# Free-Editor: Zero-shot Text-driven 3D Scene Editing

Nazmul Karim[1*] , Hasan Iqbal[2*] , Umar Khalid[1*] , Chen Chen[1] and Jing Hua[1]

[1]University of Central Florida, [2]Wayne State University

*Equal Contribution

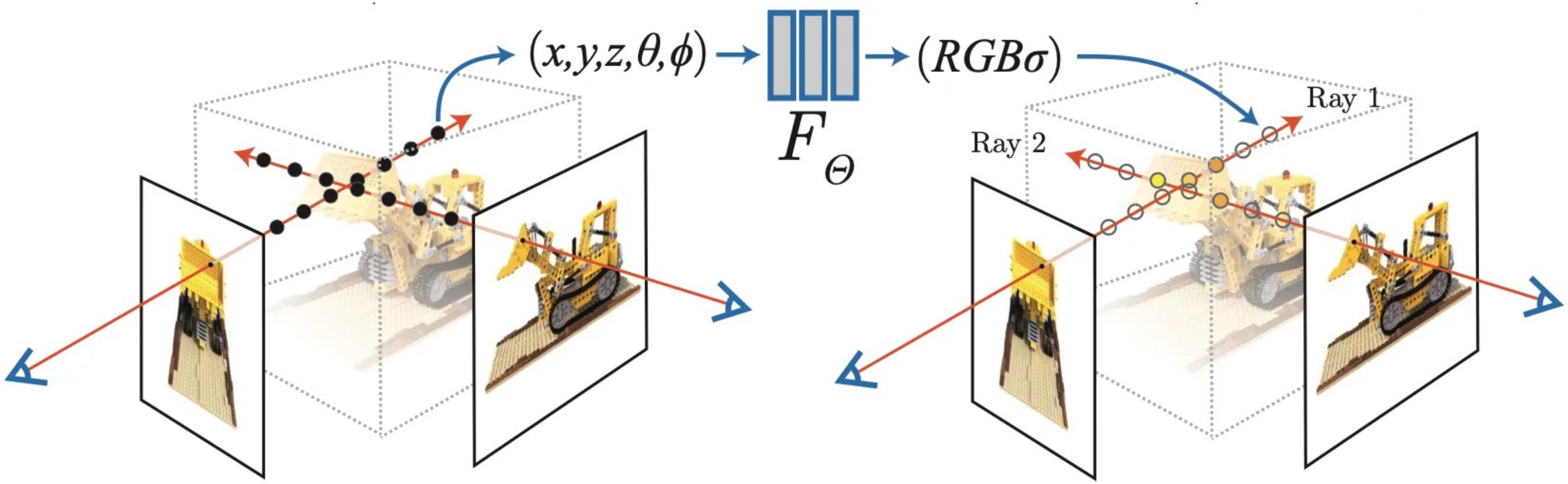
https://free-editor.github.io/

European Conference on Computer Vision (ECCV) 2024

Milan, Italy

Tue 1 Oct 10:30 a.m. - 12:30 p.m. EDT, Poster#69

# Neural Radiance Field (NeRF)*



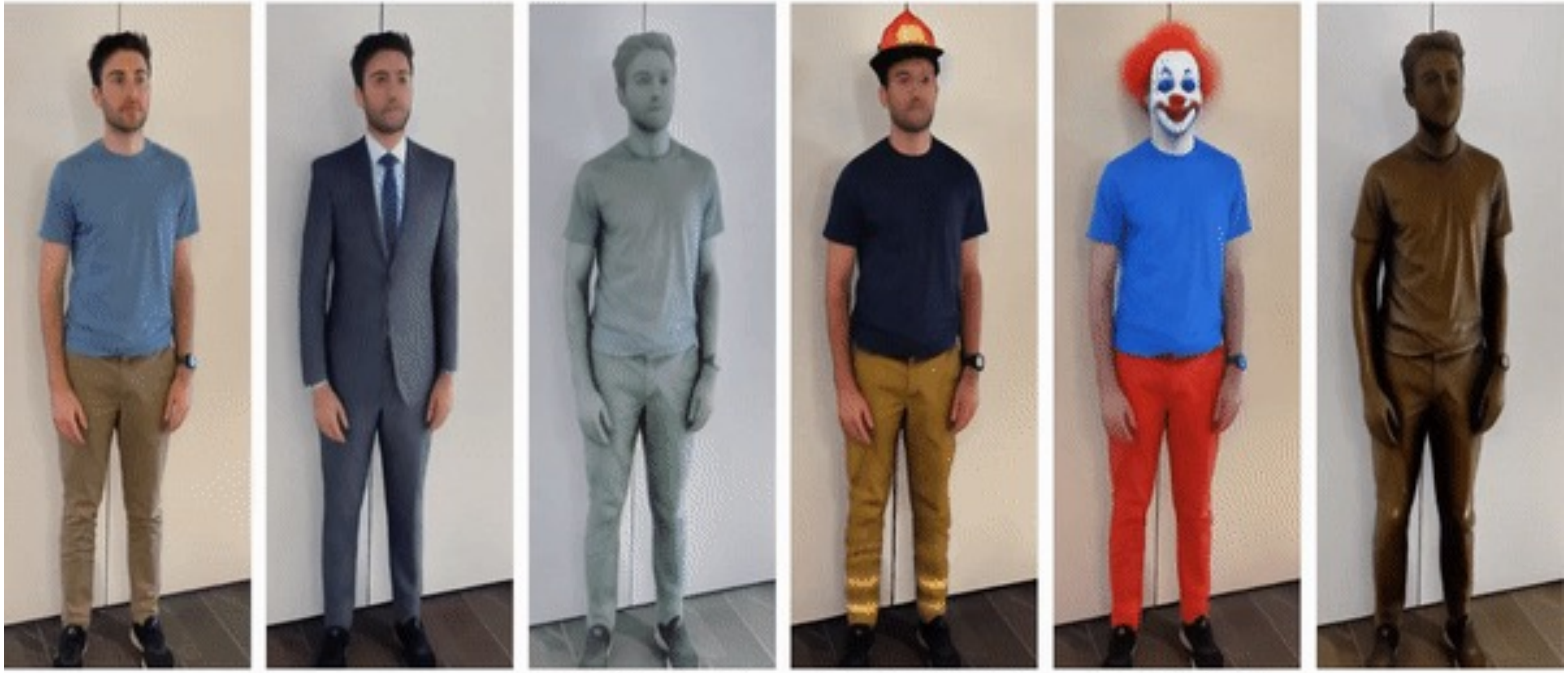
Given a set of images capturing the same object from multiple angles along with their corresponding poses-

- The network (F) learns to represent the 3D object by learning specific mappings
- New views can be synthesized in a consistent manner with the training set of views

*NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis (ECCV'20)

# Text-driven Editing of 3D NeRF Model (1)

SOTA InstructNeRF2NeRF[1] Model



Original NeRF

*"Put him in a suit"*

*"Make him into a marble statue"*

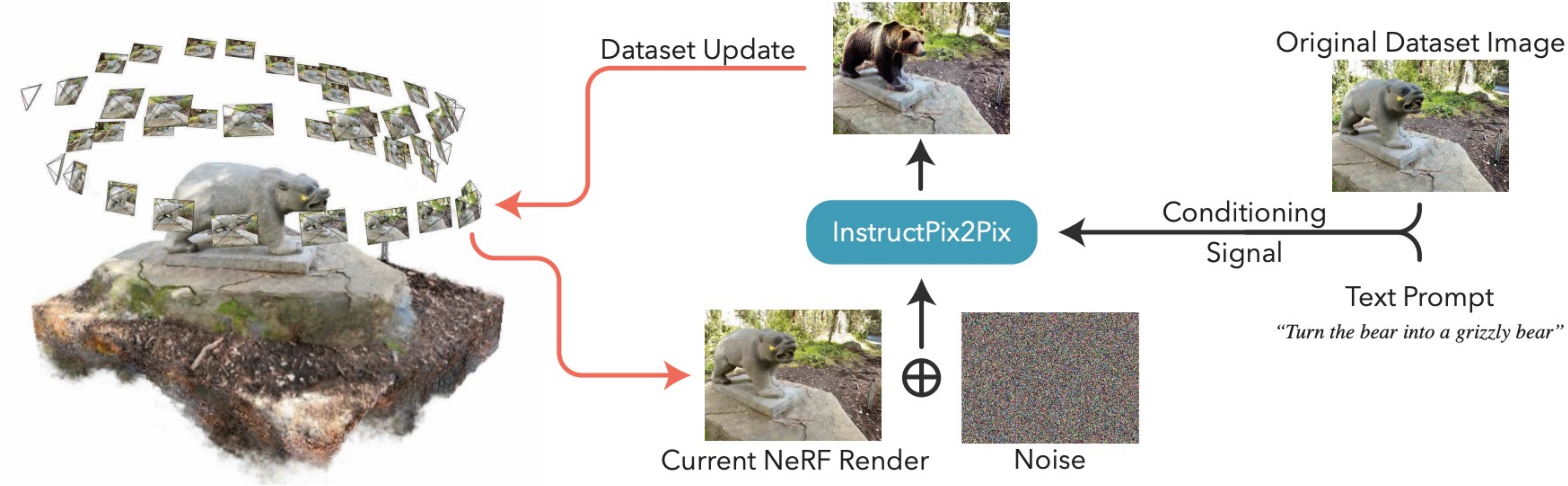
*"Turn him into a firefighter with a hat"*

*"Turn him into a clown"*

*"As a bronze statue"*

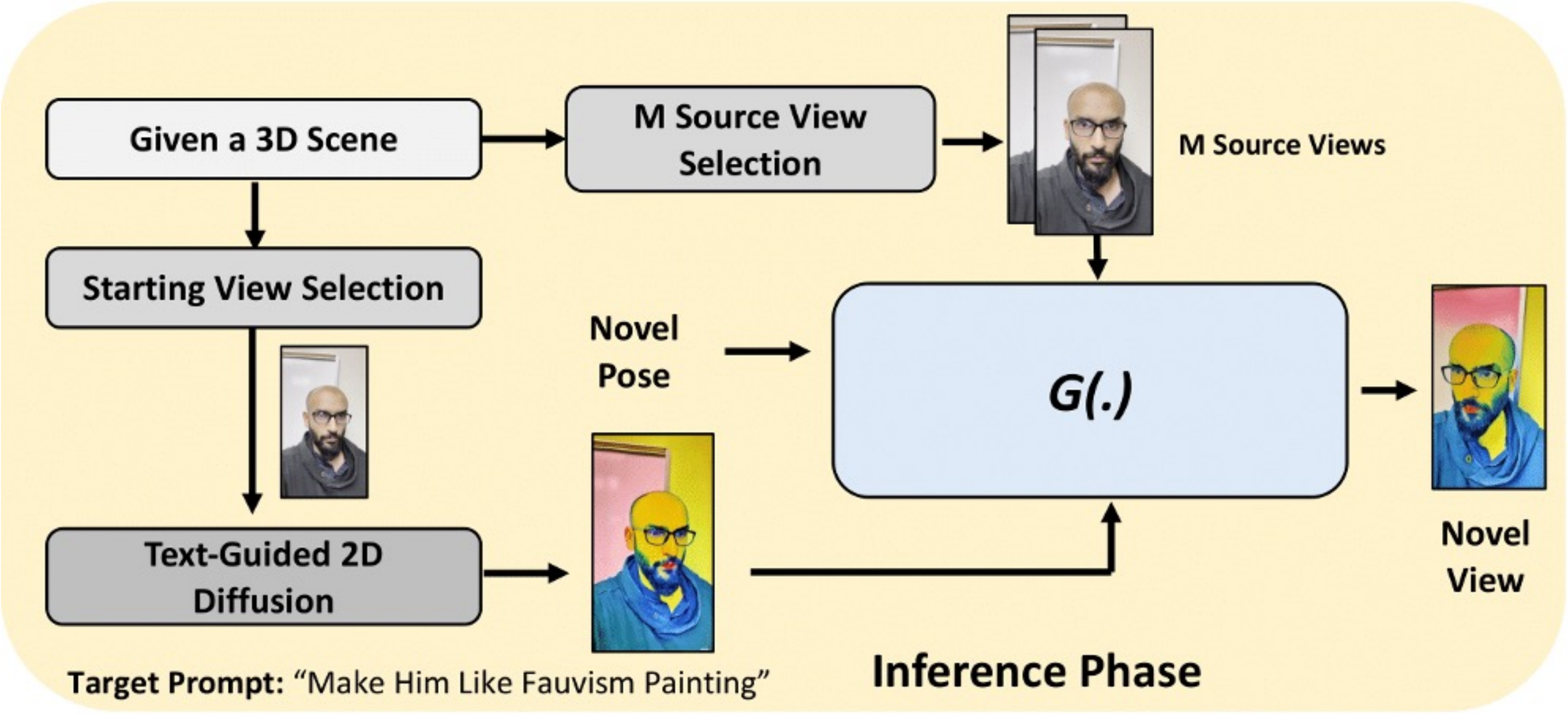
[1] Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions (ICCV'23)

# Text-driven Editing of NeRF Model (2)



**Issue:** Needs to re-train the NeRF again which is computationally inefficient

# Free-Editor: Edit Without Re-training

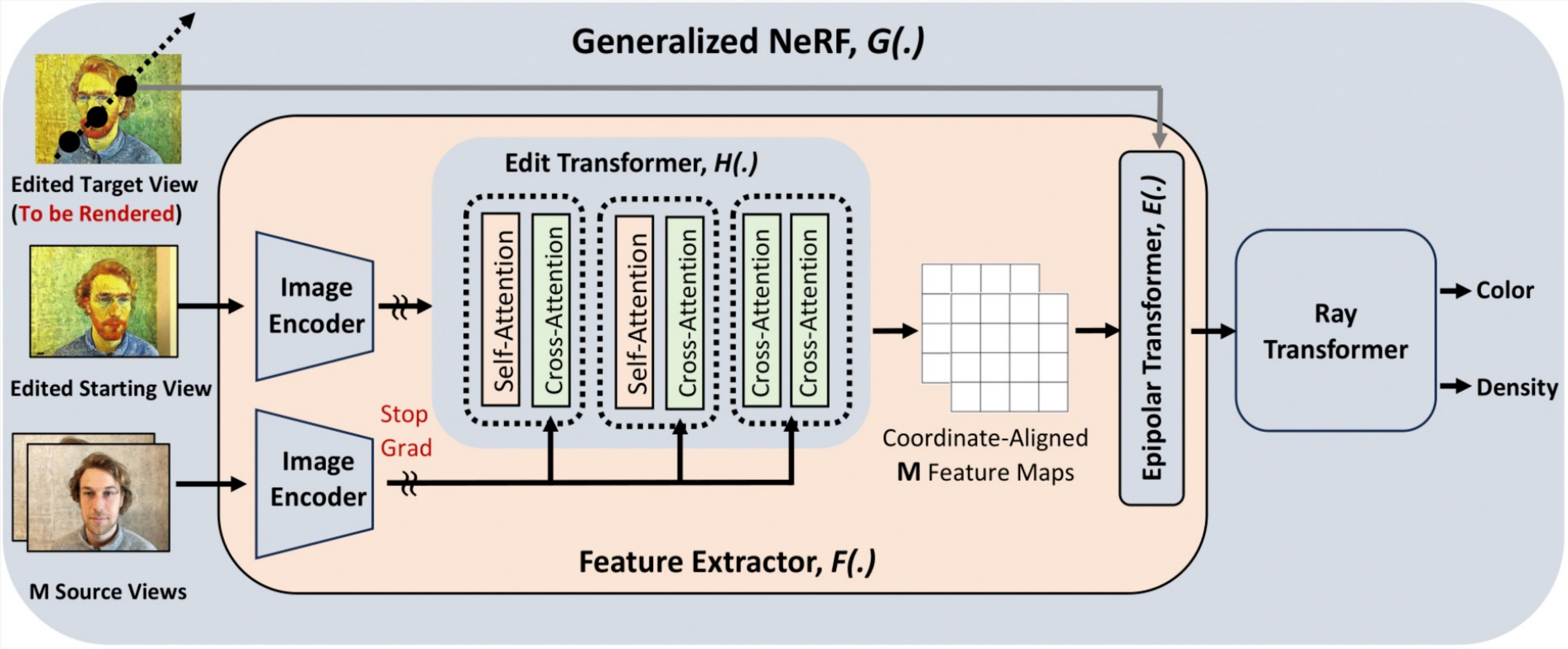


**Our Approach:** Edit only a single training image (starting view) and use a generalized NeRF (G) to obtain edited 3D Scene

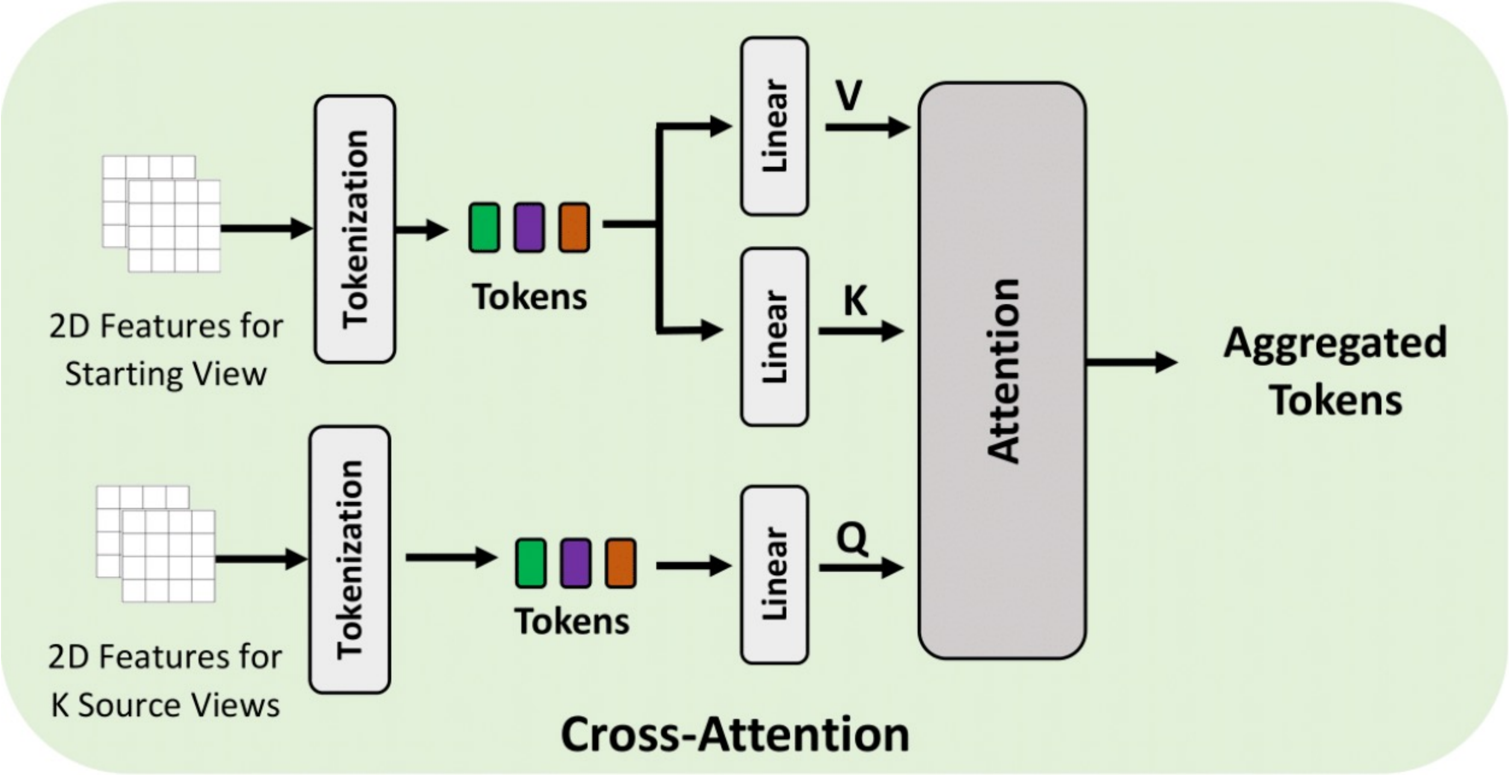
**Advantage**: 1. Unlike SOTA, No Re-training is required

2. Editing time is 3.5 minutes as compared to 70 minutes in SOTA

# Generalized NeRF (1)



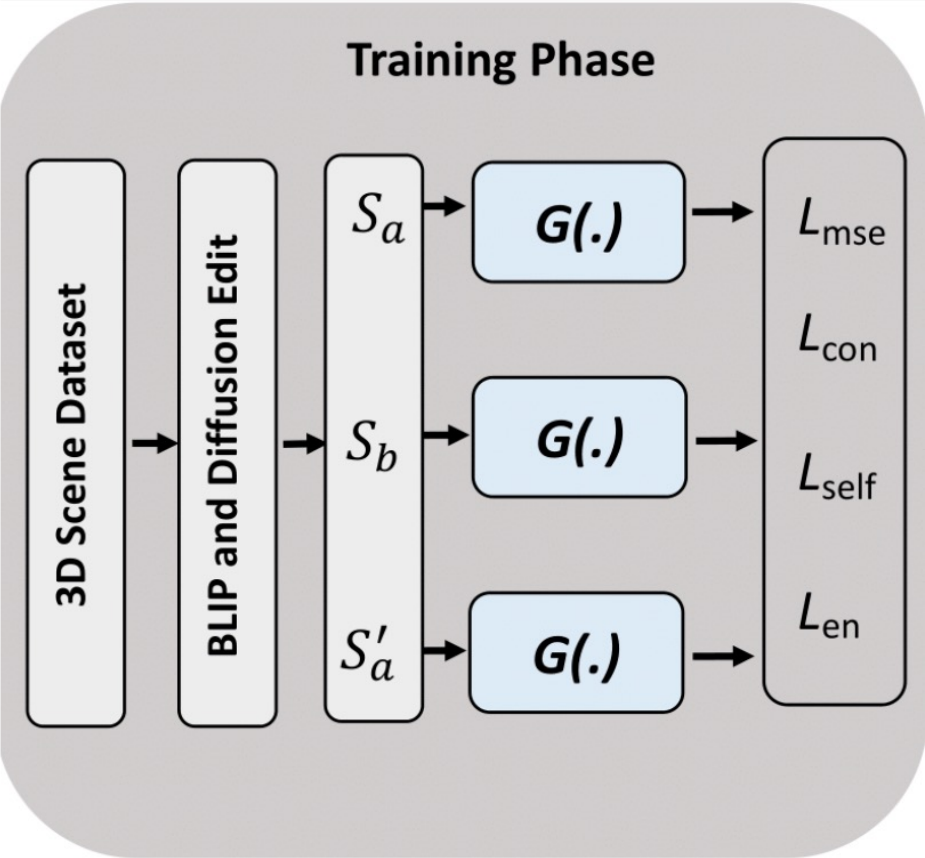
# Generalized NeRF (2)



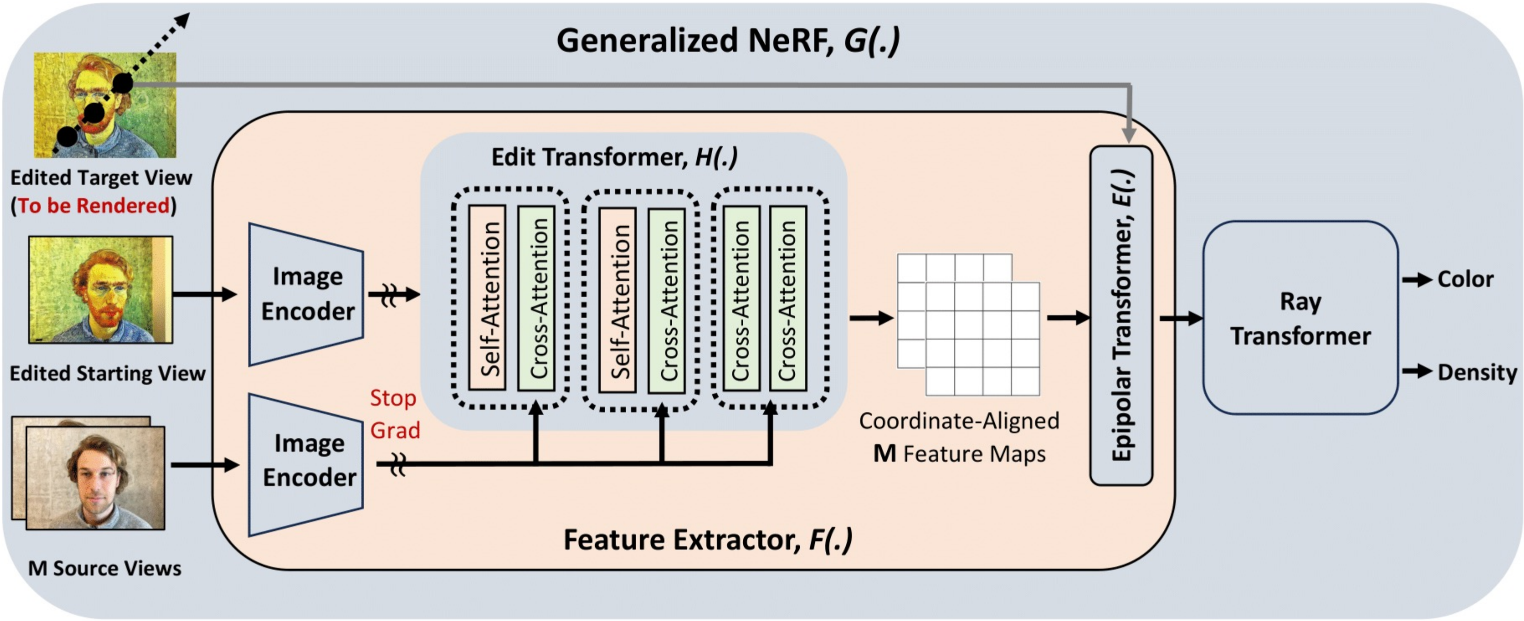
- ✓ We tokenize the 2D features for starting view
  - ○ Feed them to linear layers to produce key (K) and value (V)
- ✓ Separately, linear embeddings of 2D features for K source views used as Query (Q)
- ✓ Finally, we ge the aggregated tokens through cross-attention mechanism

# Training Details

- ✓ First, we get several 3D scene datasets
- ✓ Use BLIP to generate the description of each scene
- ✓ Generate multiple modified version of the original description
- ✓ Use a text-to-image diffusion model to edit the rendered images of each scene
- ✓ Send them to Generalized NeRF and calculate the losses



# Free-Editor: Edit Without Re-training



Generalized NeRF, *G(.)*

Edited Target View (To be Rendered)

Edited Starting View

M Source Views

Image Encoder

Image Encoder

Stop Grad

Edit Transformer, *H(.)*

Self-Attention

Cross-Attention

Self-Attention

Cross-Attention

Cross-Attention

Cross-Attention

Coordinate-Aligned **M** Feature Maps

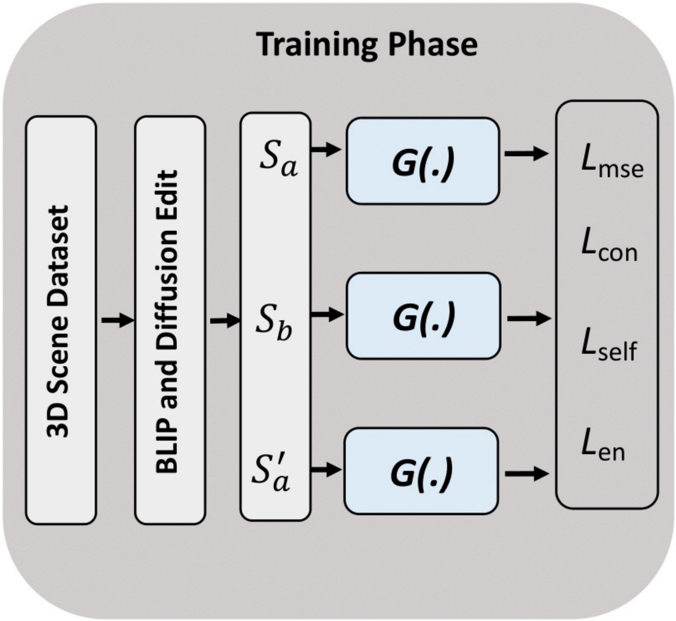
Epipolar Transformer, *E(.)*

Ray Transformer

Color

Density

Feature Extractor, *F(.)*



Training Phase

3D Scene Dataset

BLIP and Diffusion Edit

$S_a$

$S_b$

$S_a'$

*G(.)*

*G(.)*

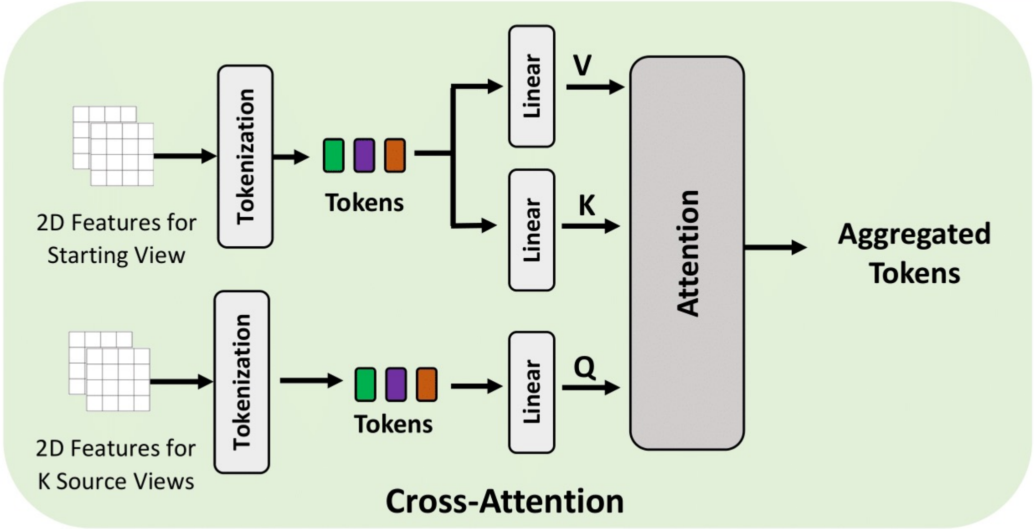
*G(.)*

$L_{mse}$

$L_{con}$

$L_{self}$

$L_{en}$



2D Features for Starting View

2D Features for K Source Views

Tokenization

Tokenization

Tokens

Tokens

Linear

Linear

Linear

V

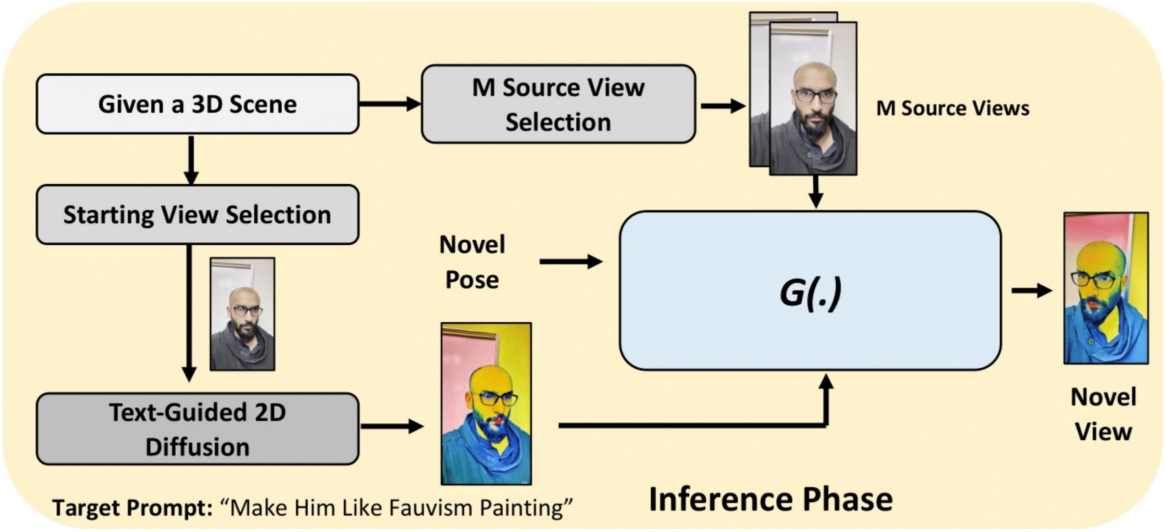
K

Q

Attention

Aggregated Tokens

Cross-Attention



Given a 3D Scene

M Source View Selection

M Source Views

Starting View Selection

Novel Pose

*G(.)*

Novel View

Text-Guided 2D Diffusion

**Target Prompt:** "Make Him Like Fauvism Painting"

Inference Phase

# Experimental Settings

## For Training, We use-

- Google Scanned Objects
- NerfStudio
- Spaces and
- IBRNet-collect
- Nerf-Art
- RealEstate10K
- OmniObject3D

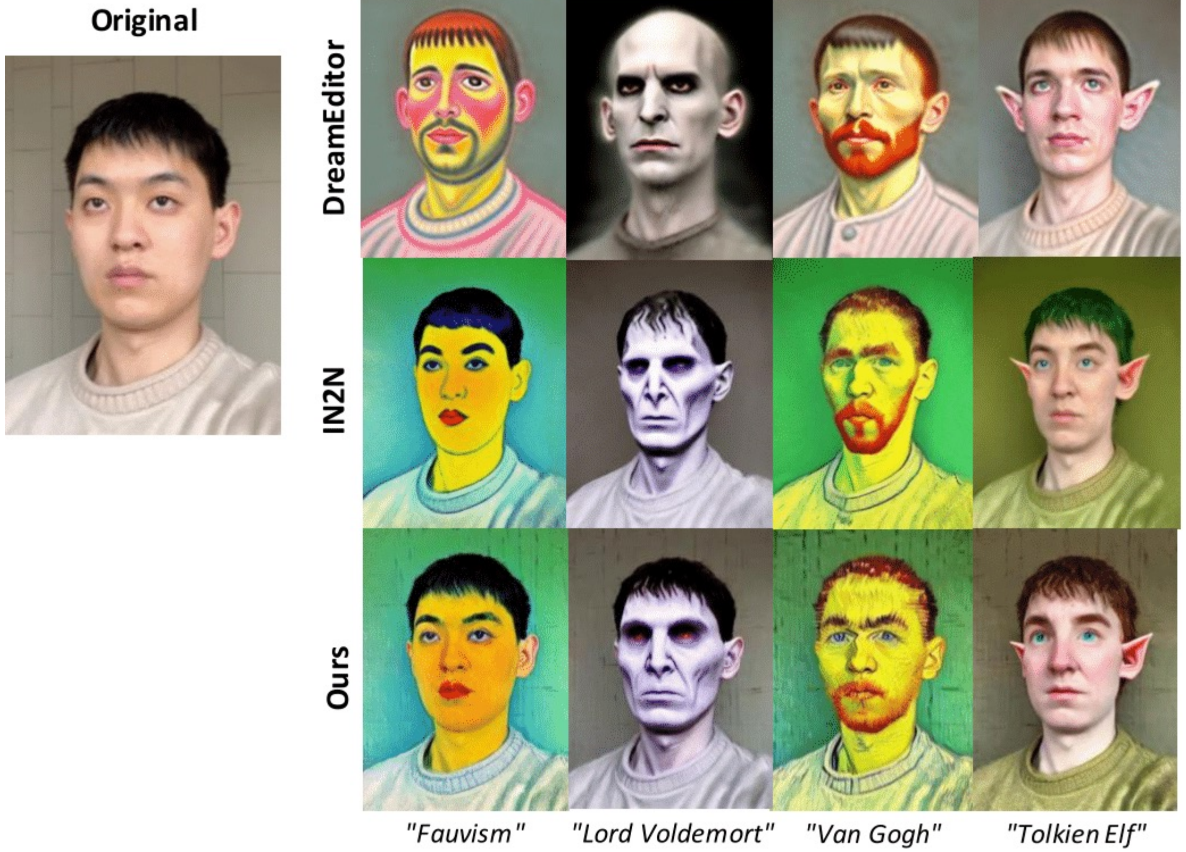
## For evaluation, we use-

- IN2N
- NeRFSynthetic
- LLFF and
- Our own dataset of four scenes.

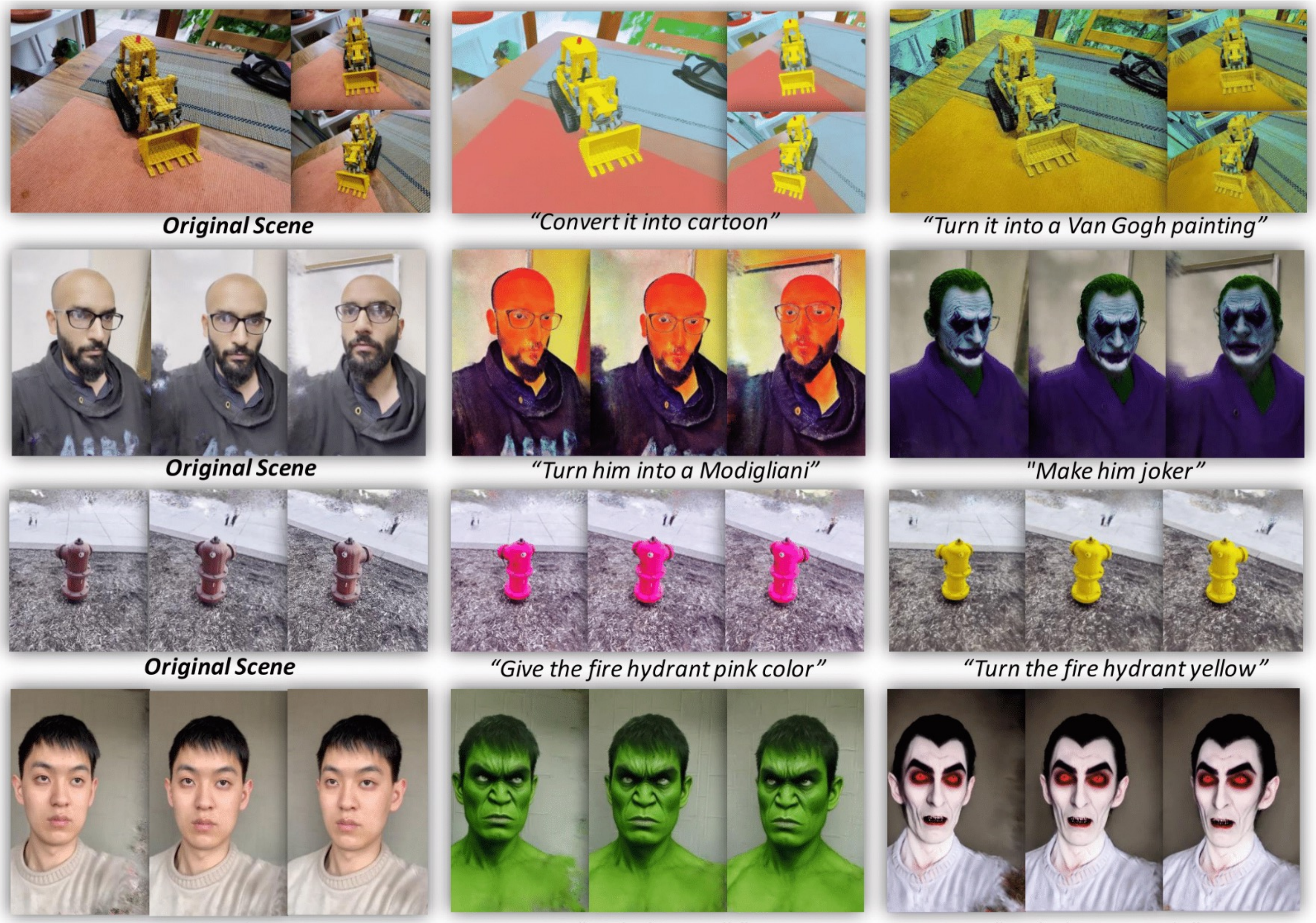
# Experimental Results

✓ Capture both the color palette and stroke patterns of the desired style.

✓ Preserve background details more effectively than IN2N



# Experimental Results



***Original Scene***

*“Convert it into cartoon”*

*“Turn it into a Van Gogh painting”*

***Original Scene***

*“Turn him into a Modigliani”*

*"Make him joker”*

***Original Scene***

*“Give the fire hydrant pink color”*

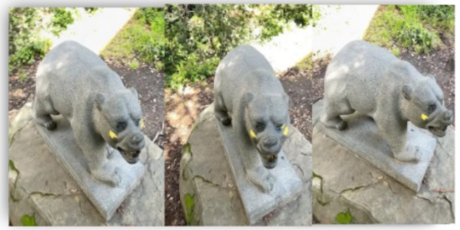
*“Turn the fire hydrant yellow”*

***Original Scene***

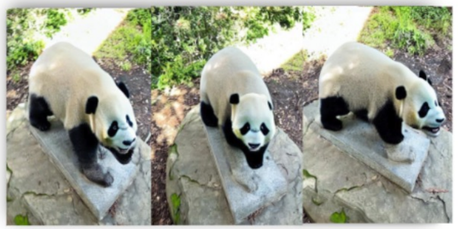
*“Make him Hulk”*

*“Make him Dracula”*

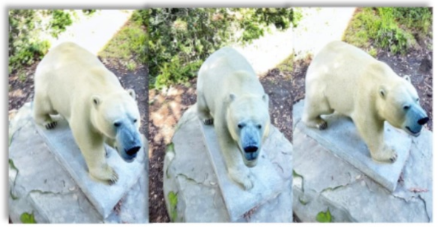
# Experimental Results



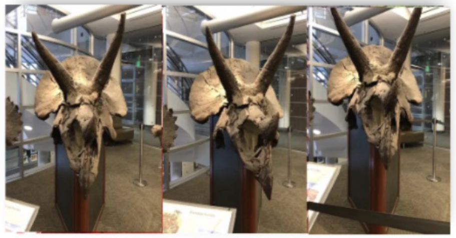
***Original Scene***



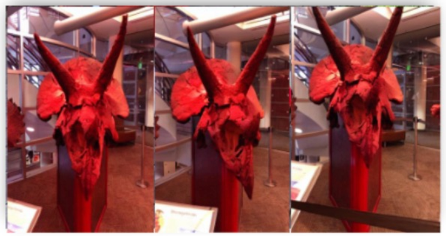
*“Turn the bear into a panda”*



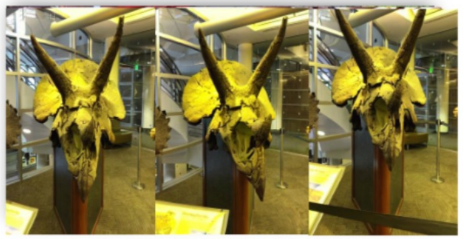
*"Turn the bear into a polar bear”*



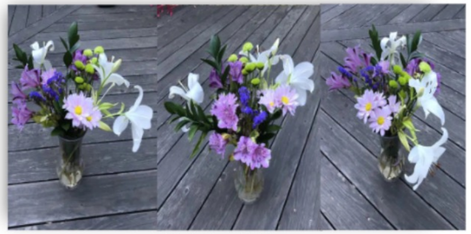
***Original Scene***



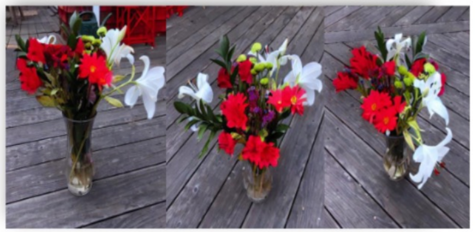
*"Turn the T-Rex Red"*



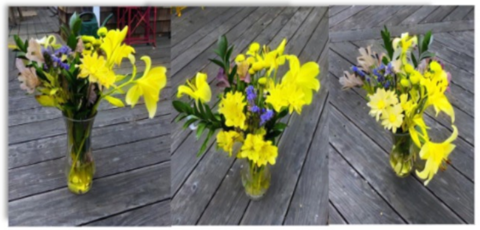
*“Turn the T-Rex Yellow”*



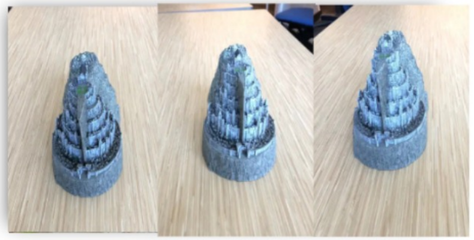
***Original Scene***



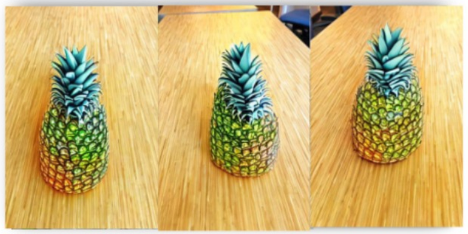
*”Turn all flowers except white into Red”*



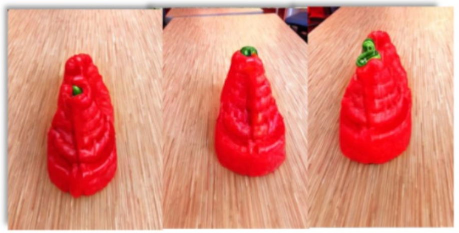
*“Turn the white flowers Yellow”*



***Original Scene***



*“Turn it into Pineapple”*



*"Turn it into Strawberry”*

Thank You