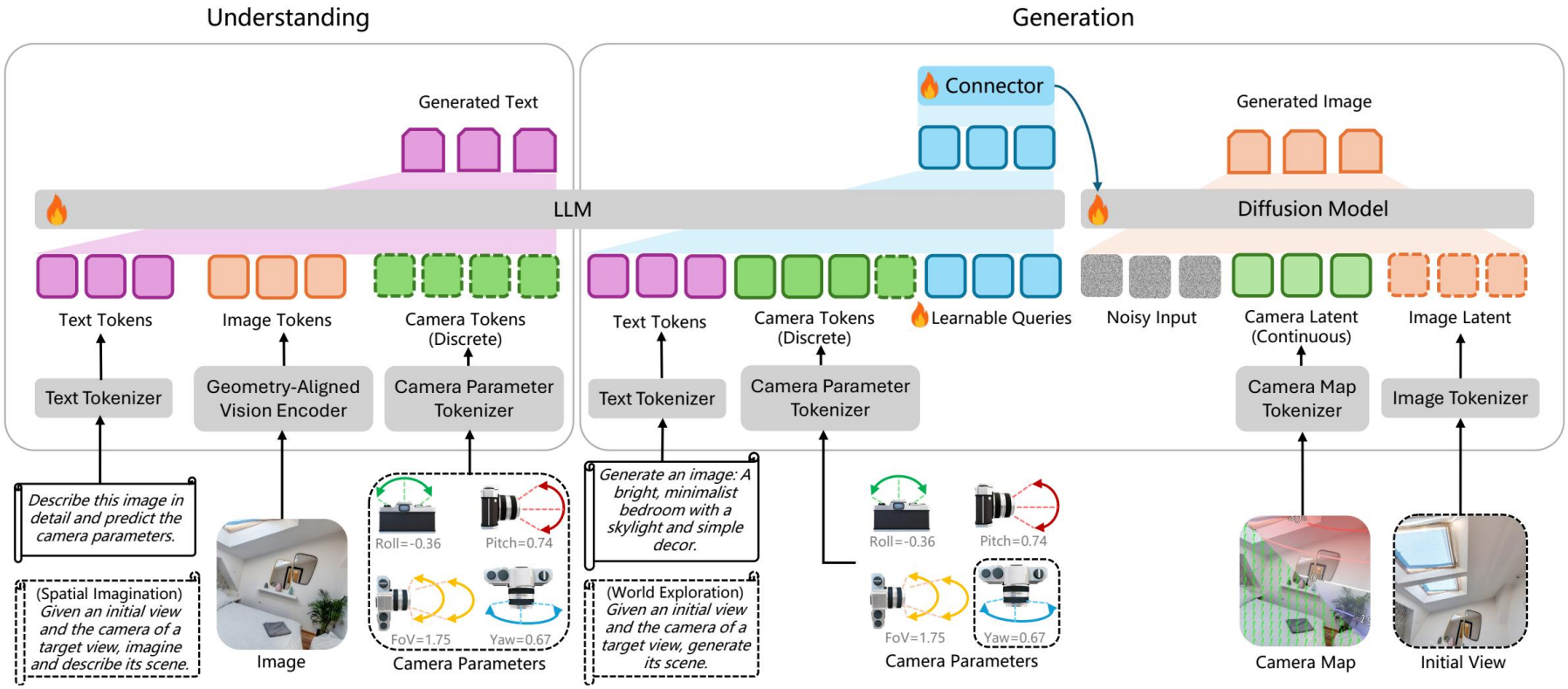
Introduction

- Motivation
  - Dataset 구축 – Puffin-4M
    - 400만 개의 vision-language-camera로 구성된 대규모 고품질 데이터셋



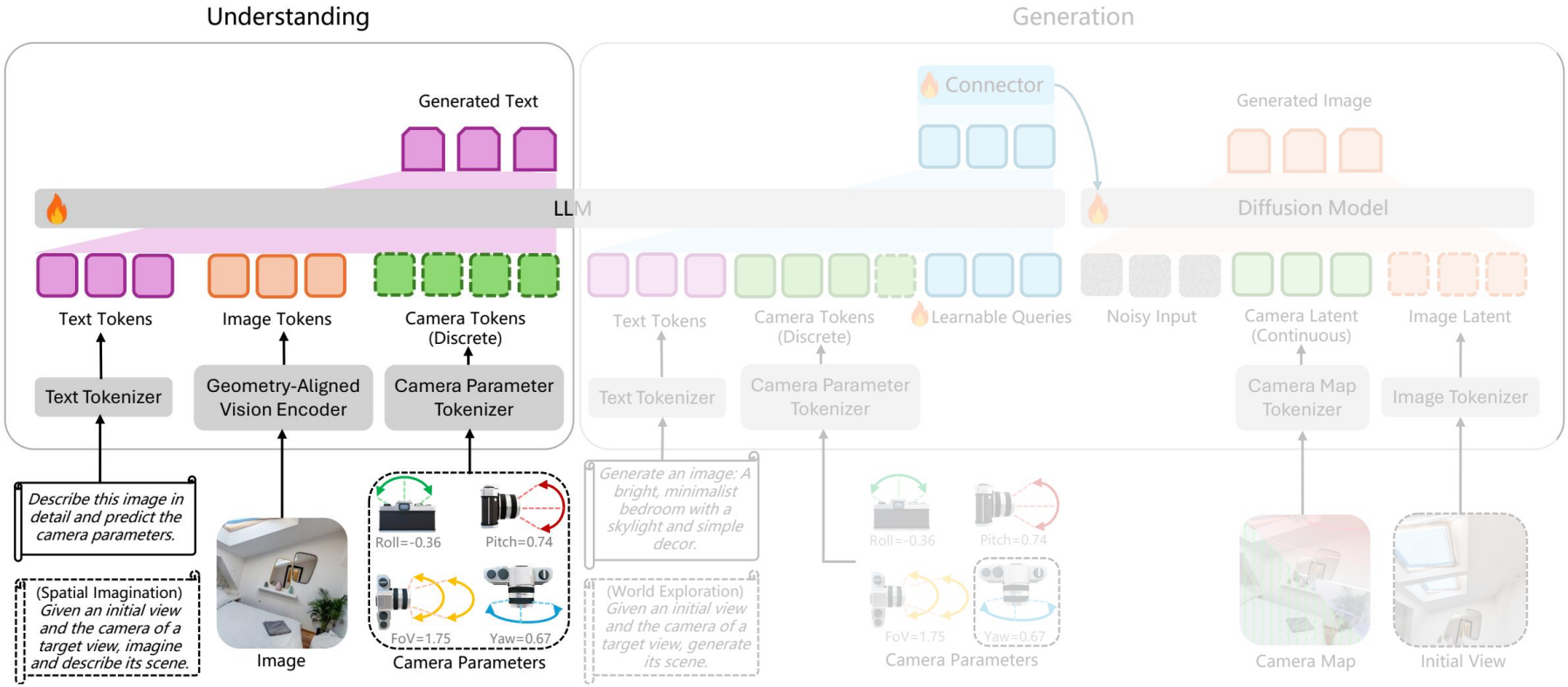
Method

- Camera-Centric Unified Multimodal Model
  - Puffin



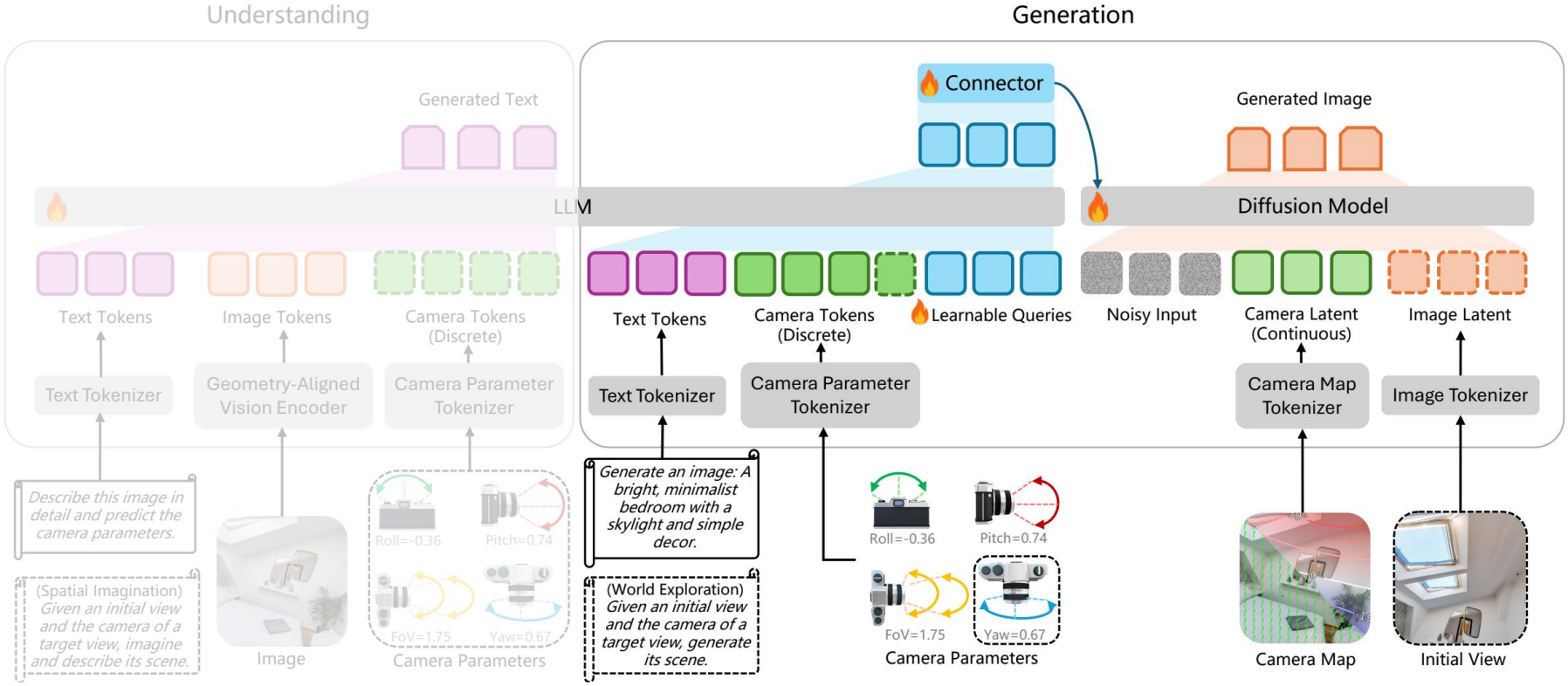
Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - Geometry-Aligned Vision Encoder
      - 기하학적 특징을 유지하고 모델의 공간 분석 역량을 강화



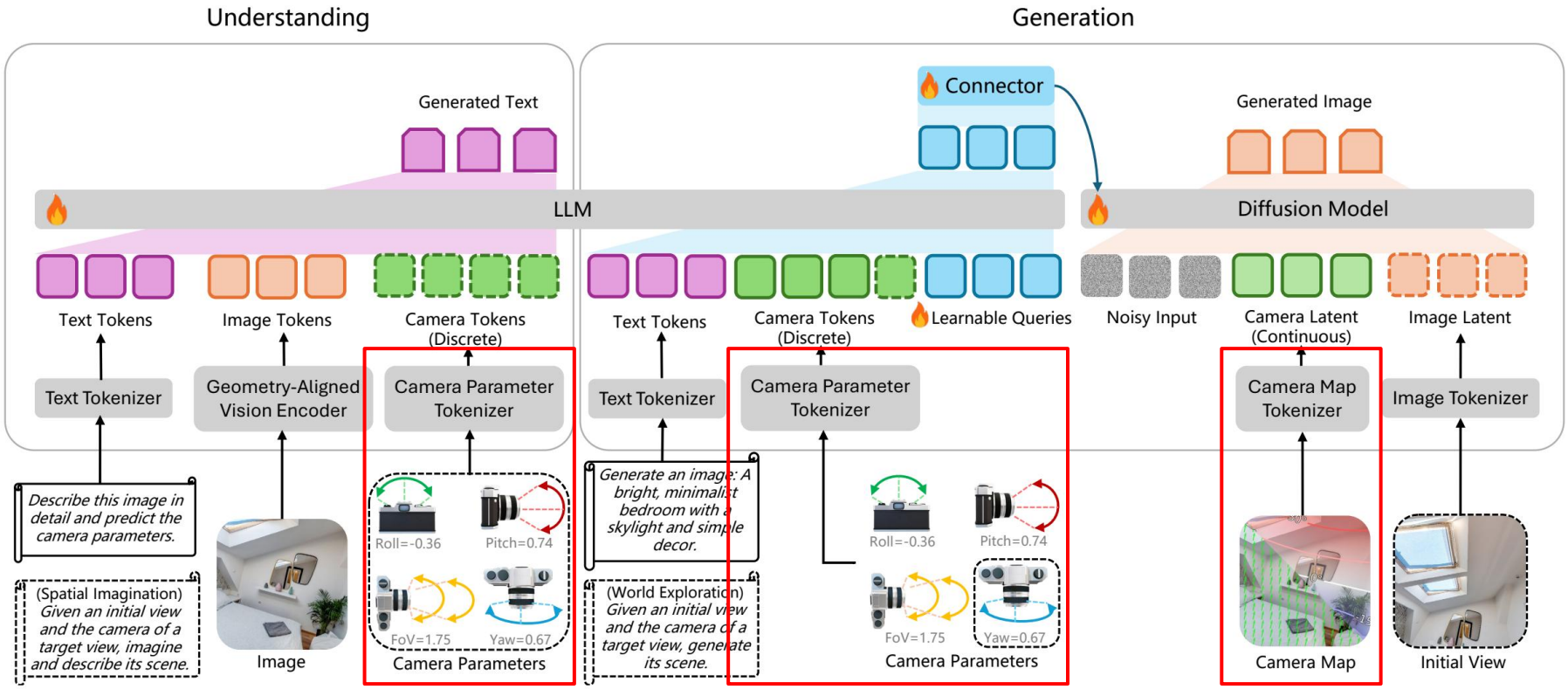
Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Generation
    - Connector Module
      - LLM의 hidden state를 learnable queries를 통해 diffusion model이 해석할 수 있는 조건 신호로 매핑하도록 학습



Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Geometry Align
    - Camera Parameter/Map Tokenizer



Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - 카메라를 언어로 해석하고 **thinking with camera**라는 개념을 도입
      - LMM이 기하학적 맥락 전반에 걸쳐 추론하는 동시에, 공간적으로 근거된 시각적 단서를 사진학적 용어와 정렬하도록 유도

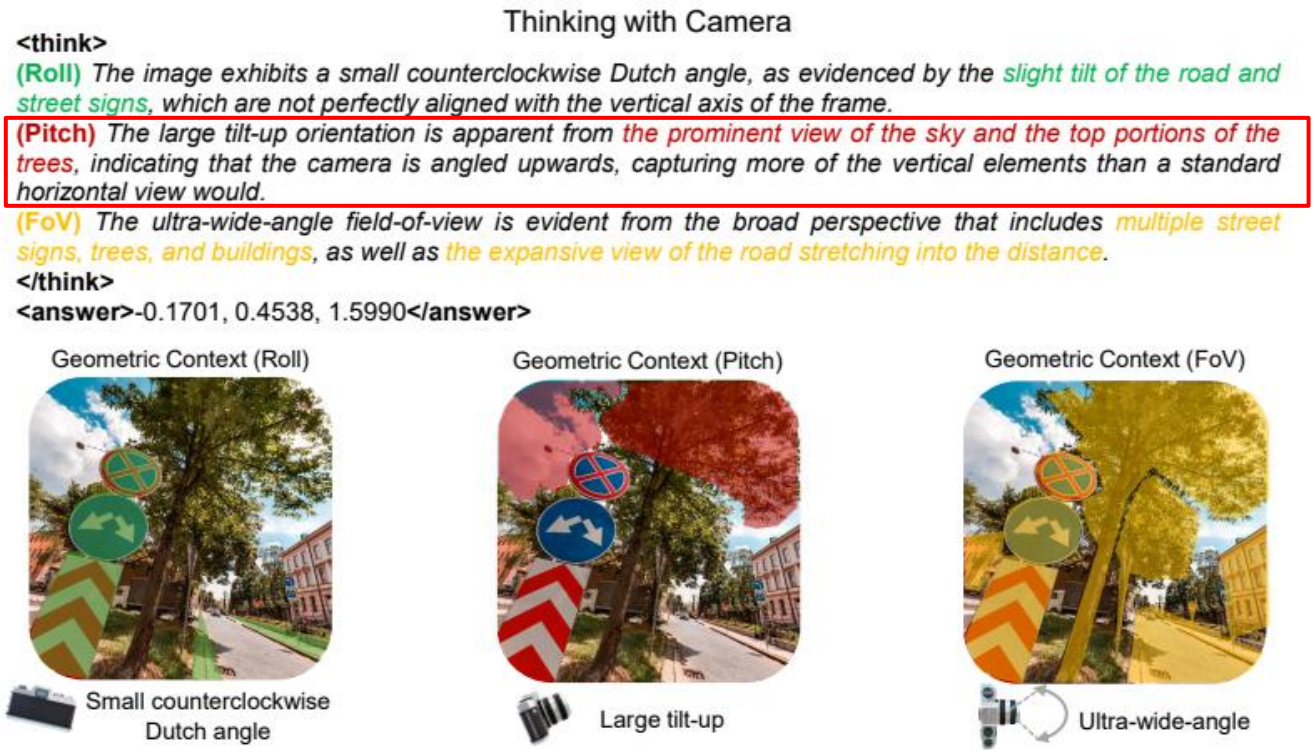
Thinking with Camera

<think>
(Roll) *The image exhibits a small counterclockwise Dutch angle, as evidenced by the slight tilt of the road and street signs, which are not perfectly aligned with the vertical axis of the frame.*
(Pitch) *The large tilt-up orientation is apparent from the prominent view of the sky and the top portions of the trees, indicating that the camera is angled upwards, capturing more of the vertical elements than a standard horizontal view would.*
(FoV) *The ultra-wide-angle field-of-view is evident from the broad perspective that includes multiple street signs, trees, and buildings, as well as the expansive view of the road stretching into the distance.*
</think>
<answer>-0.1701, 0.4538, 1.5990**</answer>**



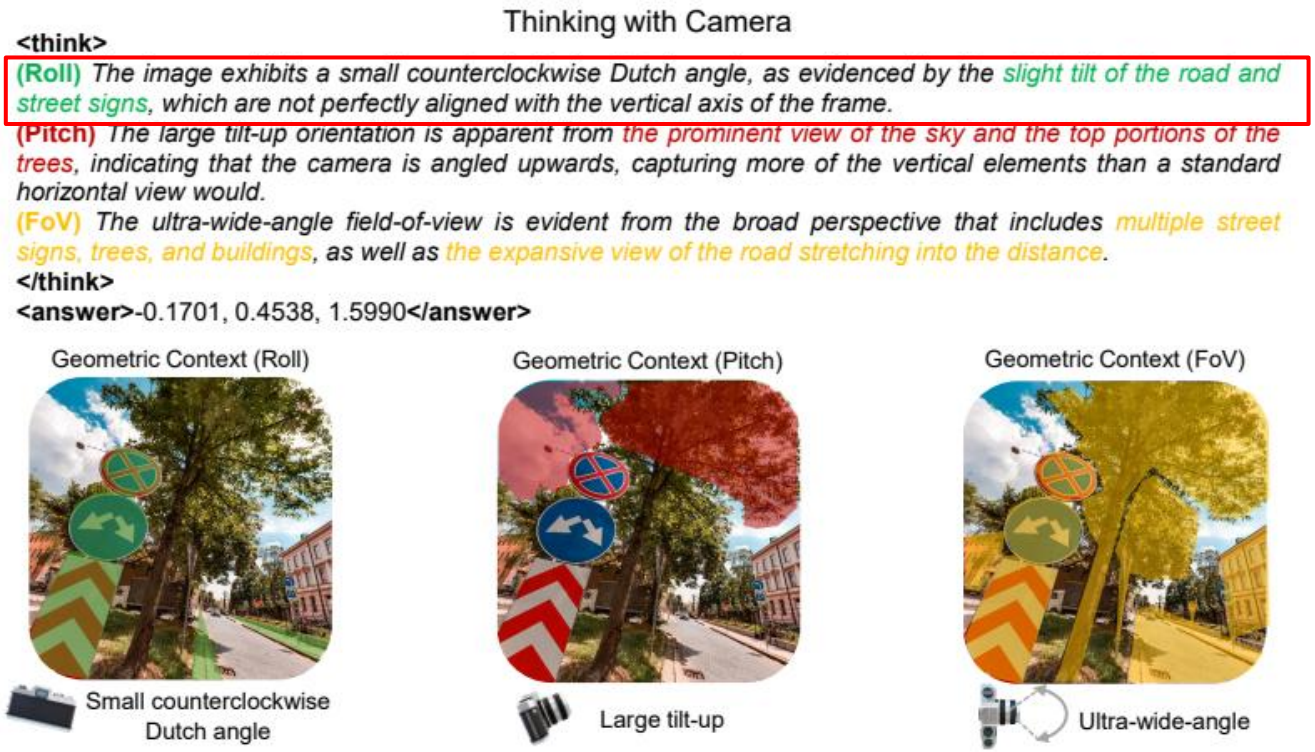
Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - 공간적 시각 단서
      - 하늘·바닥·전경–배경 비율과 같은 **공간적, 시각적 단서**를 <think> caption에 포함
      - LMM이 카메라 geometry에 대해 공간 추론을 수행하도록 함



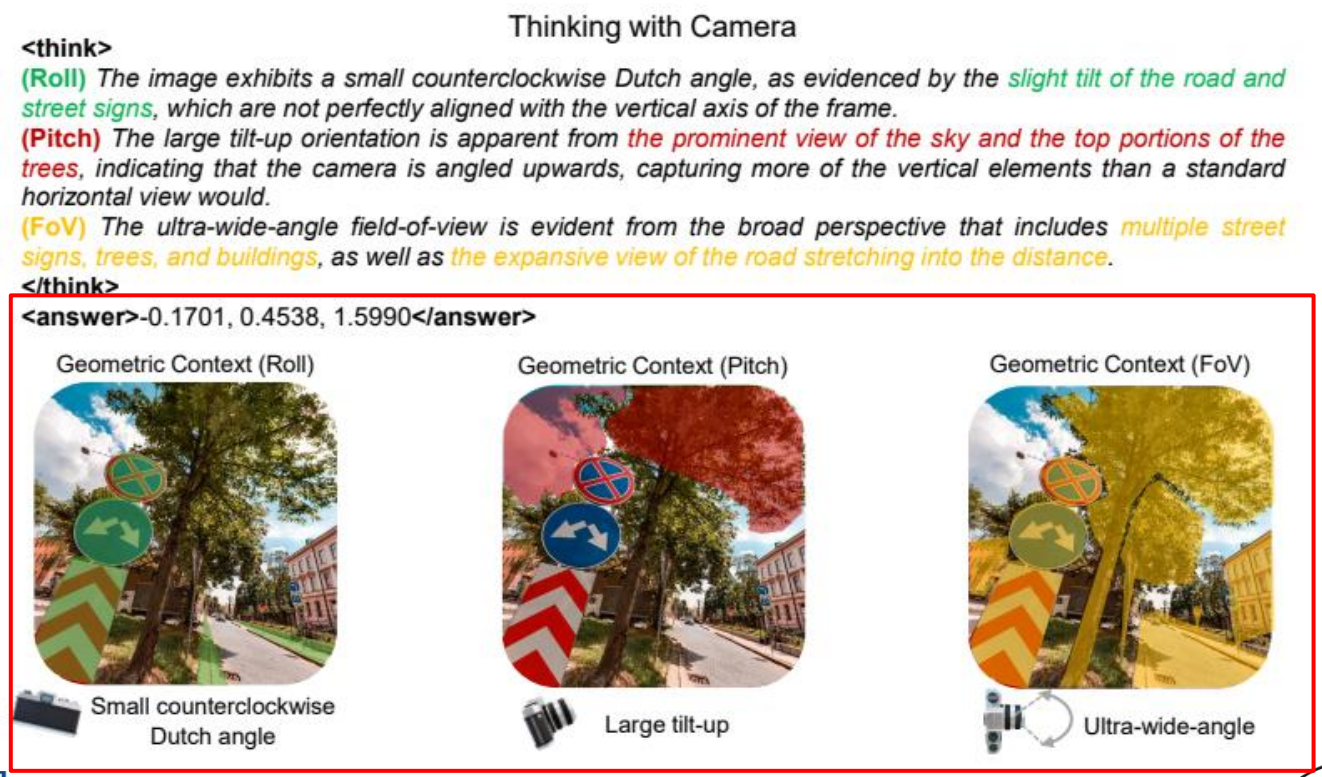
Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - Professional Photographic Terms
      - Close-up이나 dutch angle과 같은 **전문 사진학 용어**를 intermediate supervisory signals로 활용
      - 저·중수준 카메라 파라미터와 고수준 멀티모달 추론을 자연스럽게 연결



Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - Geometric Context
      - 카메라 파라미터를 roll, pitch, FoV 등 **기하학적 맥락별로 분해**, 이를 의미적으로 해석 가능한 descriptors와 정렬
      - 추상적 시각 특징과 물리적 기하를 연결

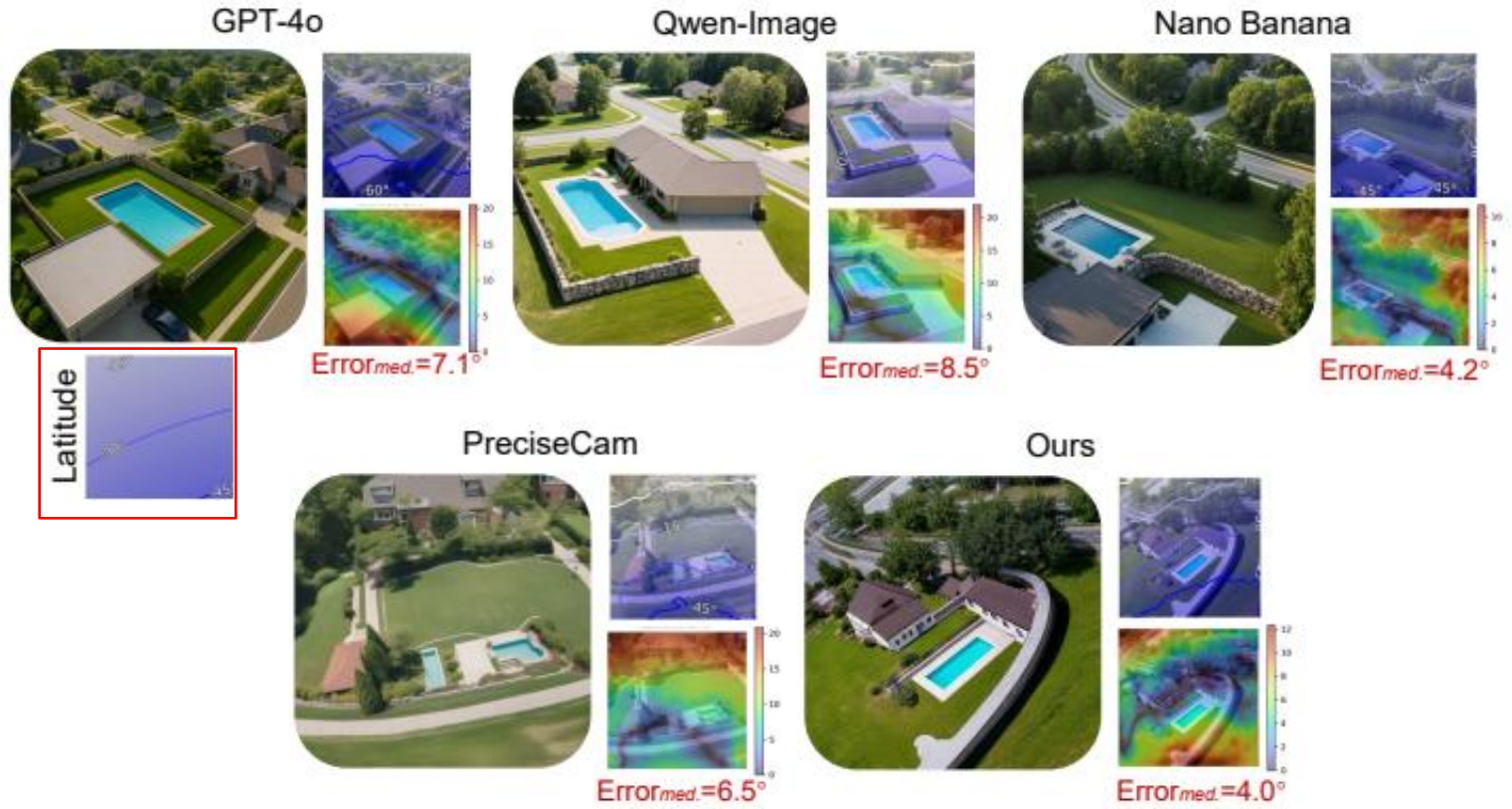


Method

- Camera-Centric Unified Multimodal Model
  - Puffin – for Understanding
    - LLM의 vision encoder가 geometry-align 특성을 갖도록 distillation 방식으로 학습
      - CLIP, SigLIP과 같은 의미 기반 teacher 모델
      - DINO, SAM과 같은 비전 중심 teacher 모델을 함께 활용

Experiments

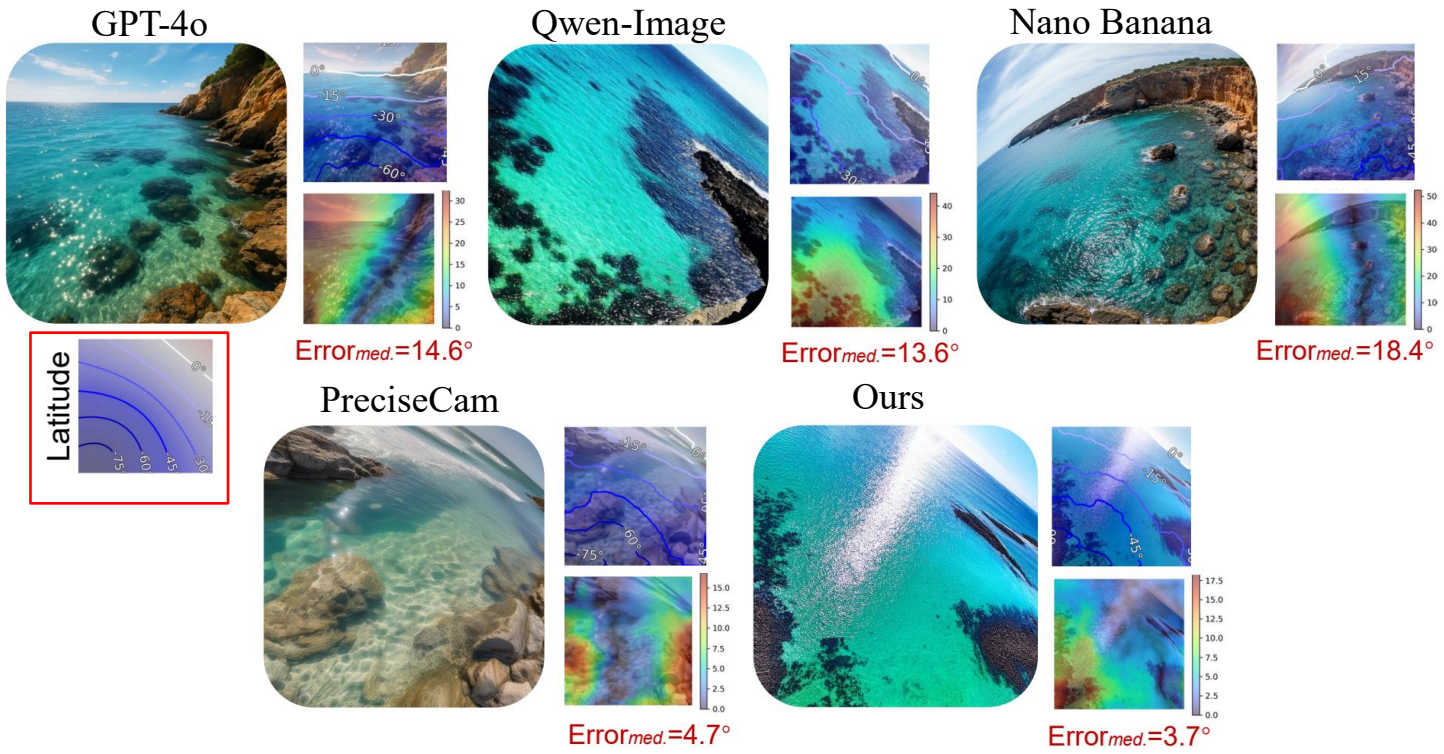
- Comparisons with Other Methods
  - Qualitative Results



Caption: *Aerial view of a well-maintained suburban neighborhood featuring a swimming pool surrounded by lush grass and bordered by a stone wall, with a flat-roofed house and driveway in the foreground, set against a backdrop of trees and parallel roads.*

Experiments

- Comparisons with Other Methods
  - Qualitative Results



Caption: *A vibrant coastal scene with clear turquoise waters reflecting sparkling sunlight, contrasting with darker rocky patches, set against a glimpse of the sky.*