ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

NEURAL
INFORMATION
PROCESSING
SYSTEMS

LLaNA: Large Language and NeRF Assistant

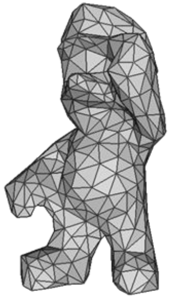
NeurIPS 2024

Andrea Amaduzzi, Pierluigi Zama Ramirez, Giuseppe Lisanti,
Samuele Salti, Luigi Di Stefano

Motivation

Multi-view images

Point cloud


Triangle mesh

Voxel grid

NeRF

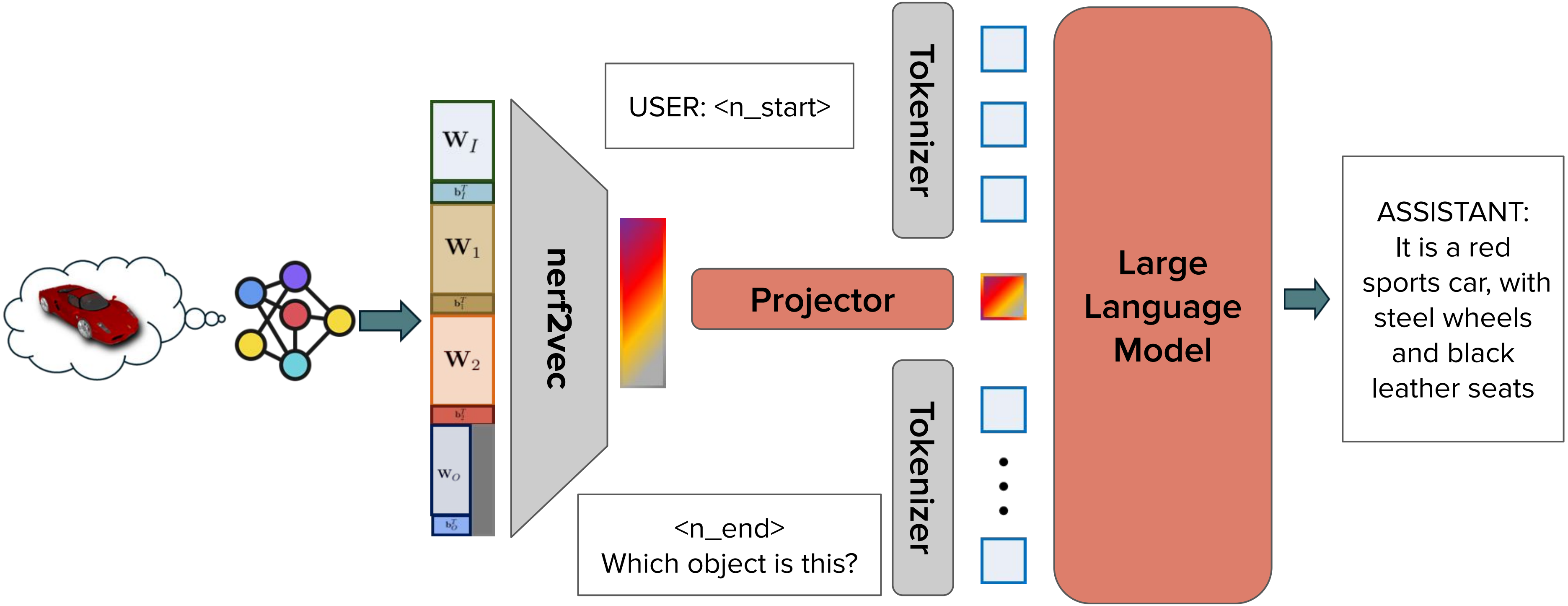
Neural Radiance Fields (NeRFs) can overcome limitations of 3D explicit representations

Neural Radiance Fields (NeRFs) can provide a compact and meaningful representation of objects' geometry and appearance

Multimodal Large Language Models (MLLMs) can process 3D explicit representations

Is it possible to perform language tasks on NeRF, without rendering?

LLaNA: Large Language and NeRF Assistant



USER: <n_start>

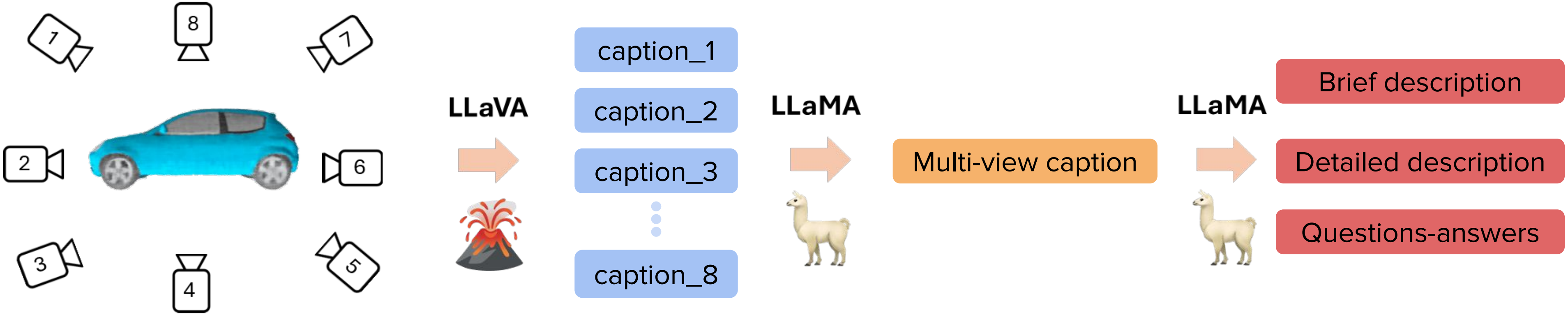
<n_end>
Which object is this?

ASSISTANT:
It is a red sports car, with steel wheels and black leather seats

LLaNA can perform the novel tasks of NeRF captioning and NeRF Q&A

nerf2vec: Ramirez, et al. “Deep Learning on Object-centric 3D Neural Fields”, ICLR 2023

ShapeNeRF-Text: the first Nerf-Language benchmark



LLaVA

caption_1

caption_2

caption_3

caption_8

LLaMA

Multi-view caption

LLaMA

Brief description

Detailed description

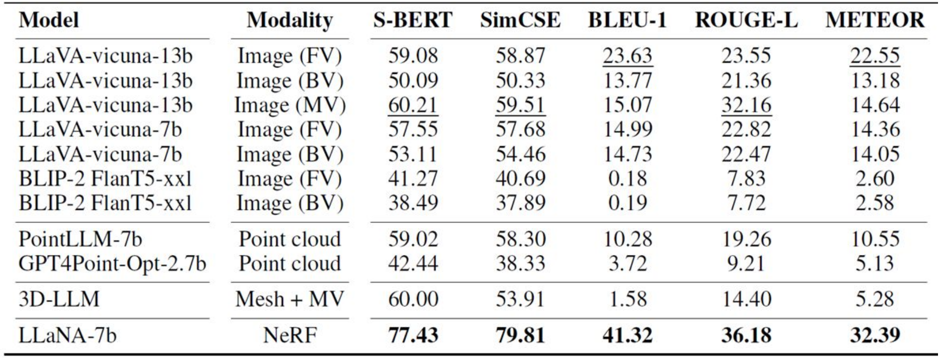
Questions-answers

Dataset size:

- 40K objects from ShapeNet
- 40K brief descriptions
- 40K detailed descriptions
- 160K QA conversations

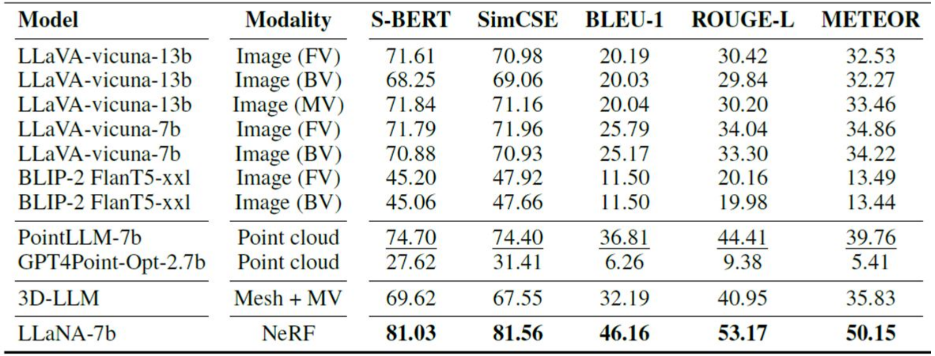
Quantitative experiments

NeRF captioning



Model	Modality	S-BERT	SimCSE	BLEU-1	ROUGE-L	METEOR
LLaVA-vicuna-13b	Image (FV)	59.08	58.87	23.63	23.55	22.55
LLaVA-vicuna-13b	Image (BV)	50.09	50.33	13.77	21.36	13.18
LLaVA-vicuna-13b	Image (MV)	60.21	59.51	15.07	32.16	14.64
LLaVA-vicuna-7b	Image (FV)	57.55	57.68	14.99	22.82	14.36
LLaVA-vicuna-7b	Image (BV)	53.11	54.46	14.73	22.47	14.05
BLIP-2 FlanT5-xxl	Image (FV)	41.27	40.69	0.18	7.83	2.60
BLIP-2 FlanT5-xxl	Image (BV)	38.49	37.89	0.19	7.72	2.58
PointLLM-7b	Point cloud	59.02	58.30	10.28	19.26	10.55
GPT4Point-Opt-2.7b	Point cloud	42.44	38.33	3.72	9.21	5.13
3D-LLM	Mesh + MV	60.00	53.91	1.58	14.40	5.28
LLaNA-7b	NeRF	**77.43**	**79.81**	**41.32**	**36.18**	**32.39**

NeRF Q&A

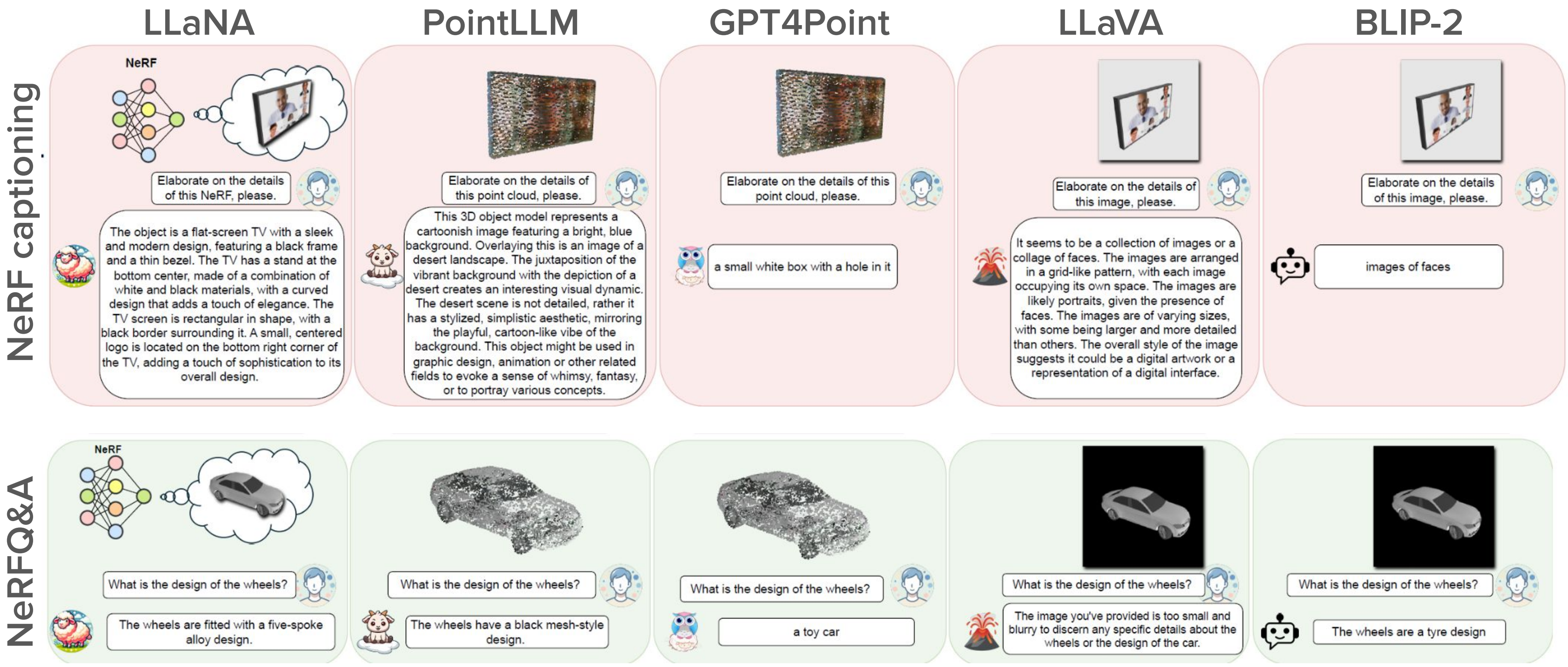


Model	Modality	S-BERT	SimCSE	BLEU-1	ROUGE-L	METEOR
LLaVA-vicuna-13b	Image (FV)	71.61	70.98	20.19	30.42	32.53
LLaVA-vicuna-13b	Image (BV)	68.25	69.06	20.03	29.84	32.27
LLaVA-vicuna-13b	Image (MV)	71.84	71.16	20.04	30.20	33.46
LLaVA-vicuna-7b	Image (FV)	71.79	71.96	25.79	34.04	34.86
LLaVA-vicuna-7b	Image (BV)	70.88	70.93	25.17	33.30	34.22
BLIP-2 FlanT5-xxl	Image (FV)	45.20	47.92	11.50	20.16	13.49
BLIP-2 FlanT5-xxl	Image (BV)	45.06	47.66	11.50	19.98	13.44
PointLLM-7b	Point cloud	74.70	74.40	36.81	44.41	39.76
GPT4Point-Opt-2.7b	Point cloud	27.62	31.41	6.26	9.38	5.41
3D-LLM	Mesh + MV	69.62	67.55	32.19	40.95	35.83
LLaNA-7b	NeRF	**81.03**	**81.56**	**46.16**	**53.17**	**50.15**

FV: “front view”, BV: “back view”, MV: “multiple views”

LLaNA outperforms MLLMs processing discrete representations

Qualitative experiments



NeRF captioning

LLaNA

Elaborate on the details of this NeRF, please.

The object is a flat-screen TV with a sleek and modern design, featuring a black frame and a thin bezel. The TV has a stand at the bottom center, made of a combination of white and black materials, with a curved design that adds a touch of elegance. The TV screen is rectangular in shape, with a black border surrounding it. A small, centered logo is located on the bottom right corner of the TV, adding a touch of sophistication to its overall design.

PointLLM

Elaborate on the details of this point cloud, please.

This 3D object model represents a cartoonish image featuring a bright, blue background. Overlaying this is an image of a desert landscape. The juxtaposition of the vibrant background with the depiction of a desert creates an interesting visual dynamic. The desert scene is not detailed, rather it has a stylized, simplistic aesthetic, mirroring the playful, cartoon-like vibe of the background. This object might be used in graphic design, animation or other related fields to evoke a sense of whimsy, fantasy, or to portray various concepts.

GPT4Point

Elaborate on the details of this point cloud, please.

a small white box with a hole in it

LLaVA

Elaborate on the details of this image, please.

It seems to be a collection of images or a collage of faces. The images are arranged in a grid-like pattern, with each image occupying its own space. The images are likely portraits, given the presence of faces. The images are of varying sizes, with some being larger and more detailed than others. The overall style of the image suggests it could be a digital artwork or a representation of a digital interface.

BLIP-2

Elaborate on the details of this image, please.

images of faces

NeRFQ&A

LLaNA

What is the design of the wheels?

The wheels are fitted with a five-spoke alloy design.

PointLLM

What is the design of the wheels?

The wheels have a black mesh-style design.

GPT4Point

What is the design of the wheels?

a toy car

LLaVA

What is the design of the wheels?

The image you've provided is too small and blurry to discern any specific details about the wheels or the design of the car.

BLIP-2

What is the design of the wheels?

The wheels are a tyre design

Conclusion

Contributions:

- LLaNA is the first MLLM to demonstrate reasoning capabilities on NeRF data
- LLaNA enables novel NeRF-language tasks (captioning, Q&A)
- ShapeNeRF-Text is the first existing dataset of paired NeRF-language data

Future directions:

- Generalization to real-world NeRFs
- Support for more complex NeRF architectures
- Extension to scene-level NeRFs