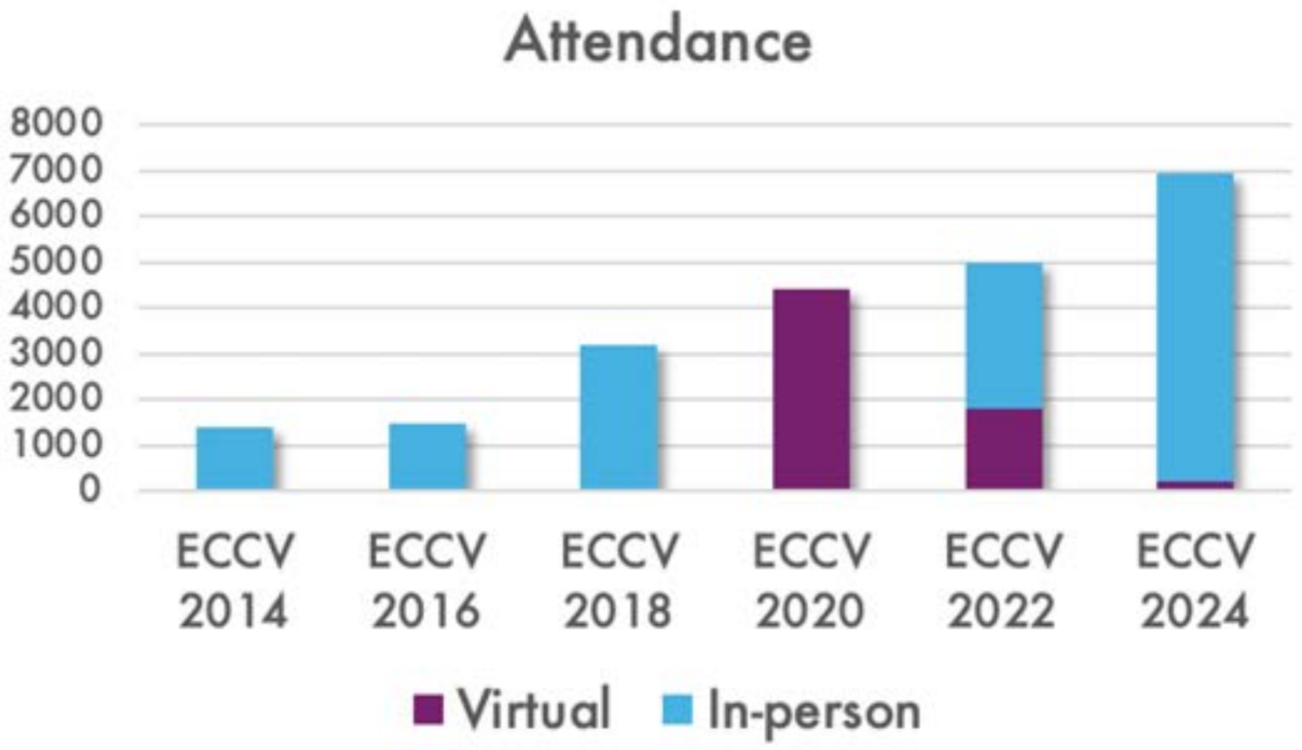
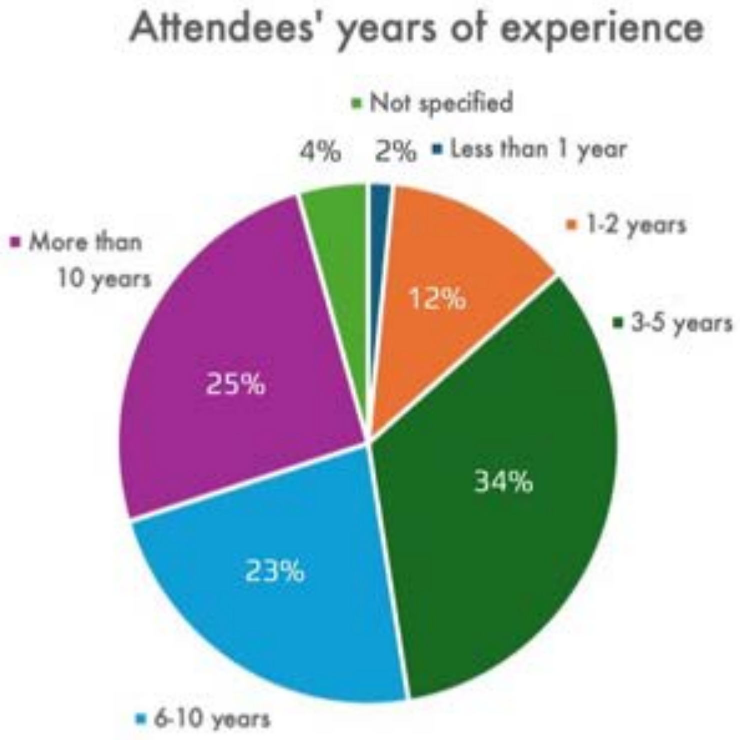
ECCV Summary

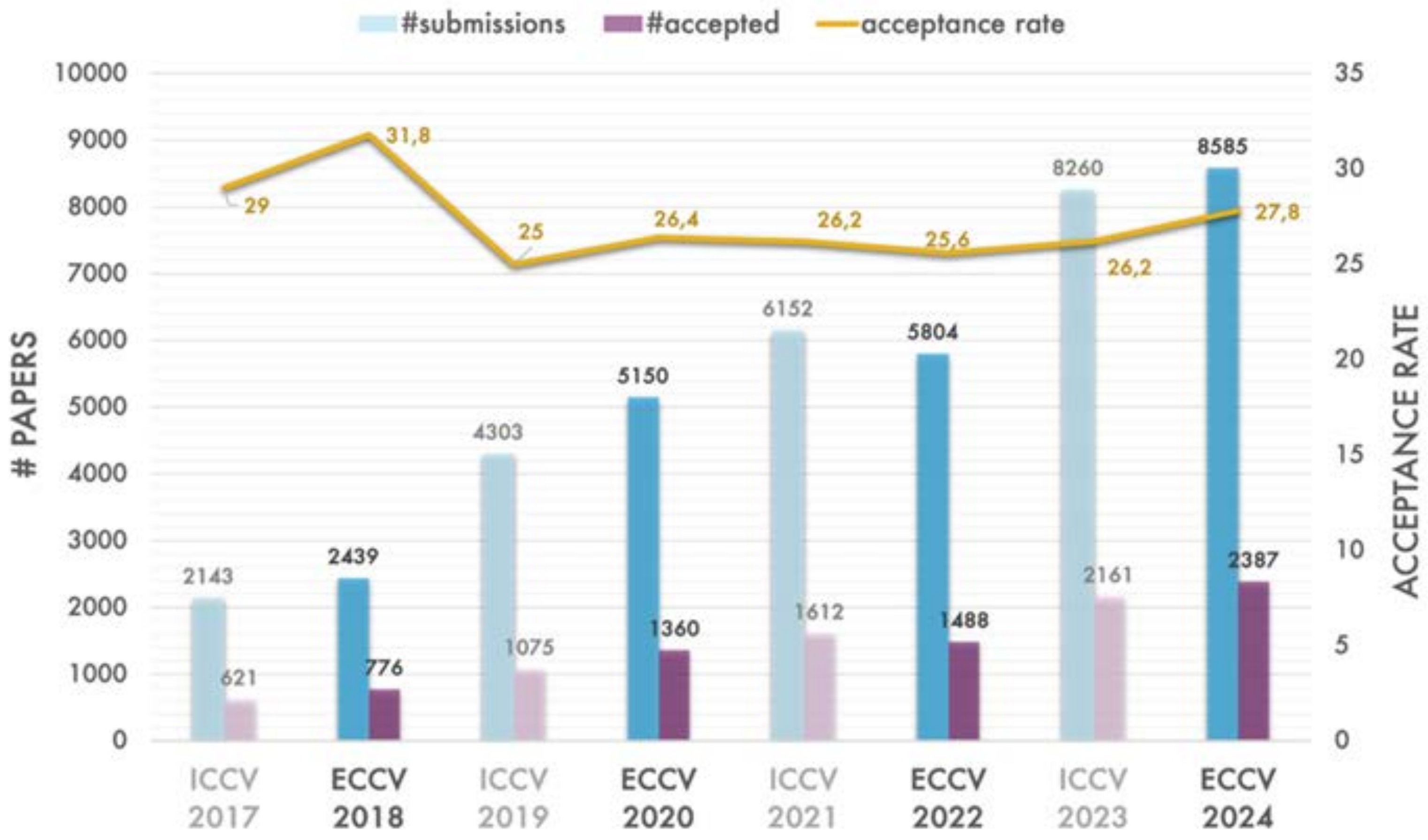
Attendance in numbers



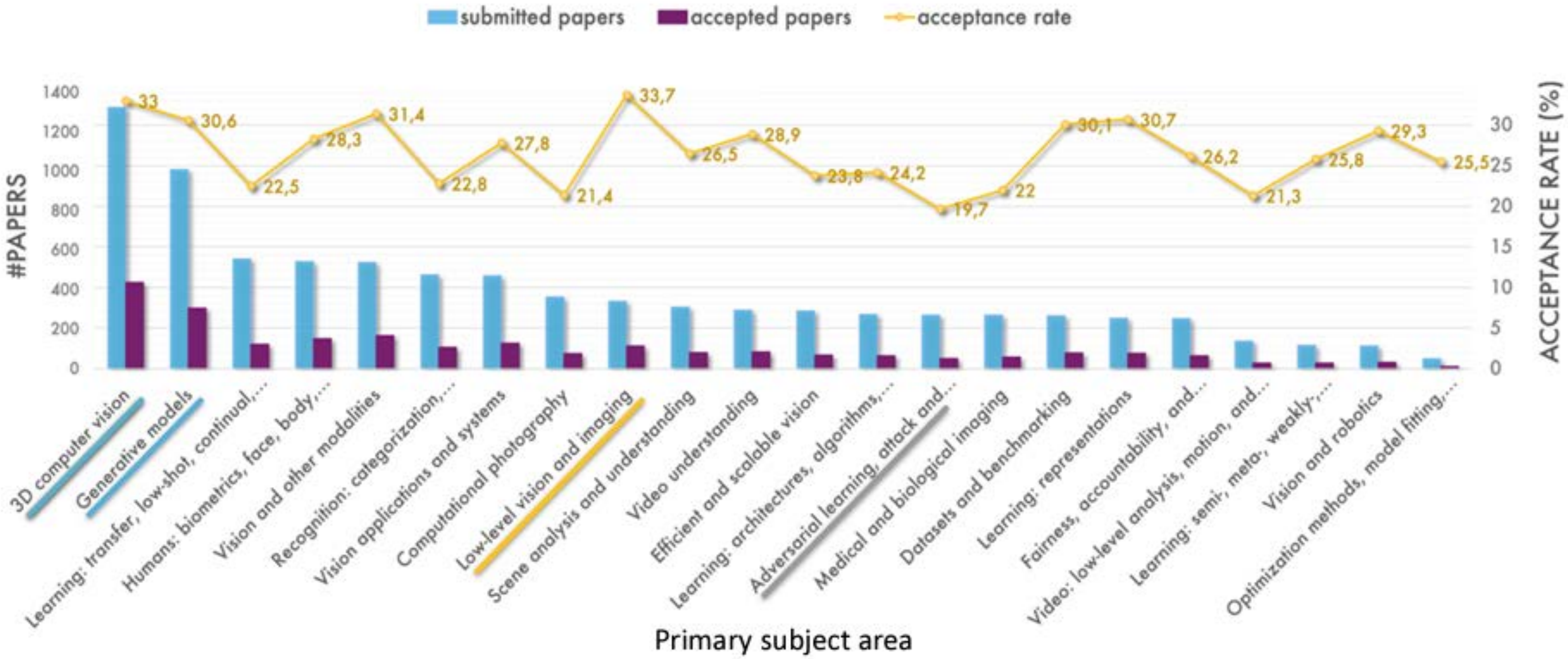


Authors by country

Submission Records



Subject Areas Distribution



ECCV Highlights

Generation

Avatar generation: Synthesia

https://www.synthesia.io/features/avatars

Generation

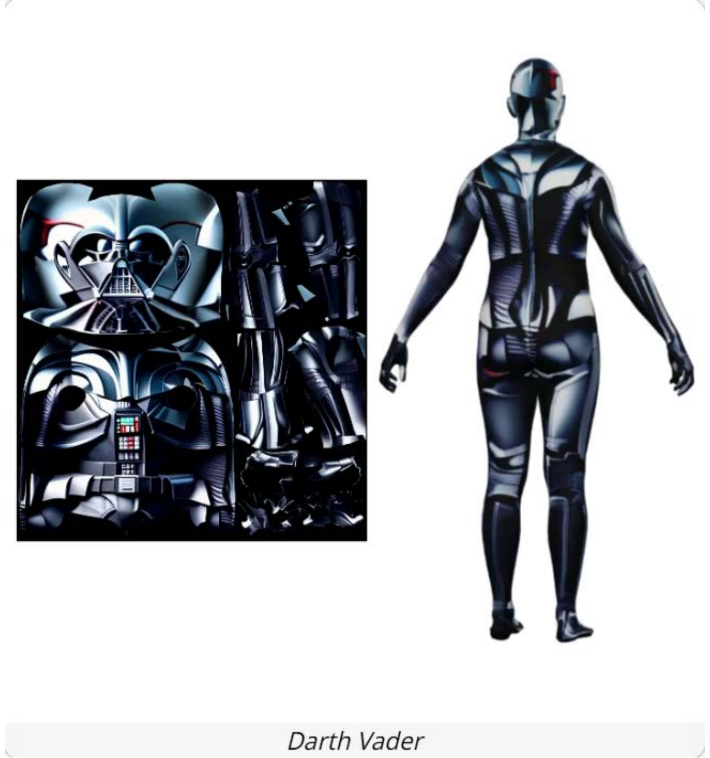
https://ggxxii.github.io/texdreamer/

3D human texture generation: TexDreamer

Image-to-Texture

Input

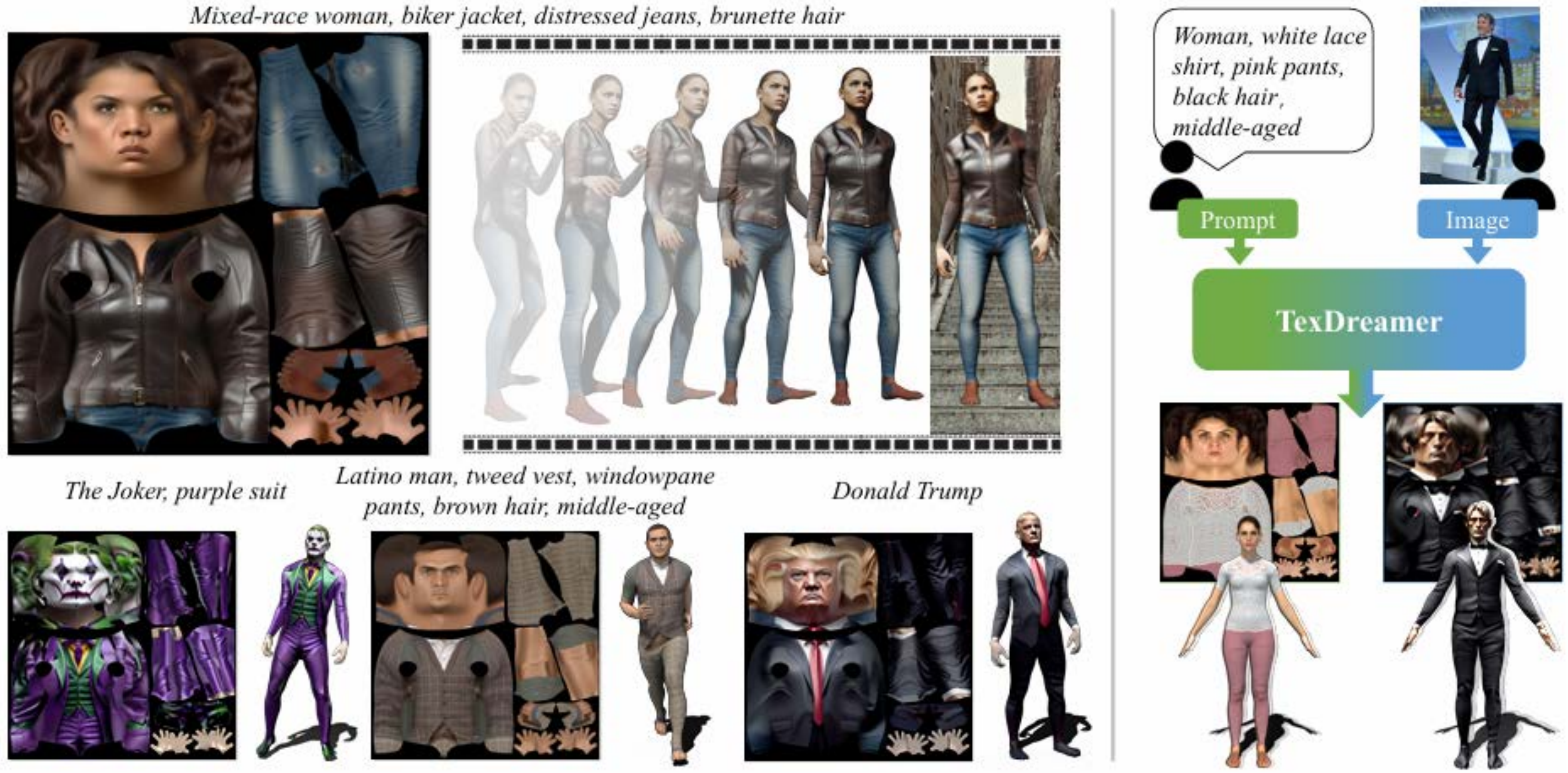
Text-to-Texture



TexDreamer: Towards Zero-Shot High-Fidelity 3D Human Texture Generation

Generation

3D human texture generation: TexDreamer

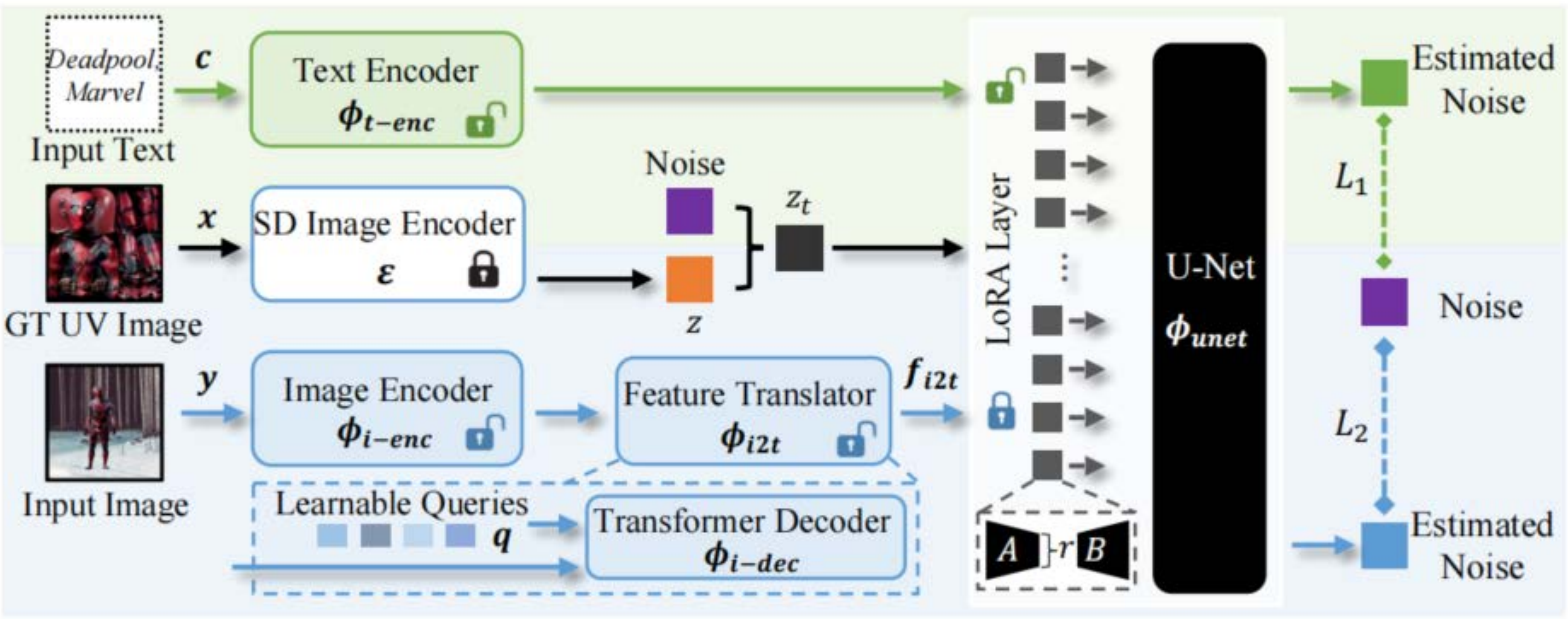


Left: Overview of the ATLAS dataset. Right: Basic structure of TexDreamer.

TexDreamer: Towards Zero-Shot High-Fidelity 3D Human Texture Generation

https://ggxxii.github.io/texdreamer/

Generation

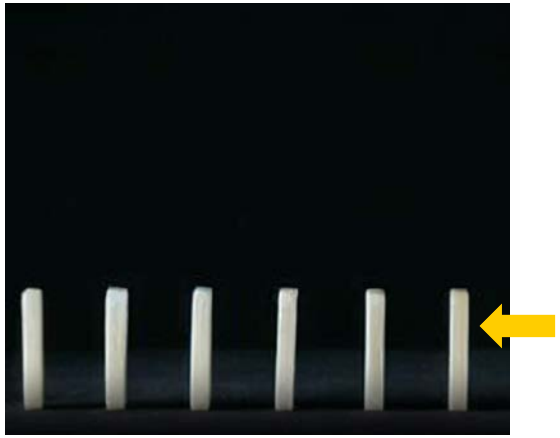


Structure of TexDreamer. Two training stages are conducted.
For T2UV (green), use L1 loss to optimize the text encoder and U-Net.
For I2UV (blue), the feature translator map the input image feature to a conditional textual feature, and use them as conditions during training process.

TexDreamer: Towards Zero-Shot High-Fidelity 3D Human Texture Generation

Generation

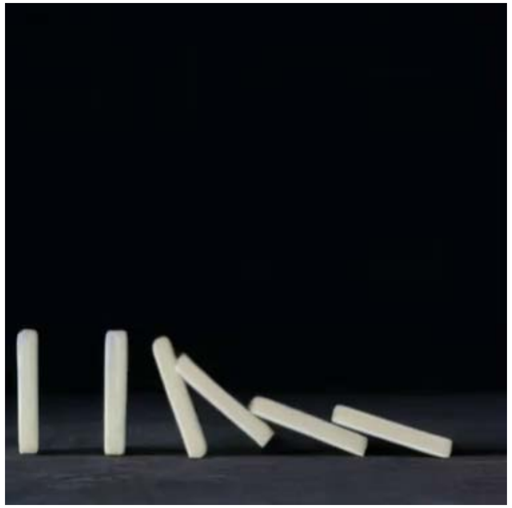
Video Generation: PhysGen



Initial state

SOTA

PhysGen

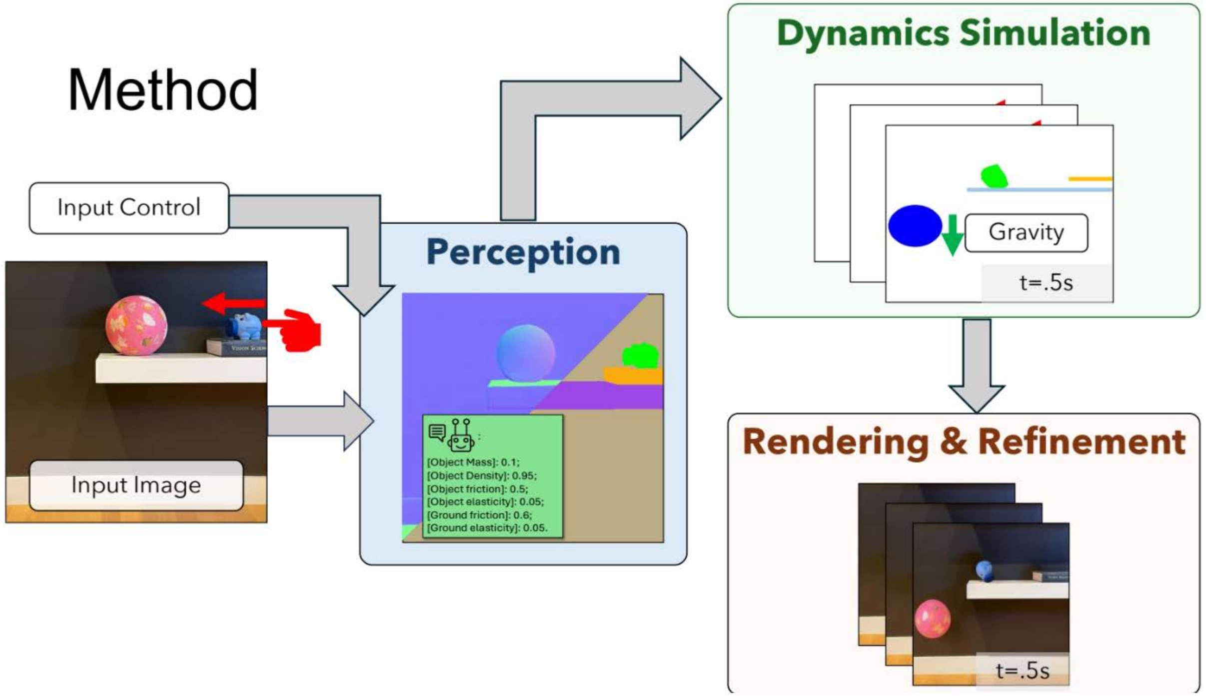


PhysGEN:rigid-body physics-grounded image-to-video generation

https://stevenlsw.github.io/physgen/

Generation

Video Generation: PhysGen

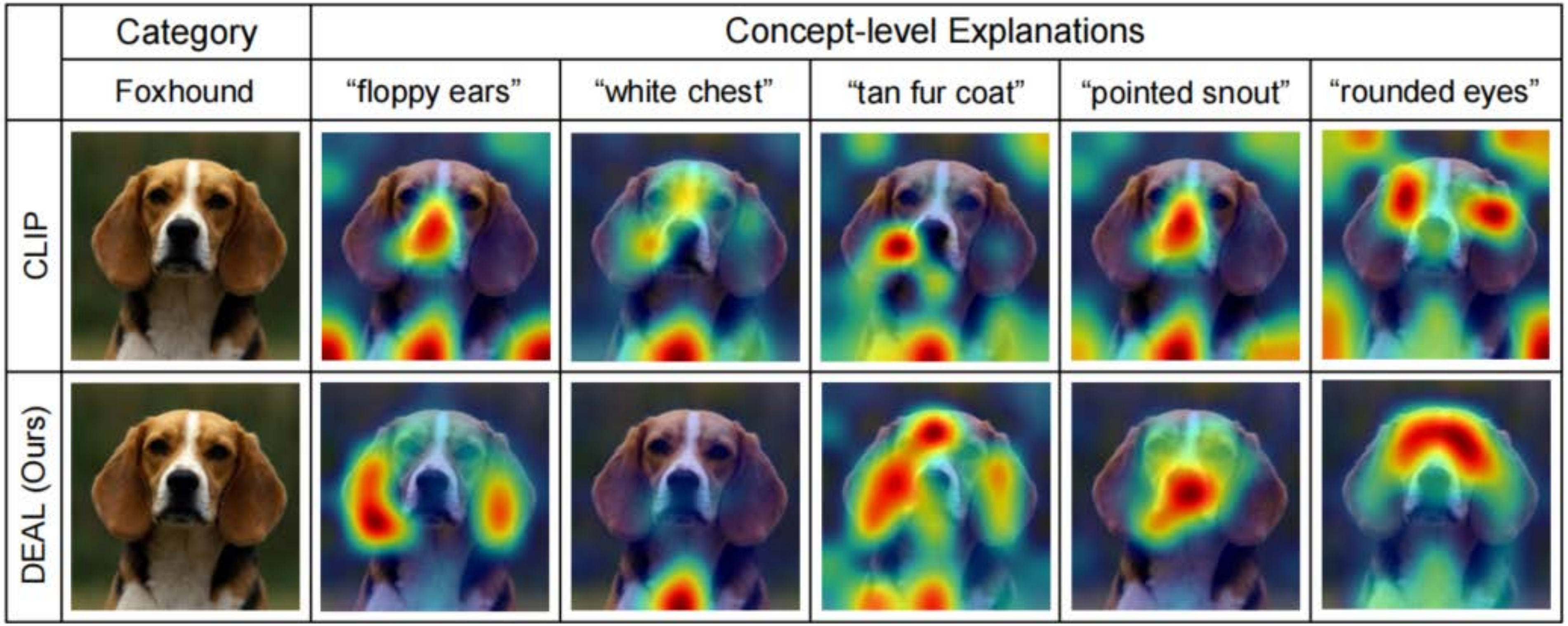


PhysGEN:rigid-body physics-grounded image-to-video generation

https://stevenlsw.github.io/physgen/

Vision-Language Models

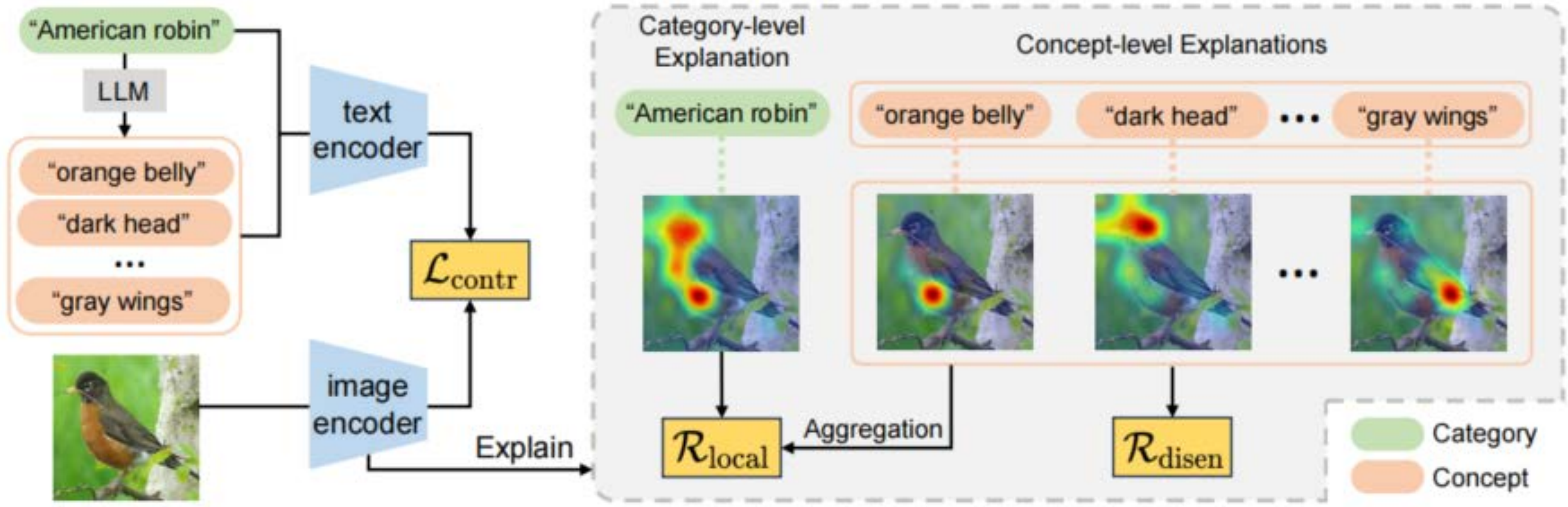
Fine-grained prediction of VLMs



!!!VLMs cannot disentangle and localize fine-grained visual evidence

DEAL: Disentangle and Localize Concept-Level Explanations for VLMs

Vision-Language Models



1. Query the Large Language Model (LLM) and obtain text embeddings
2. Calculate the explanations(GradCAM) w.r.t. the category name and each of the concepts, then constrain the disentanglement and localization with the contrastive learning process

$$\min_{f \in \mathcal{F}} \text{Risk}(f) := \mathbb{E}_{(I,T)\sim P}\left[\mathcal{L}_{\text{contr}}(f(I,T))\right] \quad \triangleleft \textbf{ Contrast}$$

$$\text{s.t. } \text{Dist}(g(\texttt{[CONCEPT]}), g(\texttt{[CONCEPT]}')) \geq \epsilon, \quad \triangleleft \textbf{ Disentangle}$$

$$\text{Dist}(\textstyle\sum g(\texttt{[CONCEPT]}), g(\texttt{[CATEGORY]})) \leq \delta. \quad \triangleleft \textbf{ Localize}$$

Vision-Language Models

Metrics	Models	#Param.	Datasets					
			ImageNet	CUB	Food101	OxfordPets	EuroSAT	Avg.
Concept-level Explanation Disentanglability ↑	CLIP [50]	151M	0.361	0.596	0.487	0.363	**0.192**	0.400
	FLAVA [60]	241M	0.298	0.541	0.463	0.429	0.115	0.369
	DeCLIP [31]	186M	0.032	0.071	0.042	0.056	0.013	0.043
	PyramidCLIP [15]	153M	0.048	0.116	0.080	0.085	0.026	0.071
	CLIPpy [2]	196M	0.300	0.557	0.449	0.461	0.162	0.386
	DEAL (Ours)	151M	**0.397**	**0.608**	**0.501**	**0.475**	**0.192**	**0.435**
Concept-level Explanation Localizability ↑	CLIP [50]	151M	0.633	0.638	0.511	0.762	0.423	0.593
	FLAVA [60]	241M	0.630	0.650	0.589	0.668	0.361	0.580
	DeCLIP [31]	186M	0.366	0.367	0.318	0.369	0.295	0.343
	PyramidCLIP [15]	153M	0.662	0.672	0.644	0.700	0.302	0.596
	CLIPpy [2]	196M	0.612	0.614	0.656	0.657	0.345	0.577
	DEAL (Ours)	151M	**0.673**	**0.718**	**0.660**	**0.809**	**0.444**	**0.661**
Prediction Accuracy (%)	CLIP [50]	151M	†63.2	52.6	†84.4	†87.0	†41.1	65.7
	FLAVA [60]	241M	55.1	49.4	79.7	57.7	28.2	54.0
	DeCLIP [31]	186M	‡66.2	35.9	57.0	59.1	27.3	49.1
	PyramidCLIP [15]	153M	46.0	43.8	49.3	36.0	20.0	39.0
	CLIPpy [2]	196M	45.3	18.9	53.8	47.5	18.9	36.9
	DEAL (Ours)	151M	**70.8**	**69.6**	**86.9**	**89.3**	**77.4**	**78.8**

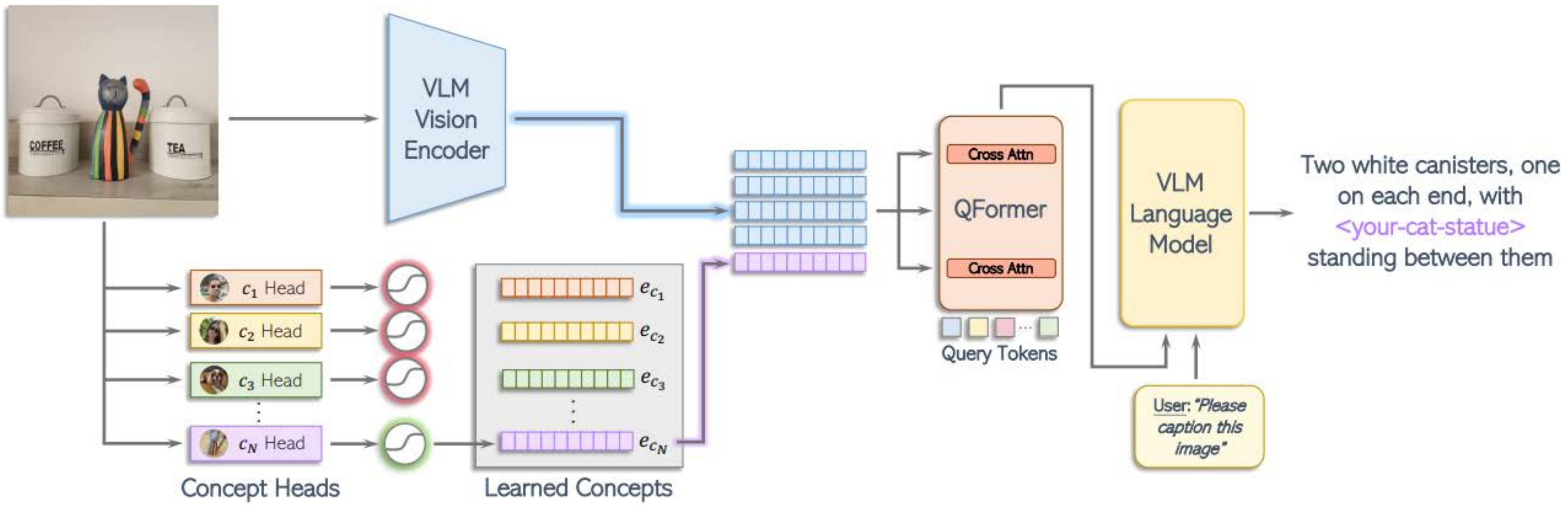
Experiment results on various datasets

Vision-Language Models

Personalized VLMs:
MyVLM

MyVLM: Personalizing VLMs for User-Specific Queries

Vision-Language Models



Step 1: Feature Extraction
Extract the frozen image features from the VLM's vision encoder.

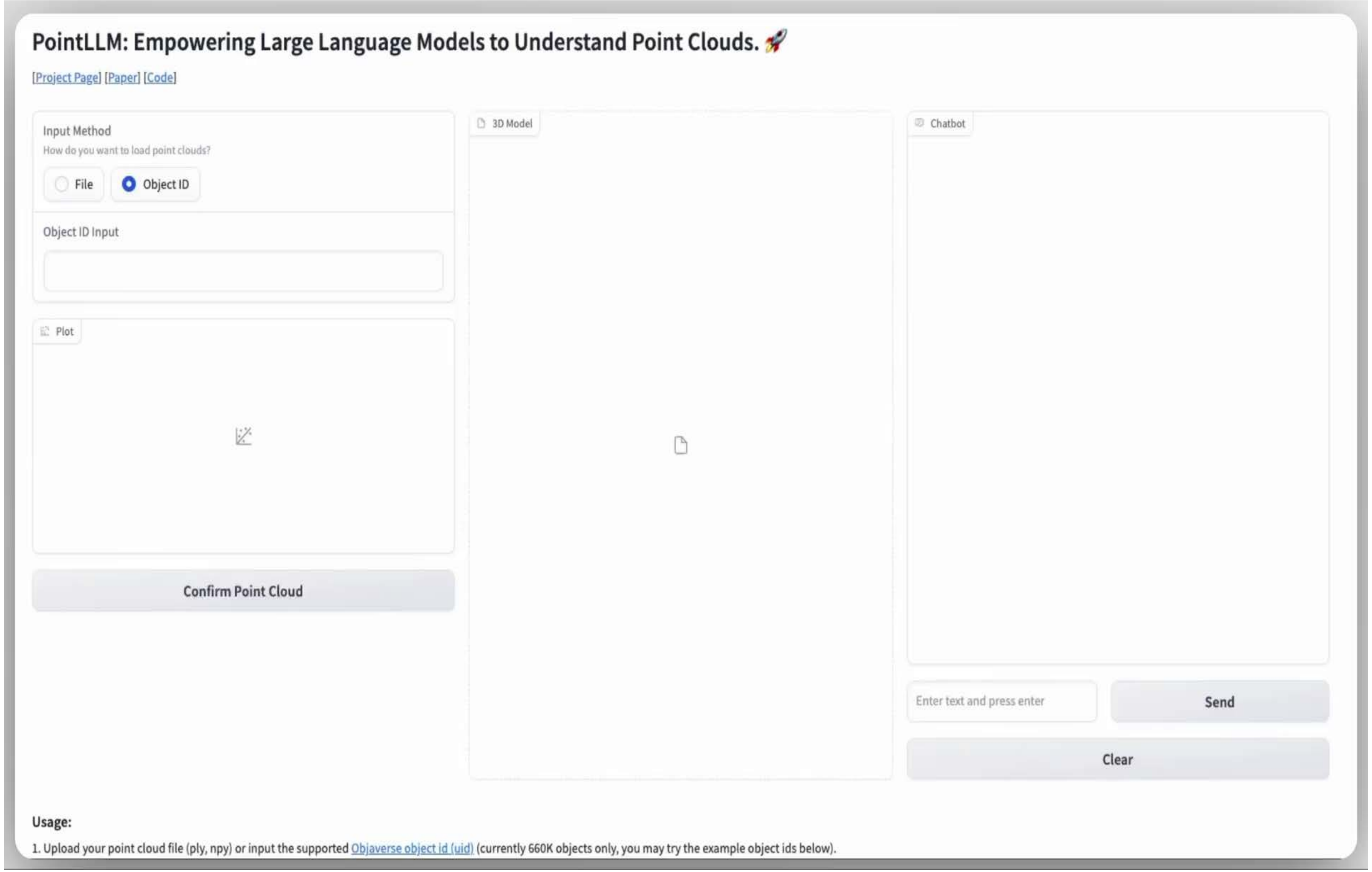
Step 2: Recognizing the Concept
Using a set of concept heads, each designed to recognize the presence of a user-specific concept within the image.

Step 3: Communicating the Concept
Train a QFormer to represent the concept and guide the LLM to incorporate the concept into its personalized response.

MyVLM: Personalizing VLMs for User-Specific Queries

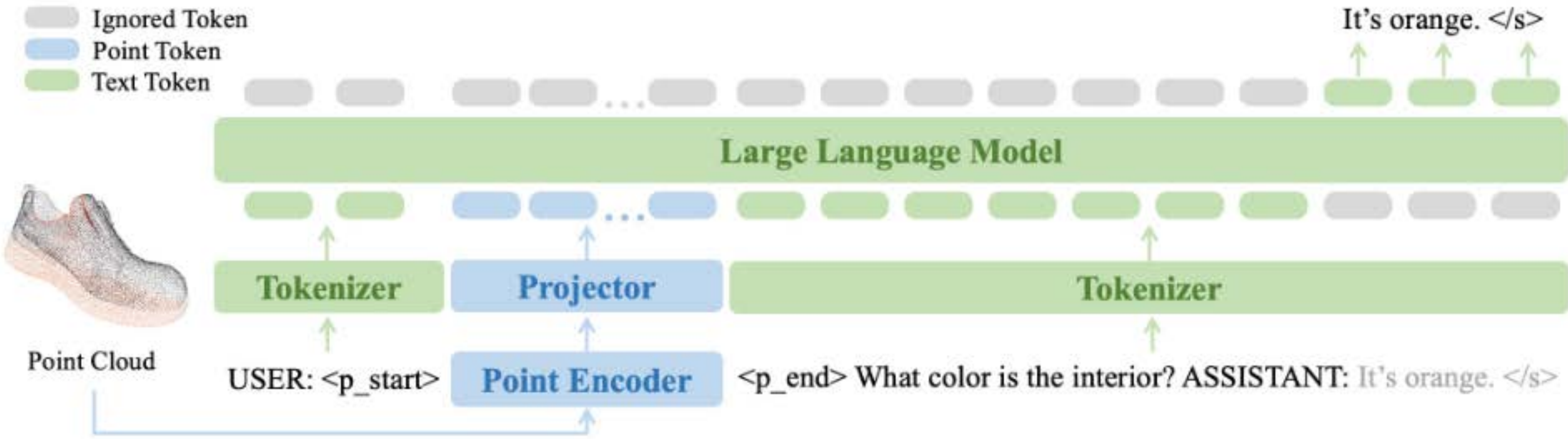
Vision-Language Models

Domain transfer:
PointLLM



PointLLM: Empowering Large Language Models to Understand Point Clouds

Vision-Language Models



An overview of PointLLM.

First stage,freeze the point encoder and the LLM, train only the projector, aiming to align point features with the text token space effectively.

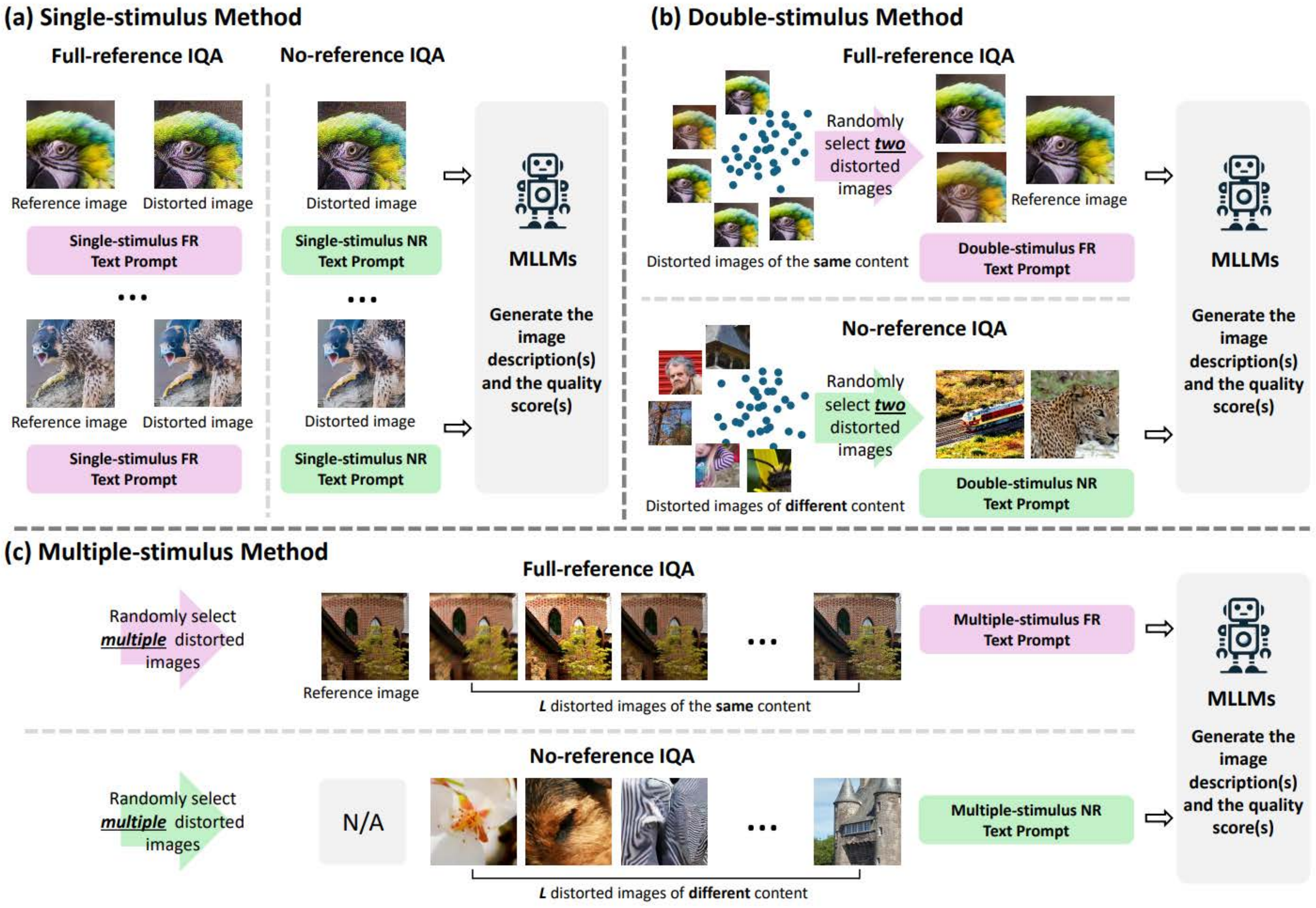
Second stage, freeze the point cloud encoder, training the projector and the LLM. This second stage helps the model build its ability to understand and respond to complex instructions including point cloud data.

3D object captioning results on Objaverse. Evaluation encompasses human (correctness, hallucination, precision) and LLM assessments

Model	Corr.	Hallu.↓	Prec.	GPT-4	S.-BERT	SimCSE	B-1.	R-L.	MET.
InstructBLIP-7B [8]	2.56	0.77	76.99	45.34	47.41	48.48	4.27	8.28	12.99
InstructBLIP-13B [8]	2.58	1.13	69.56	44.97	45.90	48.86	4.65	8.85	13.23
LLaVA-7B [31]	2.76	0.86	76.30	46.71	45.61	47.10	3.64	7.70	12.14
LLaVA-13B [31]	2.43	0.86	73.97	38.28	46.37	45.90	4.02	8.15	12.58
3D-LLM [21]	1.77	1.16	60.39	33.42	44.48	43.68	16.91	19.48	**19.73**
PointLLM-7B	3.04	**0.66**	**82.14**	44.85	47.47	48.55	3.87	7.30	11.92
PointLLM-13B	**3.10**	0.84	78.75	**48.15**	**47.91**	**49.12**	3.83	7.23	12.26
PointLLM-13B*	2.12	0.39	84.39	44.27	50.15	50.83	**17.09**	**20.99**	16.45
Human	2.67	0.22	92.46	100.00	100.00	100.00	100.00	100.00	100.00

Vision-Language Models

VLMs as evaluation tool



Three standardized psychophysical testing procedures for image quality assessment(IQA). (a) Single stimulus method. (b) Double-stimulus method. (c) Multiple-stimulus method

Instantiations of systematic prompting strategies

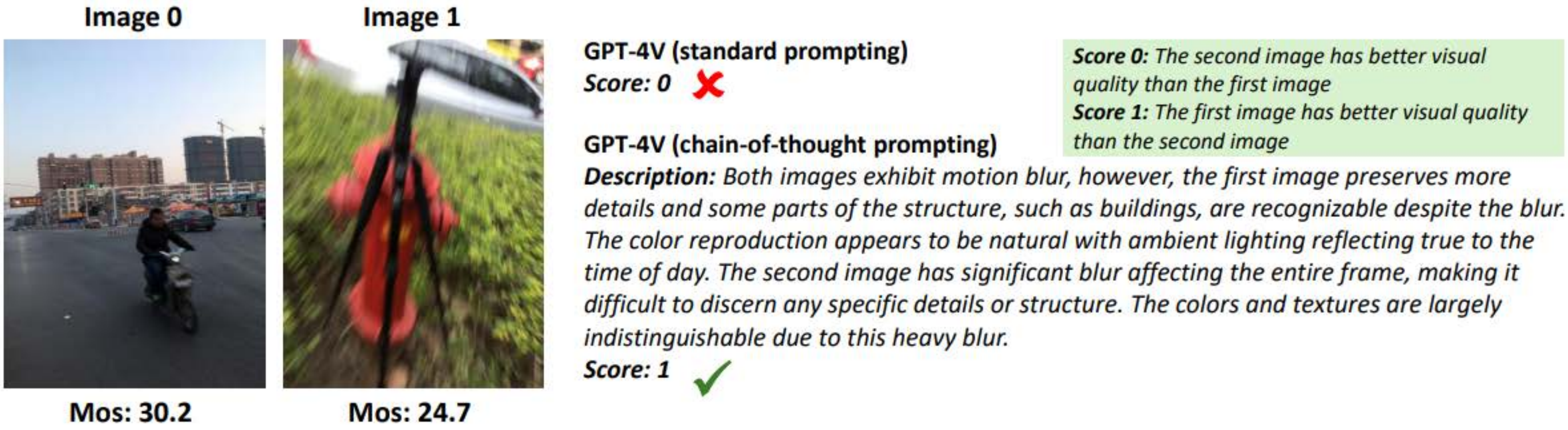
Method	FR IQA				NR IQA		
	FR-KADID	Aug-KADID	TQD	SPCD	NR-KADID	SPAQ	AGIQA-3K
Single-stimulus Method							
LLaVA-v1.6-S	0.227	0.013	0.180	0.001	0.262	0.544	**0.614**
mPLUG-Owl2-S	0.285	0.218	0.228	0.081	0.126	0.467	0.279
InternLM-XC2-VL-S	0.274	0.272	0.299	0.009	0.252	0.794	0.512
GPT-4V-S	**0.745**	**0.786**	**0.773**	**0.098**	**0.467**	**0.860**	0.420
LLaVA-v1.6-C	0.164	0.300	0.226	**0.174**	0.151	0.550	0.580
mPLUG-Owl2-C	0.387	0.361	0.278	0.122	0.179	0.455	0.409
InternLM-XC2-VL-C	0.237	0.306	0.167	0.063	0.306	0.649	0.507
GPT-4V-C	**0.809**	**0.782**	**0.809**	0.121	**0.517**	**0.869**	**0.677**
LLaVA-v1.6-I	0.249	0.194	0.222	**0.147**	0.116	0.019	0.061
mPLUG-Owl2-I	0.373	0.373	0.246	0.047	0.017	0.083	0.409
InternLM-XC2-VL-I	0.380	0.241	0.204	0.087	0.188	0.342	0.461
GPT-4V-I	**0.771**	**0.753**	**0.738**	0.028	**0.590**	**0.845**	**0.650**
Double-stimulus Method							
LLaVA-v1.6-S	0.387	0.396	0.390	0.113	0.270	0.430	0.234
mPLUG-Owl2-S	0.435	0.307	0.350	**0.117**	0.126	0.157	0.020
InternLM-XC2-VL-S	0.309	0.408	0.440	0.042	0.267	0.690	0.555
GPT-4V-S	**0.679**	**0.743**	**0.655**	0.031	**0.552**	**0.834**	**0.599**
LLaVA-v1.6-C	0.332	0.355	0.257	0.109	0.124	0.065	0.174
mPLUG-Owl2-C	0.409	0.334	0.318	0.013	0.199	0.122	0.130
InternLM-XC2-VL-C	0.332	0.411	0.267	**0.131**	0.165	0.556	0.546
GPT-4V-C	**0.818**	**0.830**	**0.786**	0.124	**0.639**	**0.881**	**0.771**
LLaVA-v1.6-I	0.379	**0.396**	0.324	0.032	0.169	0.128	0.156
mPLUG-Owl2-I	0.257	0.257	0.169	0.083	0.078	0.164	0.120
InternLM-XC2-VL-I	0.348	0.376	**0.379**	**0.144**	0.034	0.108	0.123
GPT-4V-I	**0.470**	0.244	0.340	0.122	**0.531**	**0.761**	**0.714**
Multiple-stimulus Method							
LLaVA-v1.6-S	0.349	0.351	0.315	**0.241**	0.169	0.221	0.210
mPLUG-Owl2-S	0.385	0.428	0.297	0.104	0.124	0.061	0.228
InternLM-XC2-VL-S	0.484	0.420	0.241	0.015	0.047	0.044	0.154
GPT-4V-S	**0.824**	**0.844**	**0.747**	0.037	**0.397**	**0.715**	**0.461**
LLaVA-v1.6-C	0.292	0.424	0.288	0.043	0.227	0.111	0.122
mPLUG-Owl2-C	0.377	0.406	0.376	**0.126**	0.214	0.166	0.084
InternLM-XC2-VL-C	0.500	0.466	0.273	0.038	0.031	0.037	0.148
GPT-4V-C	**0.761**	**0.806**	**0.754**	0.036	**0.537**	**0.817**	**0.679**
LLaVA-v1.6-I	0.337	0.380	0.356	**0.203**	0.152	0.033	**0.241**
mPLUG-Owl2-I	0.268	0.268	0.377	0.067	**0.196**	0.142	0.121
InternLM-XC2-VL-I	0.489	0.235	0.212	0.046	0.038	0.102	0.114
GPT-4V-I	**0.585**	**0.496**	**0.389**	0.023	0.168	**0.416**	0.201

NR (No-Reference) scenario: Evaluates image quality without a reference image.

FR (Full-Reference) scenario: Evaluates quality by comparing the input image with the reference image.

Experiment results:
Only the closed-source GPT-4V can reasonably reflect human perception of image quality,
but it performs poorly in fine-grained quality distinction and multi-image quality comparison

Mean Opinion Score(MOS). Asking multiple observers to rate the image quality and then calculating the average of ratings.



Results: Different VLMs require different hint systems to achieve optimal performance, suggesting room for further improvement in the field of IQA.

Vision-Language Models

Manuscript restoration

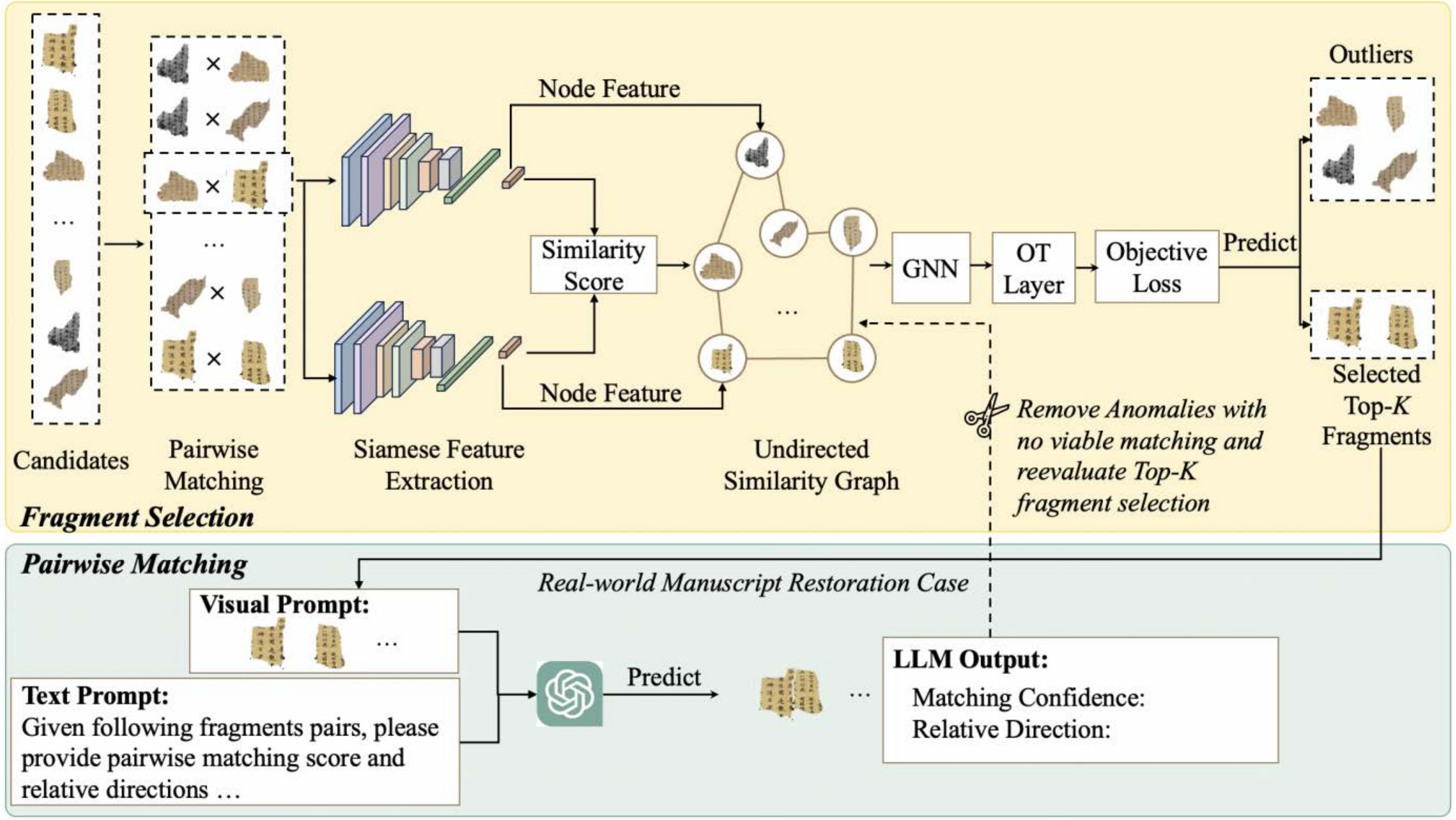
(a) conventional restoration task

Characters evolution

(b) select related small fragments and restore

LLMCO4MR: LLMs-aided Neural Combinatorial Optimization for Ancient Manuscript

Vision-Language Models



LLMCO4MR: LLMs-aided Neural Combinatorial Optimization for Ancient Manuscript

Performance on various Top-K fragments

Top-K	2×2			3×3		
Pool Size	10	15	20	20	25	30
Random Select	0.4000	0.2667	0.2000	0.4500	0.3600	0.3000
Baseline	0.7475	0.3125	0.2575	0.6756	0.5767	0.5389
OT Layer	**0.7800**	0.5825	0.4225	**0.7256**	0.6444	0.6044

Evaluatin using different methods

Top-K	2×2		3×3	
Pool Size	10	15	20	25
GPT-4V [26]	0.3250	0.1750	0.3778	0.2533
LLaVA [21]	0.3150	0.1250	0.2778	0.1556
JigsawNet [19]	0.5250	0.3750	0.4556	0.3222
Papyrus [31]	0.4250	0.3750	0.4778	0.4111
S3-Net [51]	0.3125	0.2575	0.3889	0.2111
CO Solver	0.5750	0.5250	0.5333	0.4667
LLMCO4MR (Ours)	**0.6750**	0.6250	**0.6222**	0.5556