



جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ARTIFICIAL
INTELLIGENCE
INITIATIVE

Imaginative Vision Language Models

Towards human-level imaginative
AI skills
transforming species discovery,
content creation, self-driving cars
and emotional health

Mohamed Elhoseiny
KAUST

Group / Research Focus

Vision-CAIR. Group- *Vision-CAIR stands for Computer Vision- Core AI Research* <https://cemse.kaust.edu.sa/vision-cair>

- Efficient Multimodal Learning / Zero/Few shot Learning (mainly **guided by Language**)
- Affective Vision and Language
- Continual (Zero-Shot)? Learning (**mainly guided by language** / recent foundation models)
- Creative AI and more generally Visual Generative Models (**images, videos, visual stories, recently 3D**)

Students

PhD Fall23 graduates

PhD-Alumni 2023



Eslam Bakr



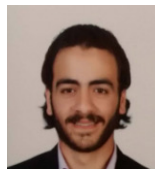
Kilich Haydarov



Wenxuan Zhang



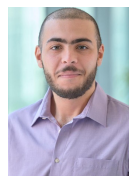
Xiaoqian Shen



Youssef Mohamed



Faizan Khan

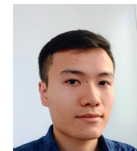


Habib Slim



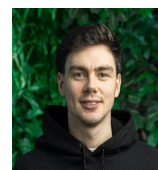
Jun Chen

-> RS @ Meta



Deyao Zhu

-> RS@ Bytedance



Ivan Skorokhodov

RS @ Snap Research

Interns



Divyansh Jha

2020-2021



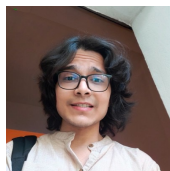
Aashiq Muhamed

2021



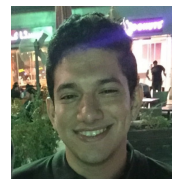
Lia Coleman

2019-2020



Aniket Agarwal

2020-2021



Sherif Abdelkarim

2020-2021



Ujjwal Upadhyay

2021-2022

MS-Alumni



Kai Yi



Yuchen Li

Imaginative AI

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short orange bill. The bird's plumage is dark above and white below.



e.g, ICCV13, CVPR17, CVPR18, ICCV19, ICLR21, CVPR22 , ECCV22, CVPR23, ICCV23



Makes me feel like birds are flying over my head.

IMAGINE TO CREATE



e.g, ICCV17, AAAI18, ECCVW18, ICCV19, CVPR21, ICCV21 , CVPR22, ICCV22

IMAGINE TO FEEL

Affective Visual Art

ArtEmis-v1@CVPR21, ArtEmis-v2@CVPR22

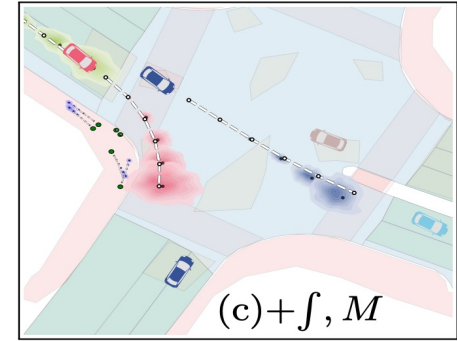
ArtELingo@EMNLP22 (long paper)

WAGA@ICCV21, CWAN@ICCV22



The pale color palette of this watercolor painting is very relaxing. I can imagine myself sitting by the water *listening* to the birds.

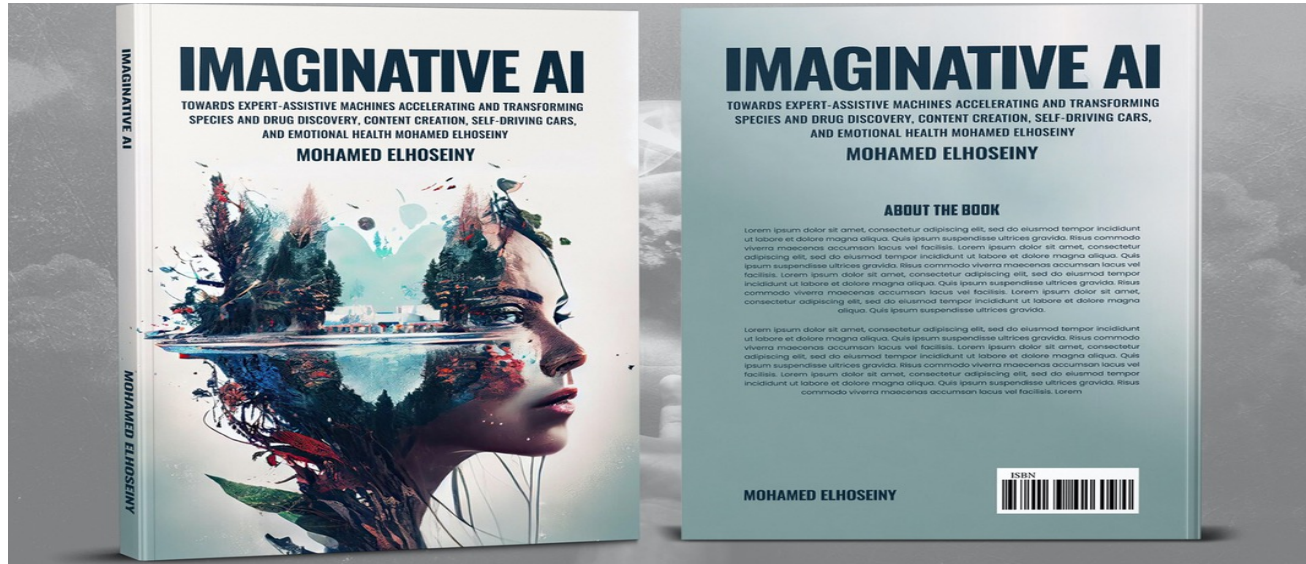
IMAGINE TO DRIVE



e.g, CVPR2020, ICLR2021, CoRL21, ECCV22, ICLR23

Imaginative AI

Imaginative AI: Towards human-level imaginative AI skill transforming species discovery, content creation, self-driving cars, and emotional health



Book in progress (feedback is most welcome)

Introduced Datasets (Recognition)

Imagine to See

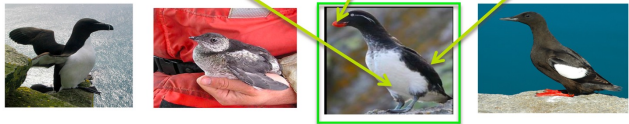
CUB-Wiki (ICCV13) and NAB-Wiki (CVPR17)

200 Wikipedia class description, 500 Wikipedia class descriptions

Parakeet Auklet is a small bird that has an short orange bill. The bird's plumage is dark above and white below.

[Elhoseiny](#), Saleh, Elgammal,
Write a Classifier

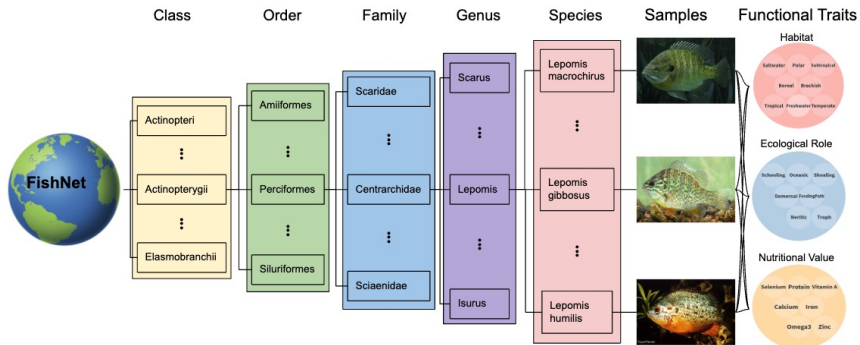
[Elhoseiny*](#), [Zhu*](#), Zhang, Elgammal,
Link the head to the "peak":



FishNet (ICCV'23)

94,523 images, 17356 species, 22 functional traits

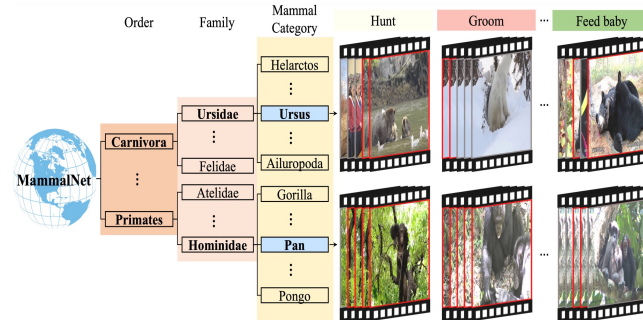
F Khan, X Li, A Temple, [Elhoseiny](#)



MammalNet (CVPR23)

500 video hours, 172 mammal categories, 12 behaviors

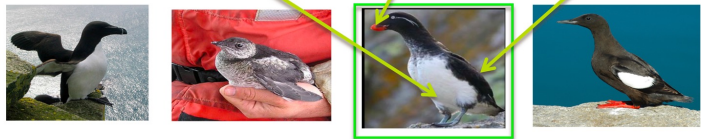
J Chen, M Hu, D Coker, M Berumen, B Costelloe, S Beery, A Rohrbach, [Elhoseiny](#)



Introduced Datasets (3D Vision)

2D (PART-LEVEL)

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



e.g, ICCV13, CVPR17, CVPR18, ICCV19, ICLR21,
CVPR22

3D (PART-LEVEL)

3D CoMPaT (ECCV22, oral)

7.2 million stylized models (unique composition)

Composition of Materials on Parts of 3D Things

Yuchen Li*, Ujjwal Upadhyay*, Habib Slim*, Ahmed Abdelreheem, Arpit Prajapati,
Suhail Pothigara Peter Wonka, Mohamed Elhoseiny

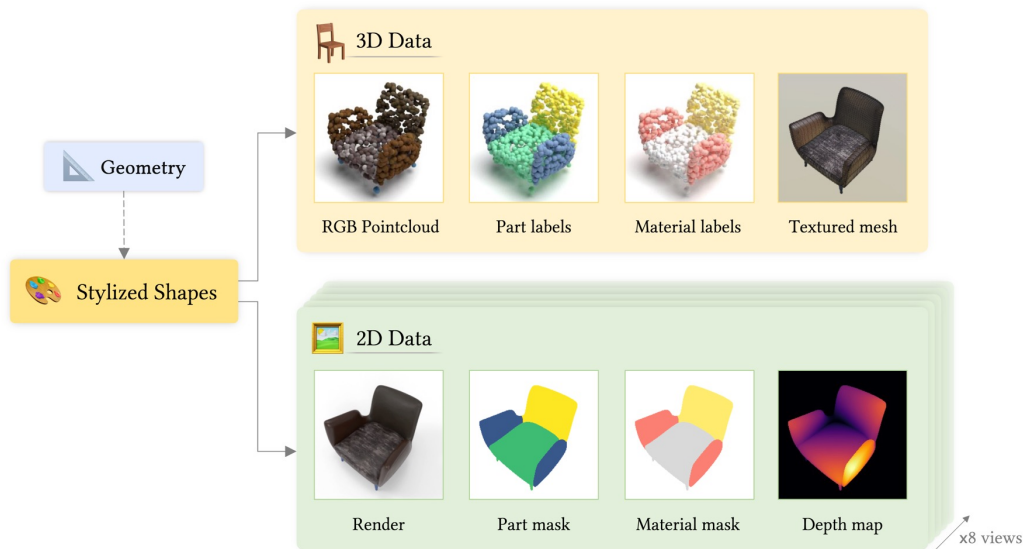


Introduced Datasets (3D Vision)

3D CoMPaT (ECCV22, oral)

7.2 million stylized models (unique composition)

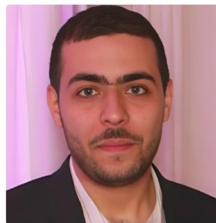
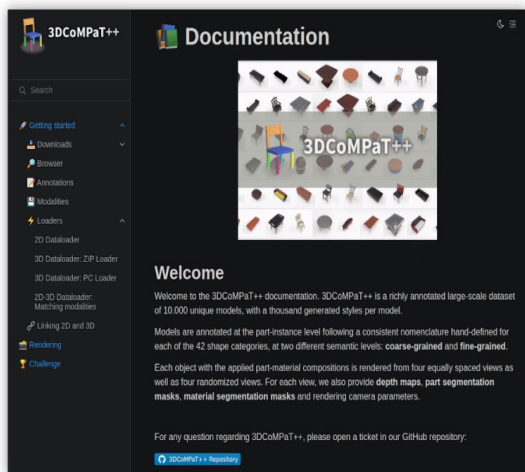
3DCoMPaT: A 2D/3D dataset of stylized CAD models for compositional 3D vision.



Introduced Datasets (3D Vision)

3DCoMPaT++ Challenge

C3DV



Habib Slim
Ph.D. Student
KAUST



Mahmoud Ahmed
Research Student
KAUST



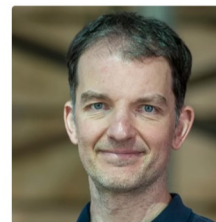
Mohamed Ayman
Research Student
KAUST



Xiang Li
Postdoctoral Researcher
KAUST



Yuchen Li
Ph.D. Student
KAUST



Peter Wonka
Professor
KAUST



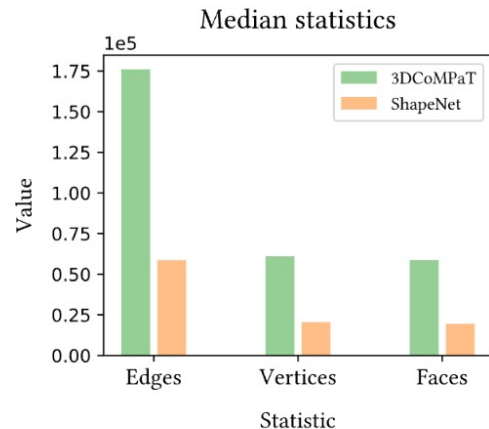
Mohamed Elhoseiny
Assistant Professor
KAUST

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Introduced Datasets (3D Vision)

3DCoMPaT++

Dataset	📦 Shapes				🍷 Materials		🧩 Parts		
	Count	Stylized	Classes	Source	Count	Classes	Parts/Shape	Multi-level	Instances
ModelNet [2]	128K	✗	662	modelled	✗	✗	✗	✗	✗
ShapeNet-Core [1]	51,3K	✗	55	modelled	?	?	✗	✗	✗
ShapeNet-Sem [1]	12K	✗	270	modelled	?	?	✗	✗	✗
PhotoShape [14]	5,8K	29K	1	modelled	658	8	✗	✗	✗
GSO [15]	1K	✗	17	scanned	?	?	✗	✗	✗
OmniObject3D [4]	6K	✗	190	scanned	?	?	✗	✗	✗
ObjectNet3D [16]	44,2K	✗	100	modelled	✗	✗	✗	✗	✗
3D-Future [6]	9,9K	✗	15	modelled	?	15	✗	✗	✗
ABO [5]	148K	✗	98	modelled	✗	✗	✗	✗	✗
Objaverse-XL [8]	10.2M	✗	?	modelled	?	?	✗	✗	✗
ShapeNet-Part [11]	31,9K	✗	16	modelled	✗	✗	2.99	✗	✗
ShapeNet-Mats [12]	3,2K	✗	3	modelled	?	6	6.2	✗	✗
PartNet [3]	26,7K	✗	24	modelled	✗	✗	18	✓	✓
3DCoMPaT++	10K	10M	41	modelled	293	13	5.21	✓	✓



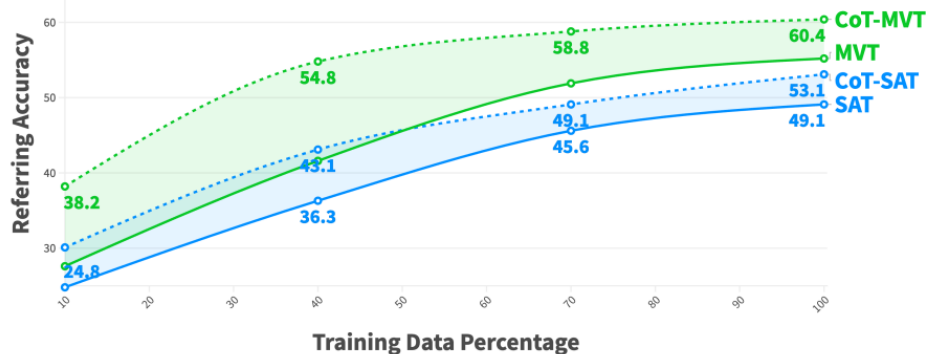
Introduced Datasets (Language & 3D Vision)

CoT3DREF: CHAIN-OF-THOUGHTS DATA-EFFICIENT 3D VISUAL GROUNDING, Arxiv, 2023

Eslam Mohamed Bakr, Mohamed Ayman, Mahmoud Ahmed, Habib Slim, Mohamed Elhoseiny



CoT Performance on NR3D of Referit3D



ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes [ECCV 2020 (Oral)]

[Panos Achlioptas](#), [Ahmed Abdelreheem](#), [Fei Xia](#), [Mohamed Elhoseiny](#), [Leonidas Guibas](#)

Introduced Datasets (Affective Vision and Language)

ArtEmis (CVPR21)

450K utterances

ArtEmis: Affective Language for Visual Art



Panos Achlioptas¹

panos@cs.stanford.edu

Maks Ovsjanikov²

maks@lix.polytechnique.fr



Kilichbek Haydarov³

kilichbek.haydarov@kaust.edu.sa



Mohamed Elhoseiny^{3,1}

mohamed.elhoseiny@kaust.edu.sa

Leonidas Guibas¹

guibas@cs.stanford.edu



¹Stanford University

²LIX, Ecole Polytechnique, IP Paris

³King Abdullah University of Science and Technology (KAUST)

ArtEmis 2.0 (CVPR22)

260K additional utterances



It is Okay to Not Be Okay:

Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection

www.artemisdataset-v2.org



جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology



Youssef Mohamed



Faizan Khan



Kilichbek Haydarov



Mohamed Elhoseiny













ArtELingo (EMNLP 2022, long paper)

ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on

Diversity over Language and Culture

Youssef Mohamed, Shyma Yaser Alhuwaidar, Mohamed Abdelfattah, Feifan

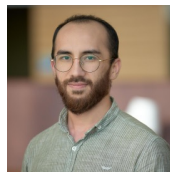
Li, Kenneth Ward Church, Xiangliang Zhang and **Mohamed Elhoseiny**

a)	 <p>شلال طبيعي جميل. مشاعر النمو والحيوية والطاقة موجودة. Translation: Beautiful natural waterfall. Feelings of growth, vitality and energy.</p> <p>Excitement Arabic</p> 	<p>The water that's rushing downward looks like a bride's wedding veil.</p> <p>Awe English</p> 	<p>瀑布就像四蹄生风的白马如潮水涌来，非常的壮观 Translation: The waterfall is like a white horse and wind, it is spectacular.</p> <p>Contentment Chinese</p> 
b)	 <p>Translation: Girls sitting with their mother outside the house, exchanging love and affection, pigeons flying over a tree.</p> <p>Contentment Arabic</p> 	<p>The women relaxing while birds are flying about makes me feel relaxed and calm as well.</p> <p>Contentment English</p> 	<p>Translation: Three sisters lying on a bench and watching the birds fly comfortably.</p> <p>Contentment Chinese</p> 
c)	 <p>Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction.</p> <p>Contentment Arabic</p> 	<p>The trees are dead and exposing their roots due to erosion and lack of water.</p> <p>Sadness English</p> 	<p>Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed.</p> <p>Sadness Chinese</p> 

Introduced Datasets (Affective Vision and Language)

Affective Visual Dialog (Arxiv23)

50,000 dialogs, 27,180 working hours : thanks Jack Urbanek, Mephisto team)



Kilichbek Haydarov, Xiaoqian shen, Avinash Madasu



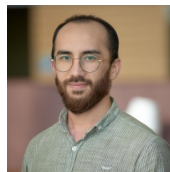
Jia Li



Introduced Datasets (Affective Vision and Language)

Affective Visual Dialog (Arxiv23)

50,000 dialogs, 27,180 working hours : thanks Jack Urbanek, Mephisto team)



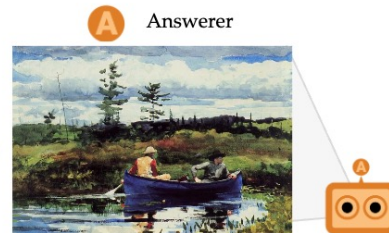
Kilichbek Haydarov, Xiaoqian shen, Avinash Madasu



Jia Li



Pos: The men seems to be happy and enjoying the time together here. undisturbed.



Neg: The man is looking over his shoulder because he sees a threat

Affective emotion explanations:



The two men are thinking about how to go home, I feel scared because the clouds are coming and it looks like the rain will fall and storms will come into the lake.



fear



No, It is not that I'm thinking it is cool weather and it seems they doing boating in the lake and enjoying their time



contentment



contentment

The men are sitting in the boat and seeing somewhere very interesting I think they are waiting for their friend to sail in the lake in the beautiful whether alone with the lake is surrounded by a lot of trees which looks very excellent so it gives me a good feeling



<https://affective-visual-dialog.github.io/>

Introduced Metrics (text->image)

HRS-Bench

Holistic, Reliable and Scalable

Benchmark for Text-to-Image Models



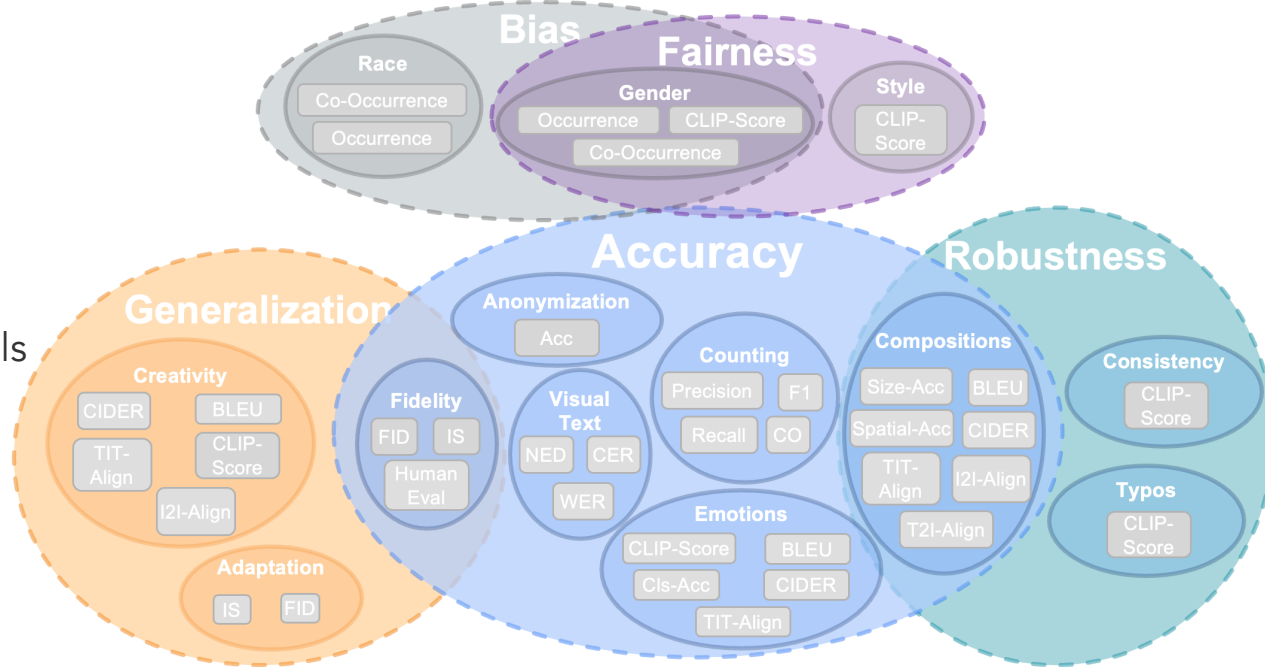
Eslam Bakr



Xiaoqian Shen



Faizan Khan



Inspired by HELM bench by Stanford for evaluating Language Models



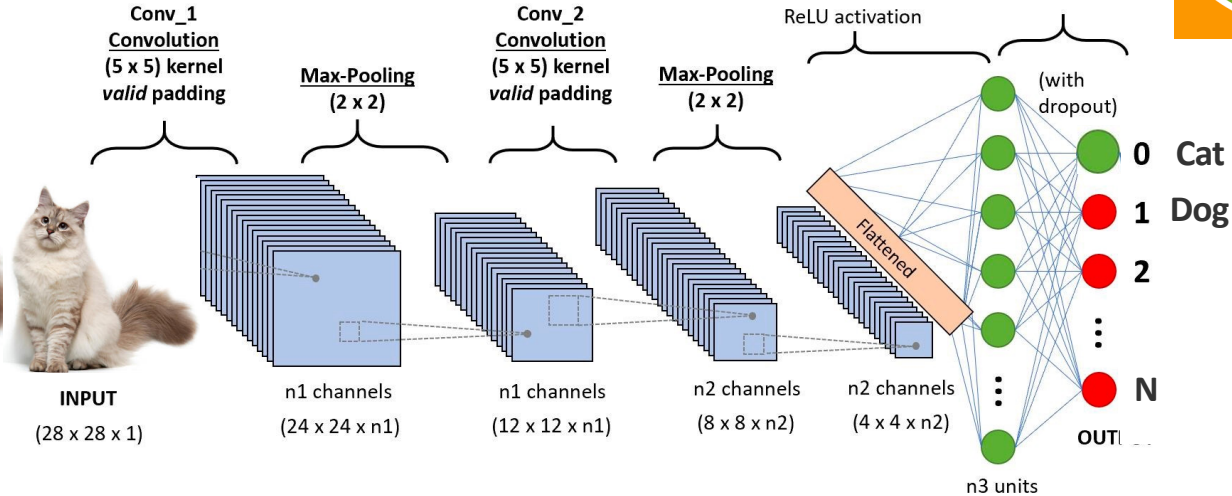
جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology

ARTIFICIAL
INTELLIGENCE
INITIATIVE

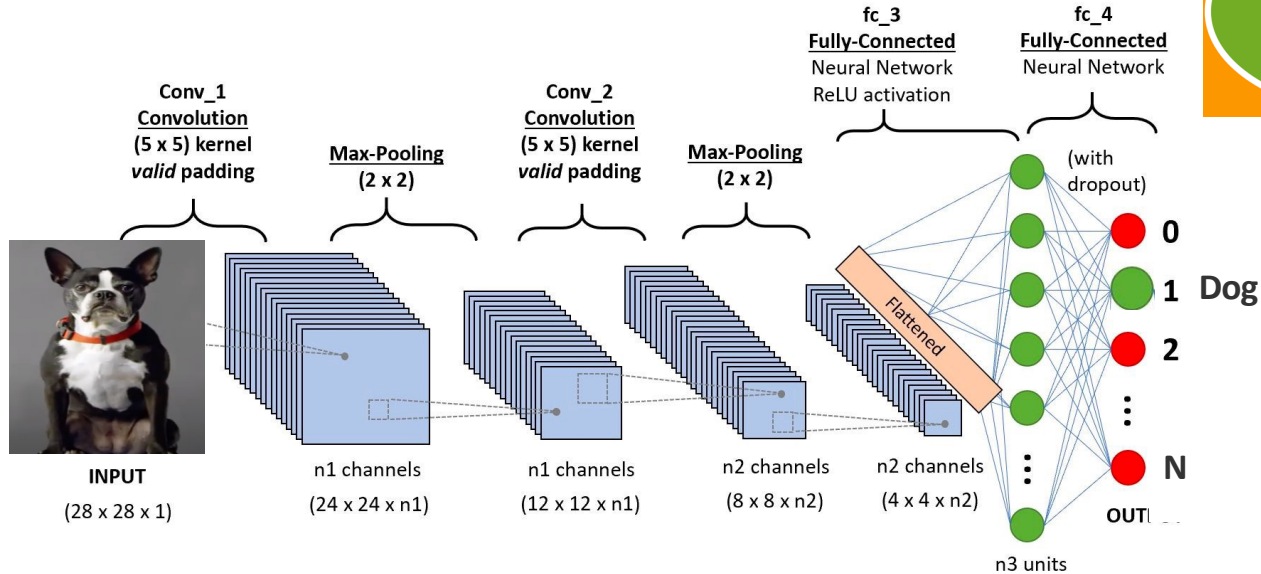
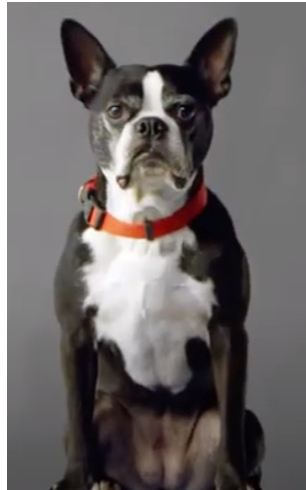
Imaginative Vision Language Models

Mohamed Elhoseiny
CS Faculty, KAUST

Deep Learning for Visual Recognition



Deep Learning for Visual Recognition



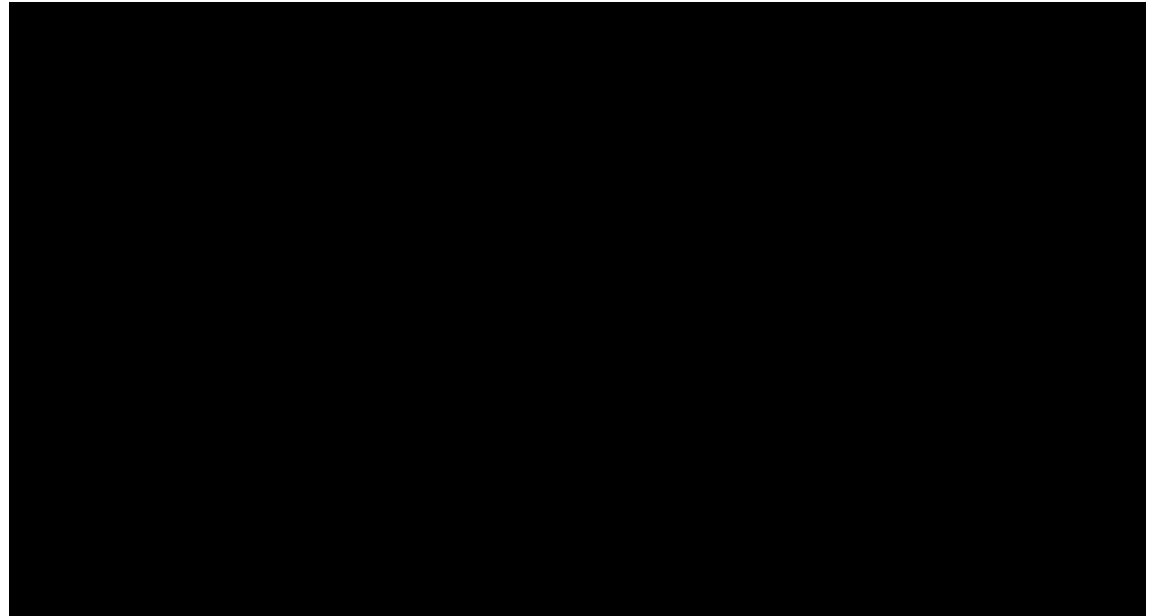
Fukushima, Kunihiko, and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition." *Competition and cooperation in neural nets*. Springer, Berlin, Heidelberg, 1982. 267-285.

Y LeCun B Boser J S Denker D Henderson R E Howard W Hubbard and L D Jackel Backpropagation applied to handwritten zip code recognition *Neural Computation*

Imaginative AI for Understanding Species

Video Credit: MammalNet (CVPR2023)

<https://mammal-net.github.io/>



Jun Chen, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, Mohamed Elhoseiny, MammalNet: A Large-Scale Video Benchmark For Mammal Recognition And Behavior Understanding , CVPR, 2023

- Animals are extremely diverse (estimated 7.7 million animal species, <1 million catalogued). [1]. ~ 90% are neither named nor discovered yet [2]
- We are interested in identifying them and understanding their behavior and activities (e.g., running, riding, sitting, etc.). However, with this scale, scientists argue that some of these species will become extinct before they can even be studied [2]

[1] Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. "How many species are there on Earth and in the ocean?." *PLoS biology* 9, no. 8 (2011): e1001127.
[2] Costello, Mark J., Robert M. May, and Nigel E. Stork. "Can we name Earth's species before they go extinct?." *Science* 339.6118 (2013):

Imaginative Vision and Vision LLM

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



e.g, ICCV13, CVPR17, CVPR18, ICCV19, ICLR21,
CVPR22

Imaginative AI for Species Discovery?

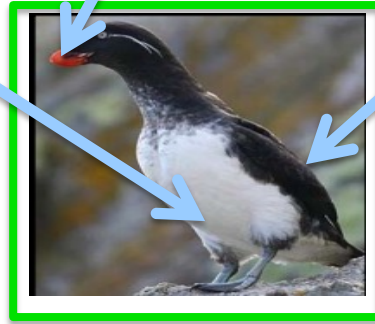
- Animals are extremely diverse (estimated 7.7 million animal species, <1 million catalogued). [1]. ~ 90% are neither named nor discovered yet [2]
- We are typically interested in identifying them and understanding their behavior and activities (e.g., running, riding, sitting, etc.). However, with this scale, scientists argue that some of these species will become extinct before they can even be studied [2]



[1] Mora, Camilo, Derek P. Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. "How many species are there on Earth and in the ocean?." *PLoS biology* 9, no. 8 (2011): e1001127.
[2] Costello, Mark J., Robert M. May, and Nigel E. Stork. "Can we name Earth's species before they go extinct?." *Science* 339.6118 (2013):

Recognition of Unseen Species

Parakeet Auklet is a small bird that has a short orange beak. The bird's plumage is dark above and white below.



Imaginative Visual Classifier from Text

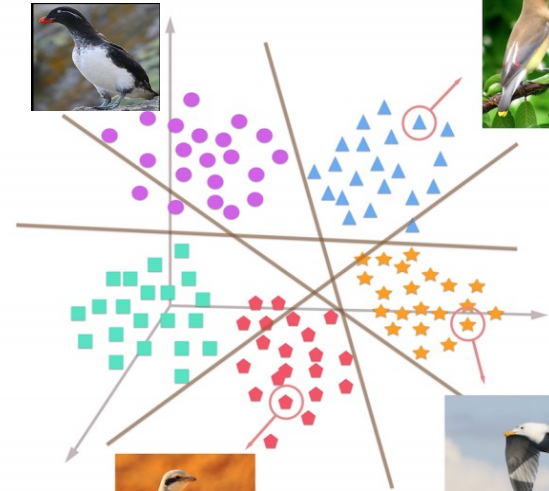
Data generation

Parakeet Auklet is a small bird that has a short **orange** beak. The bird's plumage is **dark** above and **white** below.

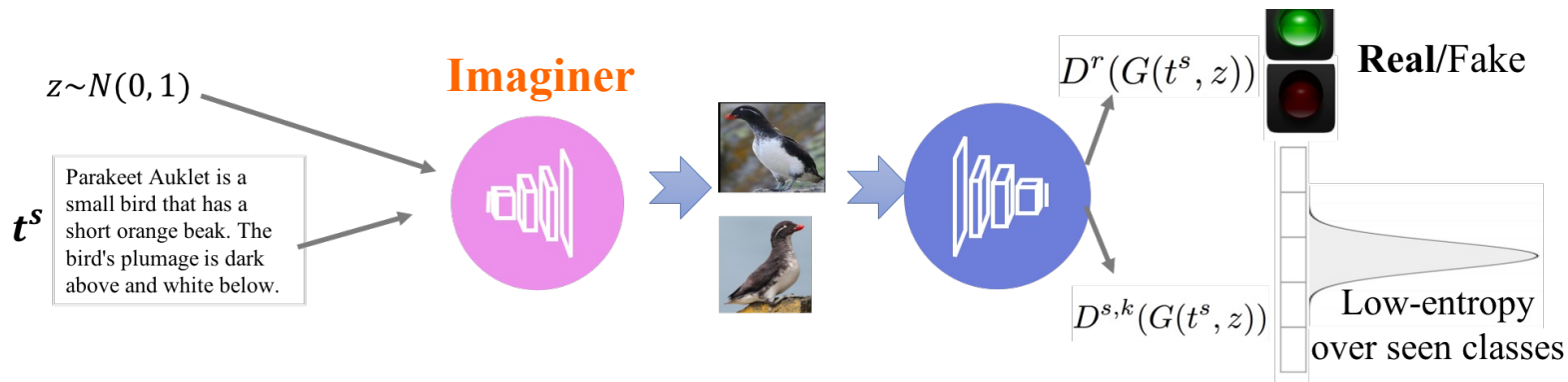


GAN

Supervised classifier



Generative Zero-shot learning success



Parallel work in related problems

[Xian, etal, 2018]

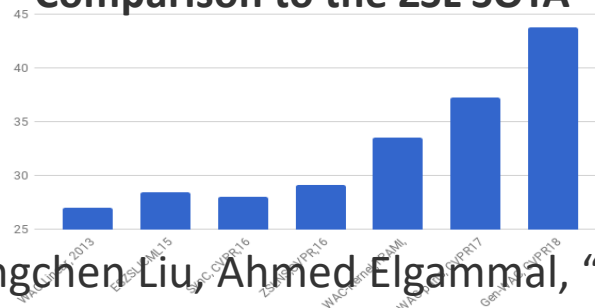
[Bharath, etal, 2017]

[Wang etal, 2018]

[YZ etal, 2018]

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Ahmed Elgammal, "Imagine it for me: Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts". CVPR,

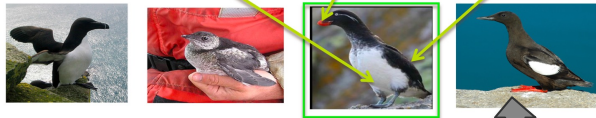
Comparison to the ZSL SOTA



Imagination to See and Creation

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



e.g. ICCV13, CVPR17, CVPR18, ICCV19, ICLR21,
CVPR22, ICCV23

IMAGINE TO CREATE

Art



Fashion



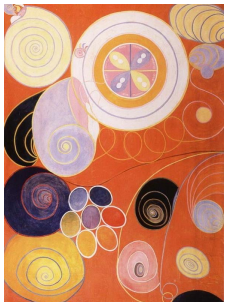
e.g. ICCV17, AAAI18, ECCVW18, ICCV19, CVPR21,
ICCV22, CVPR22, ICCV22

QUICK DETOUR

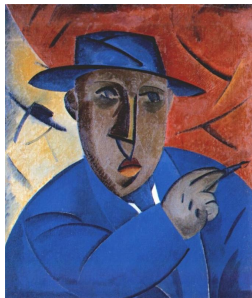
Creative Adversarial Networks, 2017

Wiki Art 20 Style Classes and Modeling the deviation

Abstract Art



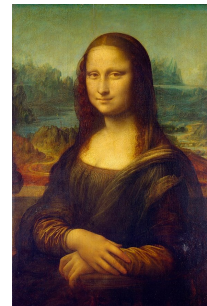
Cubism



Impressionism



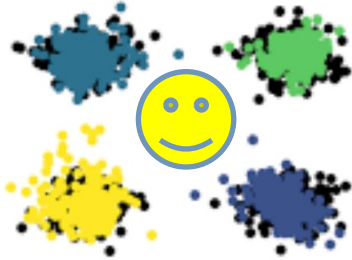
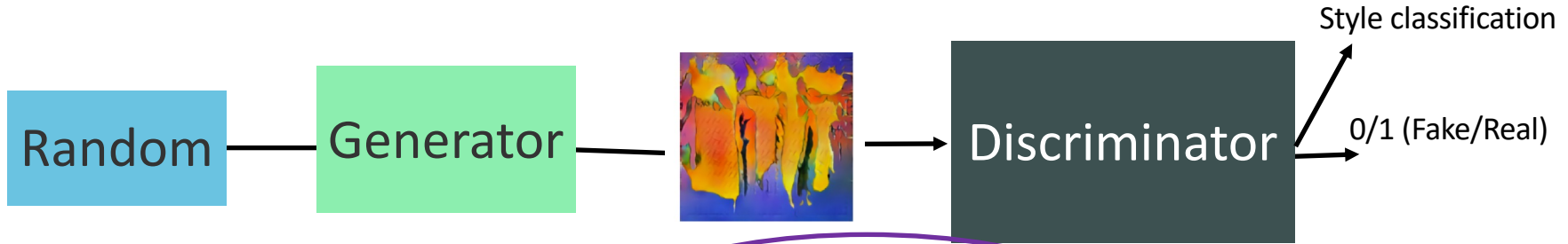
High Renaissance



.....

Creative Adversarial Networks, 2017

High Style Ambiguity (high Entropy)= high Creativity



$$\mathcal{L}(z) = - \sum_{k=1}^K \left(\frac{1}{K} \log(D_c(c_k|G(z))) + \frac{1}{K} \log(1 - D_c(c_k|G(z))) \right),$$

Creative Adversarial Networks, 2017

Qualitative Examples



Creative Walk Adversarial Networks (CWAN), ICCC, 2022

Creative Walk Adversarial Networks: Novel Art Generation with Probabilistic Random Walk Deviation from Style Norms

Divyansh Jha*, **Kai Yi**, **Ivan Skorokhodov**, **Mohamed Elhoseiny***

King Abdullah University of Science and Technology (KAUST)

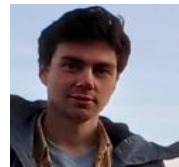
{divyansh.jha, kai.yi, ivan.skorokhodov, mohamed.elhoseiny}@kaust.edu.sa, * denotes Equal contribution.



Divyansh Jha



Kai Yi



Ivan Skorokhodov

Creative Walk Adversarial Networks, ICCV, 2022



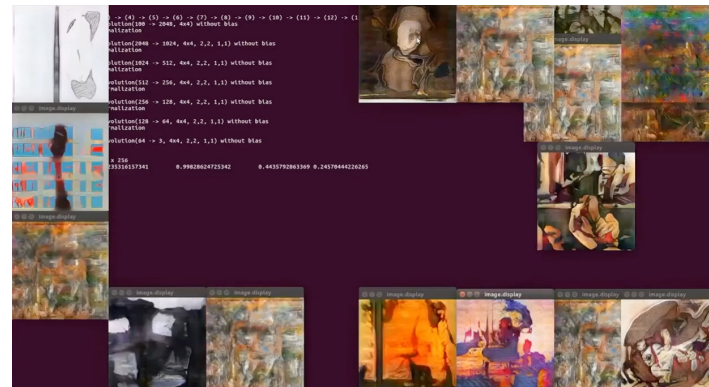
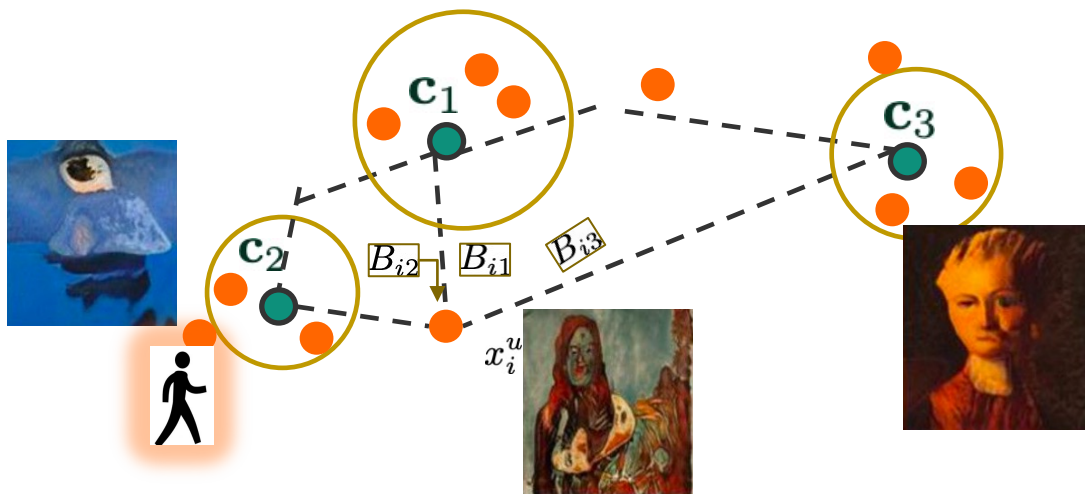
Generated artworks by CWAN



Generated
artworks by
CWAN

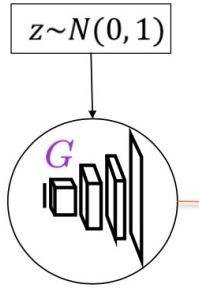
Nearest Neighbors from
training data

Creative Walk Adversarial Networks, ICCV, 2022

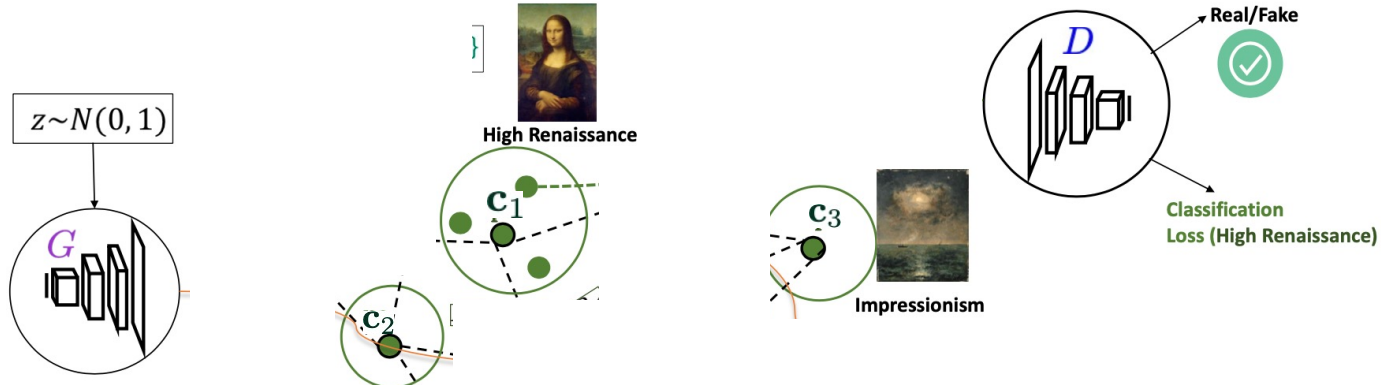


Goal: operates at the minibatch level producing generations that are message-passing to each other to facilitate better deviation of unseen classes/styles from seen ones. **Orange** are generated data

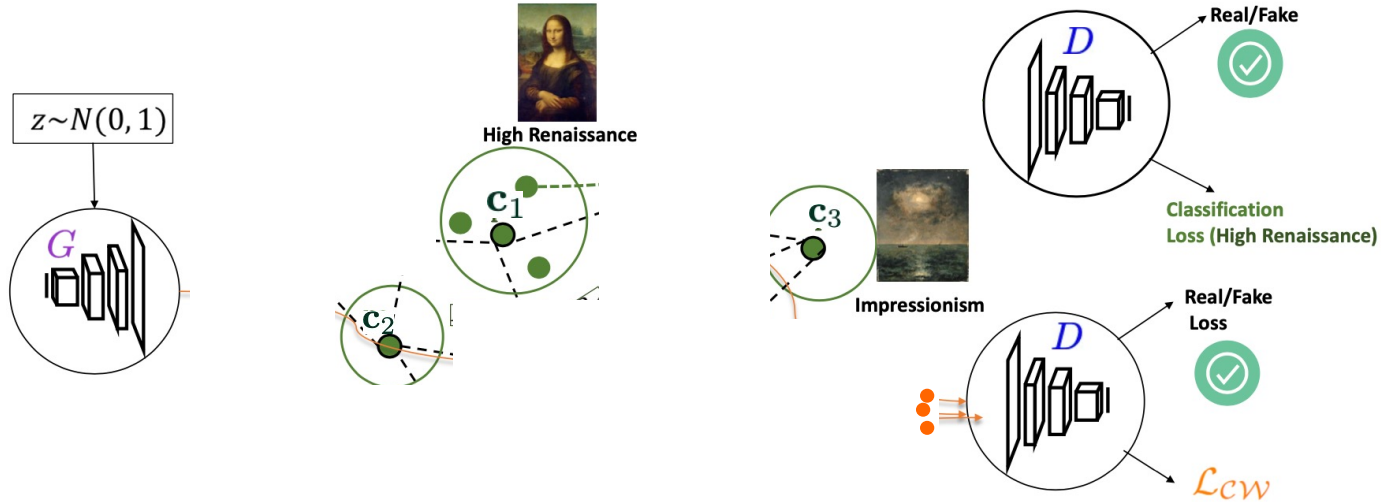
Creative Walk Adversarial Networks, ICCV, 2022



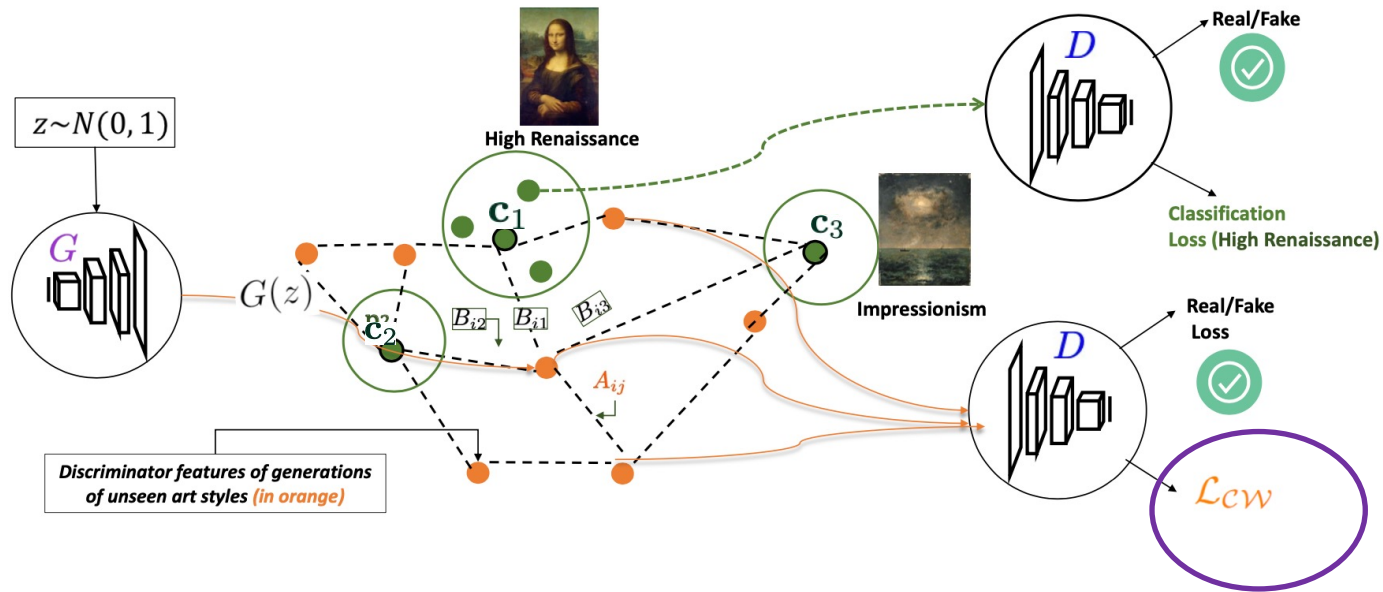
Creative Walk Adversarial Networks, ICCV, 2022



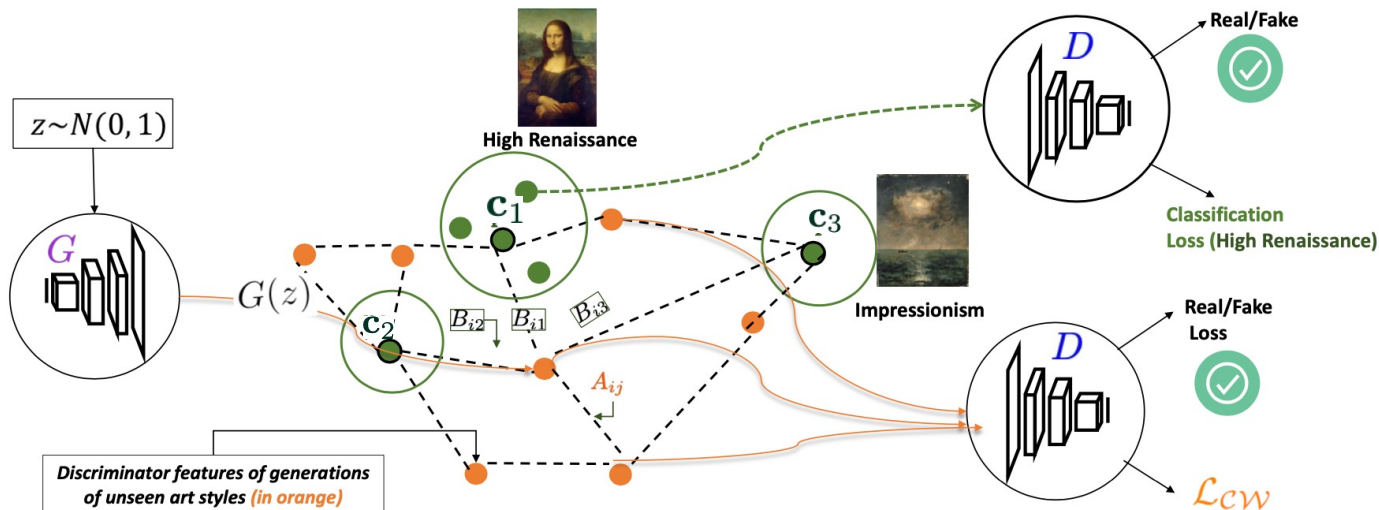
Creative Walk Adversarial Networks, ICCV, 2022



Creative Walk Adversarial Networks, ICCV, 2022



Creative Walk Adversarial Networks, ICCV, 2022



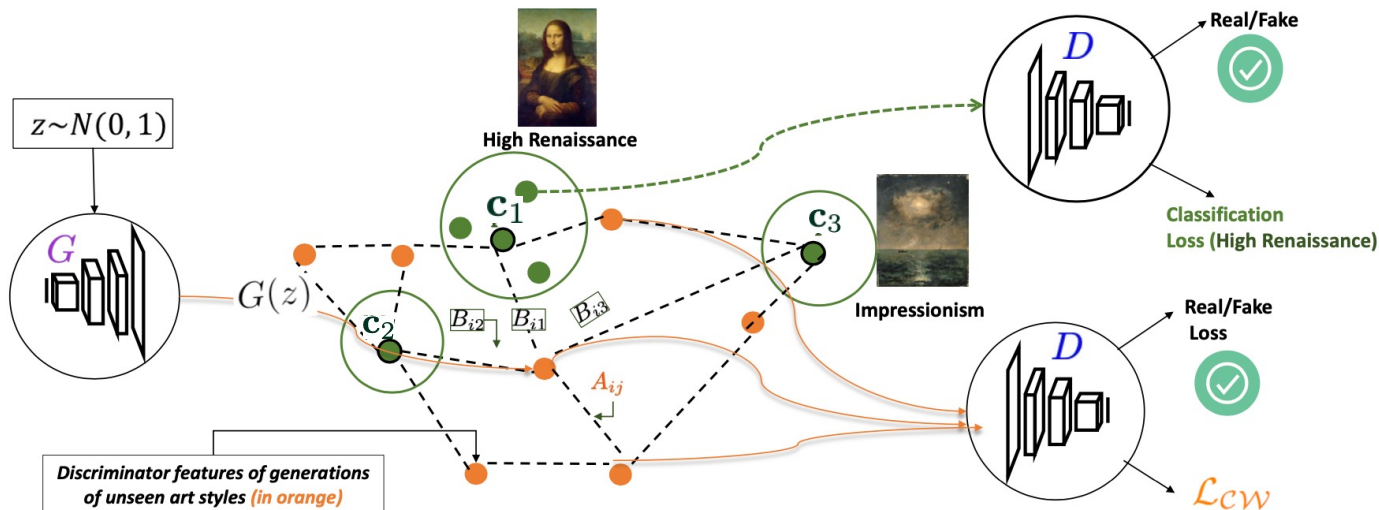
L_{cw}

Transition probabilities from seen class centers to unseen generations	Transition probabilities between unseen generations for t steps	Transition probabilities from unseen generations to seen class centers
--	---	--

$$P_{C \rightarrow C}(t, \tilde{X}) = \text{softmax}(B^T) \cdot \text{softmax}(A)^t \cdot \text{softmax}(B)$$

$$\sum_{t=0}^T \gamma^t \cdot H \left(H(t, \tilde{X}), \text{Uniform distribution} \right) \Rightarrow \sum_{i=1}^K \sum_{j=1}^K P_{C \rightarrow C}^{i,j}(t, \tilde{X})$$

Creative Walk Adversarial Networks, ICCV, 2022



L_{cw}

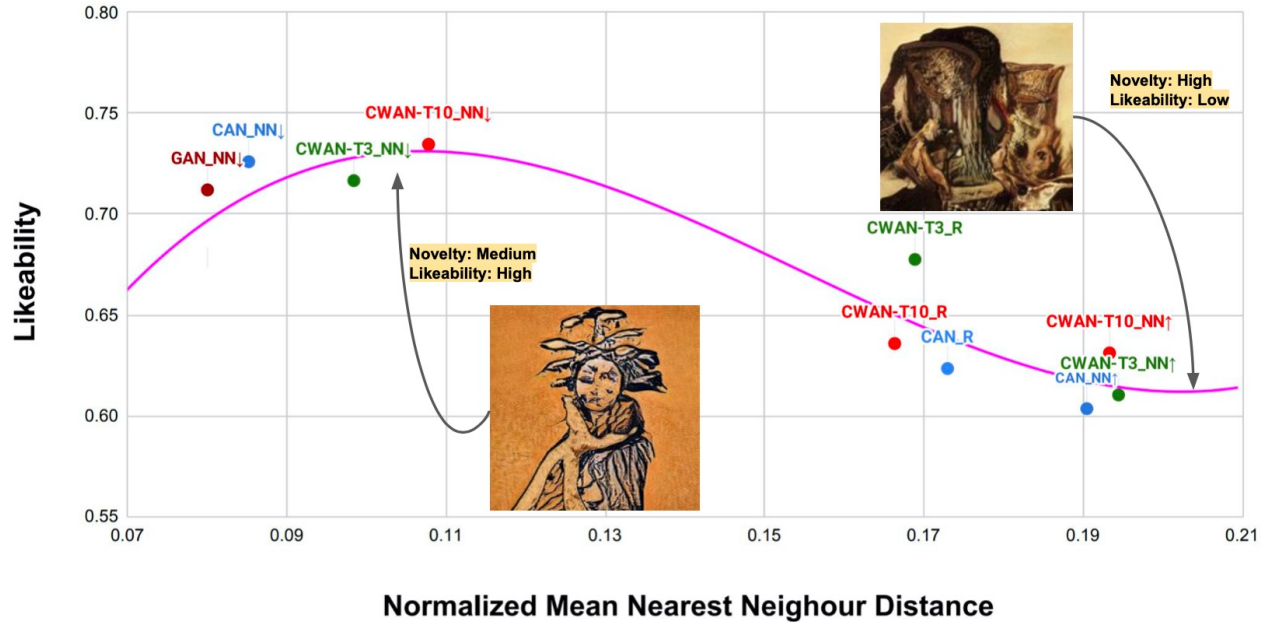
Transition probabilities from seen class centers to unseen generations Transition probabilities between unseen generations for t steps Transition probabilities from unseen generations to seen class centers

$$P_{C \rightarrow C}(t, \tilde{X}) = \text{softmax}(B^T) \cdot \text{softmax}(A)^t \cdot \text{softmax}(B)$$

$$\sum_{t=0}^T \gamma^t \cdot H \left(\text{Uniform distribution} \right) \rightarrow \sum_{i=1}^K \sum_{j=1}^K P_{C \rightarrow C}^{i,j}(t, \tilde{X})$$

Creative Walk Adversarial Networks, ICCV, 2022

Wundt Curve Reconstructed (Also Creative, see distance from NN)



Creative Walk Adversarial Networks, ICCV, 2022

CWAN (Constructing Emotional Experience)

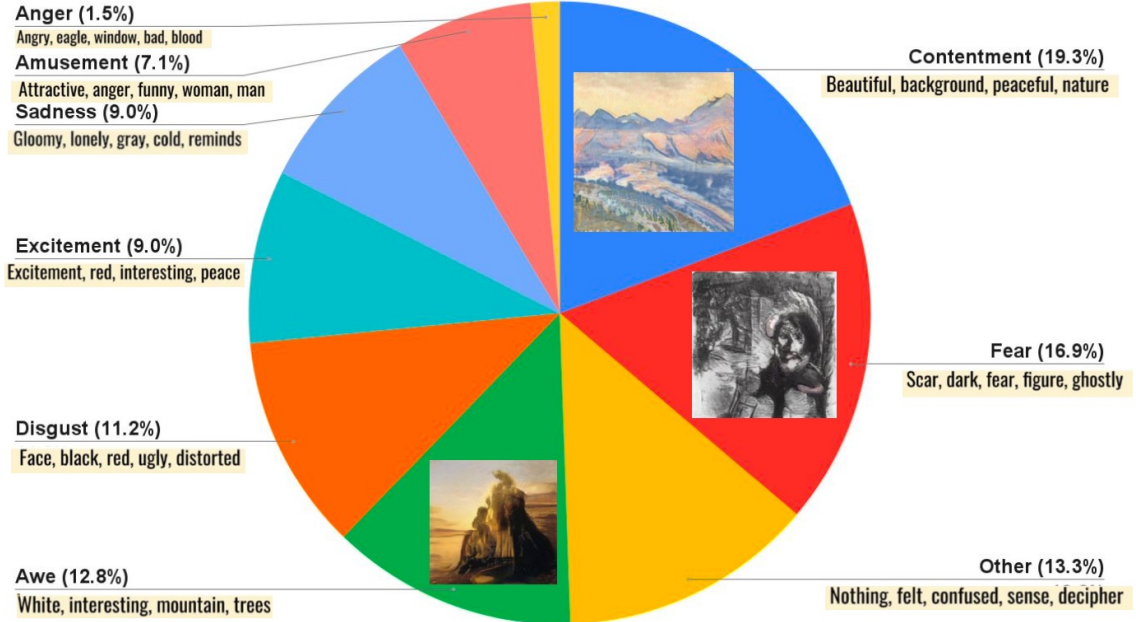
Contentment



Sunset piece that's super relaxing. Great piece with animals and trees in the background



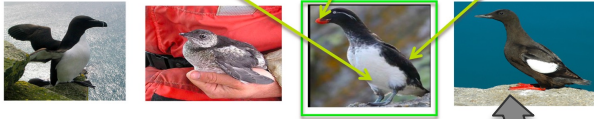
The rogue look of the man and the shirt used are just pretty good



Creativity Loss loops back to help ZSL, 2019

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



e.g, ICCV13, CVPR17, CVPR18, ICCV19, ICLR21, CVPR22

IMAGINE TO CREATE

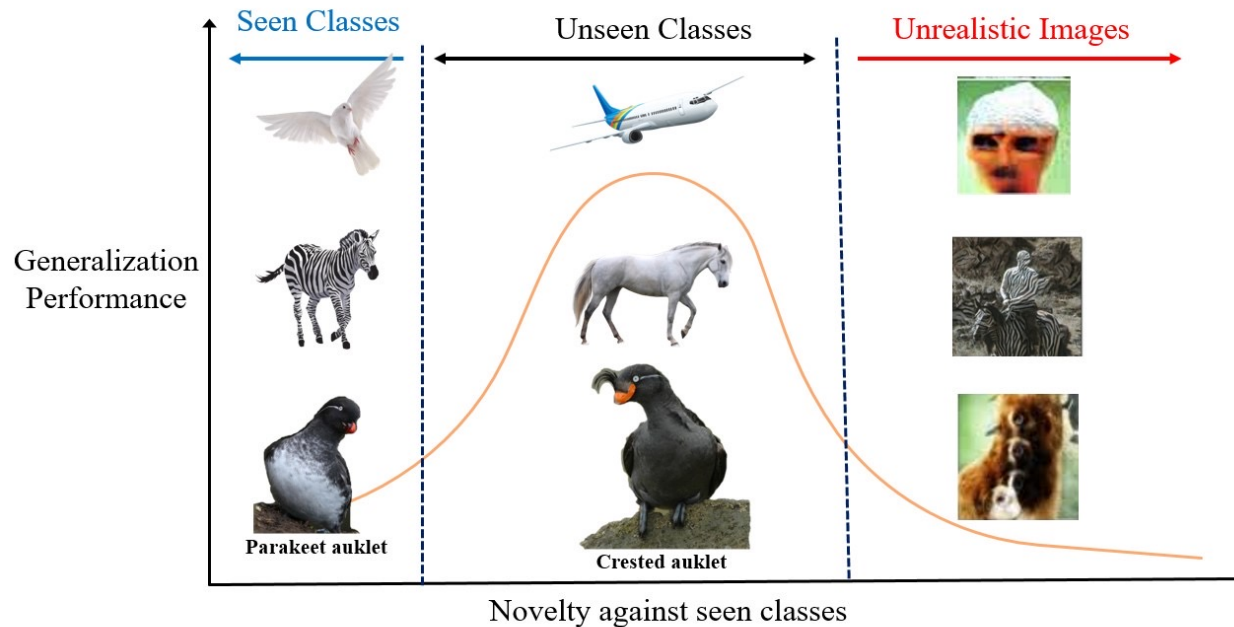
Art



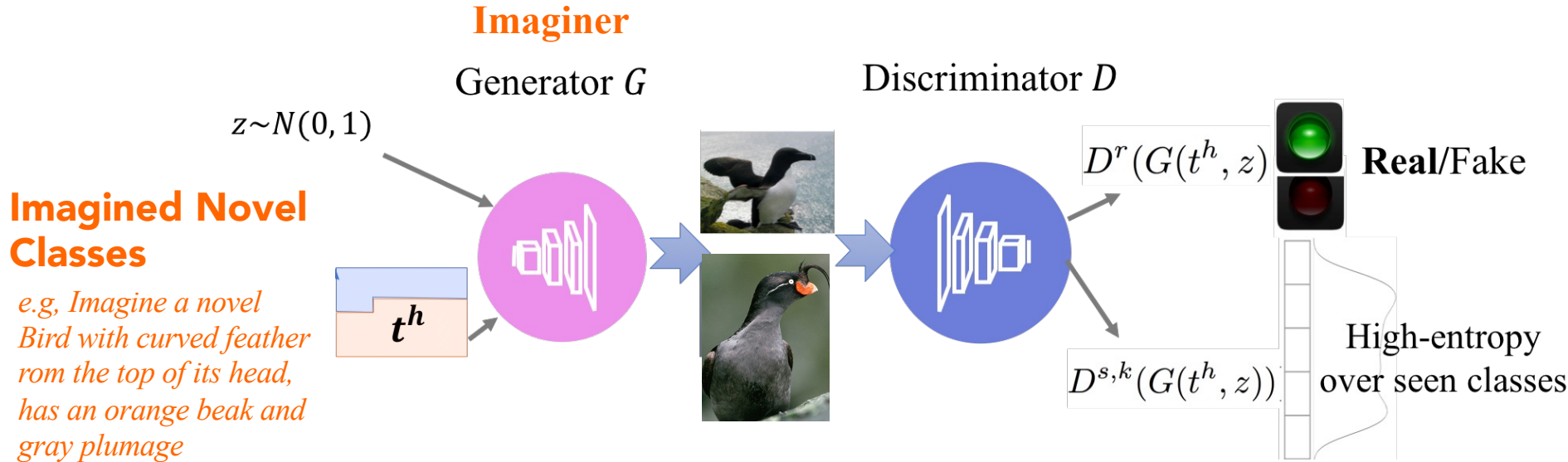
Fashion

e.g, ICCV17, AAAI18, ECCVW18, ICCV19, CVPR21, ICCV22, CVPR22, ICCV22

Creativity Inspired Zero-Shot Learning, ICCV19



Creativity Inspired Zero-Shot Learning, ICCV19



What to Teach AI to Do? Encourage its visual generations to be distinguishable from seen classes

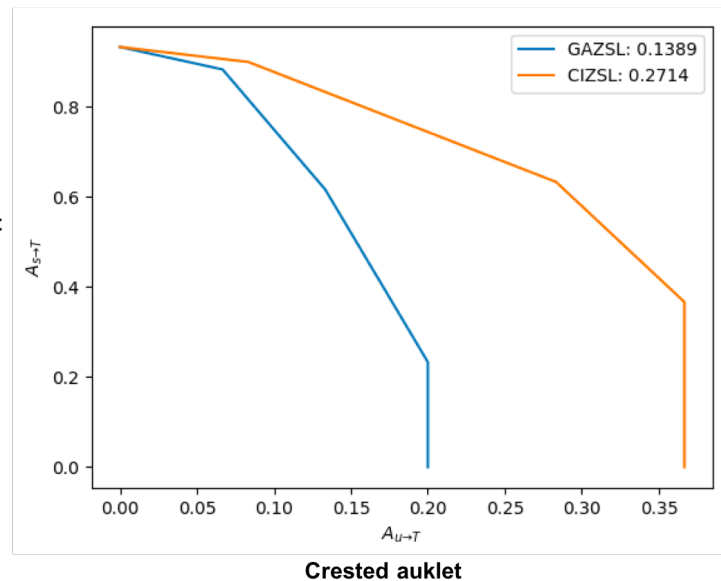
Creativity Inspired Zero-Shot Learning, ICCV19

- Notice the gap between the two curves on the unseen dimension (x-axis).

Parakeet auklet



- The only difference is our loss



Crested auklet

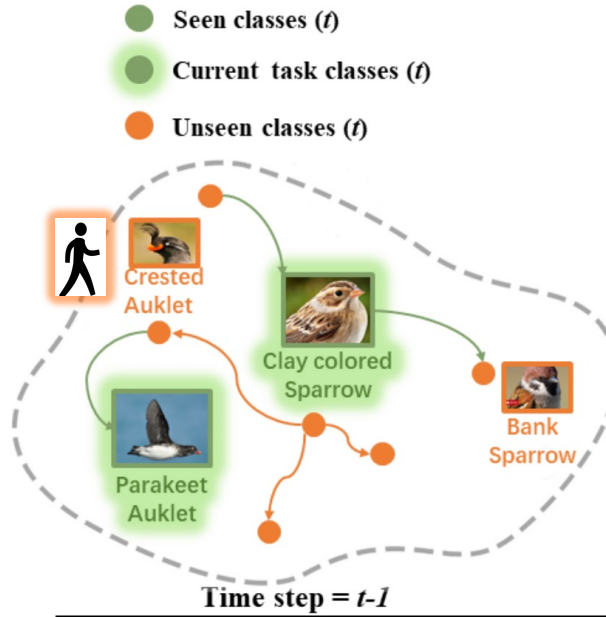


Creative Walks can be adapted Continual Learning of Species and Potential to discovering new ones, ICCV,2023



Wenxuan Zhang Paul Janson

(Vision-CAIR)



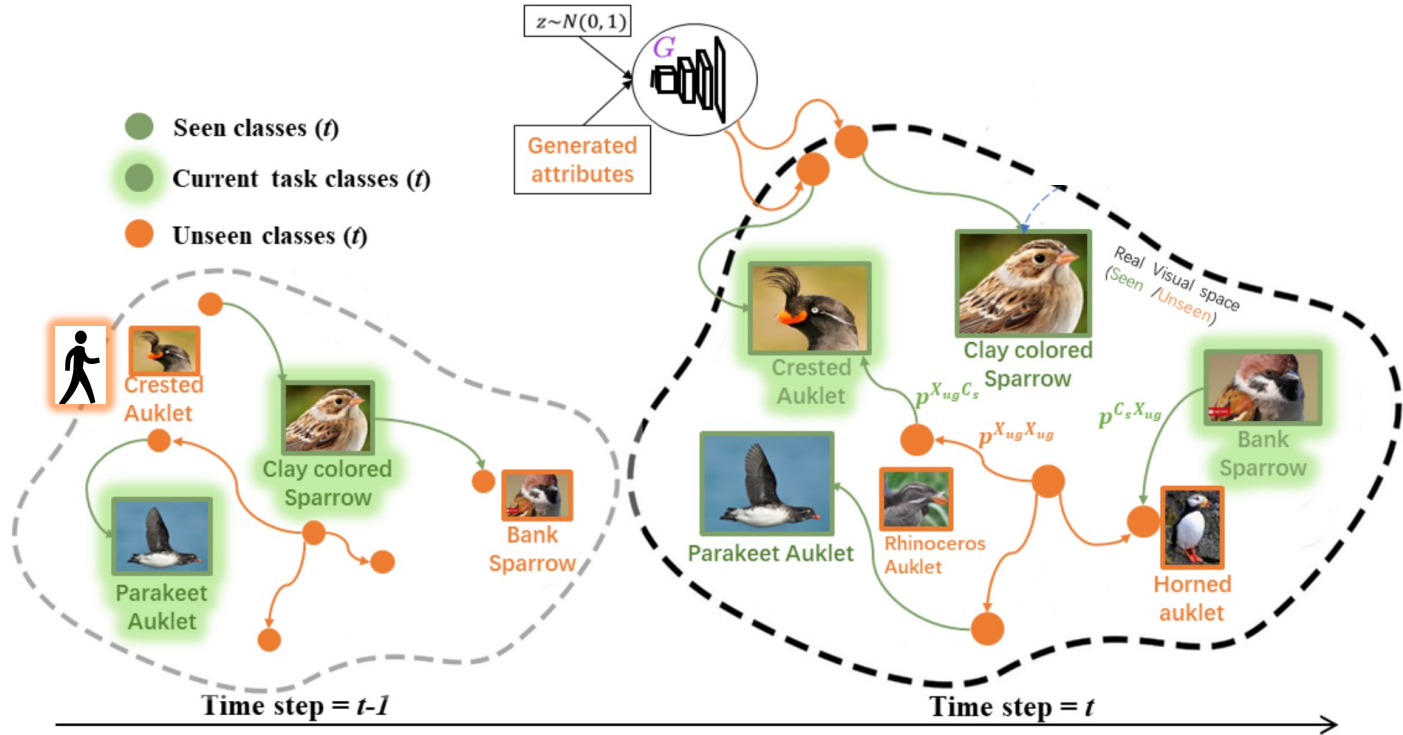
ICCV23: Creative Walks can be adapted Continual Learning of Species and Potential to discovering new ones,



Wenxuan Zhang Paul Janson

(Vision-CAIR)

جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology



ICCV23: Creative Walks can be adapted Continual Learning of Species and Potential to discovering new ones,



Wenxuan Zhang Paul Janson

(Vision-CAIR)

$$\mathcal{L}_D = -\mathcal{L}_{\text{real-fake}} + \lambda_{\text{cls}} \mathcal{L}_{\text{classification}} + \lambda_{\text{rd}} \mathcal{R}_D,$$

$$\mathcal{L}_G = \mathcal{L}_{\text{real-fake}} + \lambda_{\text{cls}} \mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{inductive}} + \lambda_{\text{rg}} \mathcal{R}_G.$$

$$\mathcal{L}_{\text{inductive}} = \lambda_c L_{\text{creativity}} + \lambda_i L_{\text{GRW}} + \lambda_i R_{\text{GRW}}$$

Unseen classes $L_{\text{GRW}} = \sum_{r=0}^R \gamma^r L_e(P^{C_s X_h C_s}(r), \mathcal{U}) + L_e(\mathbf{P}_v(C_s, X_h), \mathcal{U}_v)$

Seen classes $R_{\text{GRW}} = \sum_{r=0}^R \gamma^r L_e(P^{C_s X_{sg} C_s}(r), \mathcal{I}) + L_e(\mathbf{P}_v(C_s, X_{sg}), \mathcal{U}_v),$

ICCV23: Creative Walks can be adapted Continual Learning of Species and Potential to discovering new ones



Wenxuan Zhang, Paul Janson

(Vision-CAIR)



	interpolation	dictionary
with $R_{GRW} + L_{GRW}$	28.4	28.8
- $L_{creativity}$	27.72	27.66
w/o R_{GRW}, L_{GRW}	19.07	20.75
- $L_{creativity}$	14.43	14.43
with L_{GRW}	26.73	27.39

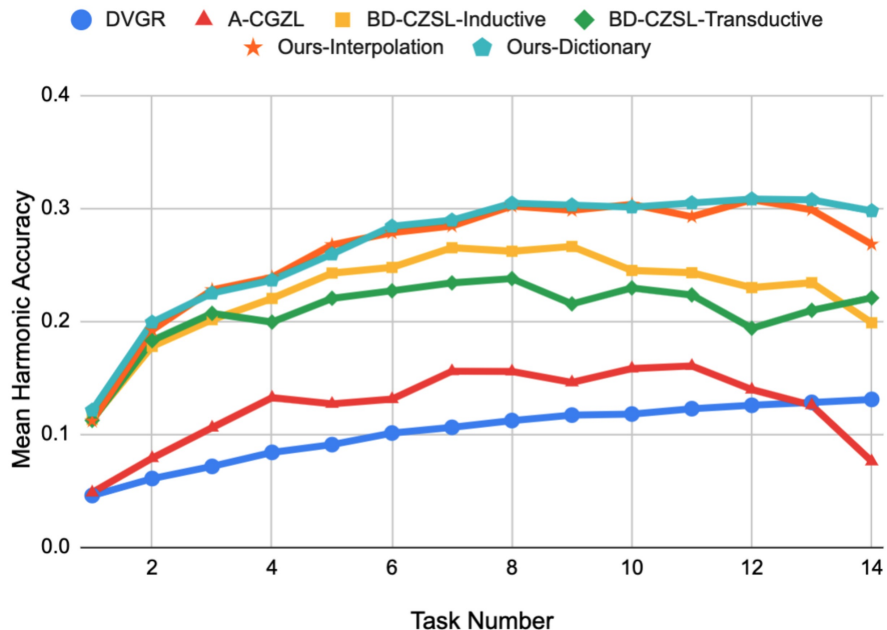
Table 3. Effect of the random walk-based penalty with mH measure on CUB dataset.

ICCV23: Creative Walks can be adapted Continual Learning of Species and Potential to discovering new ones



Wenxuan Zhang, Paul Janson

(Vision-CAIR)

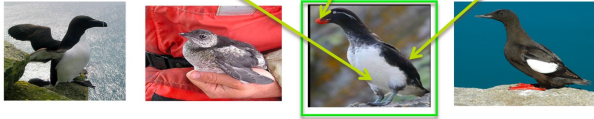


Average Harmonic accuracy up until each task on SUN dataset. Creative Walk outperforms existing methods

Quick Detour

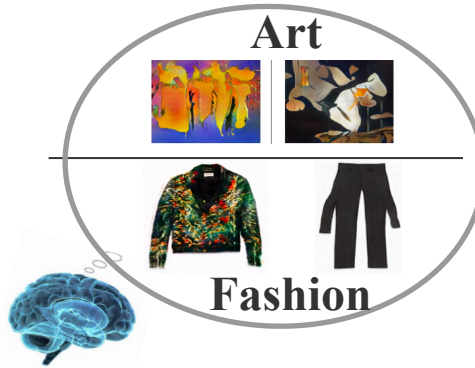
IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



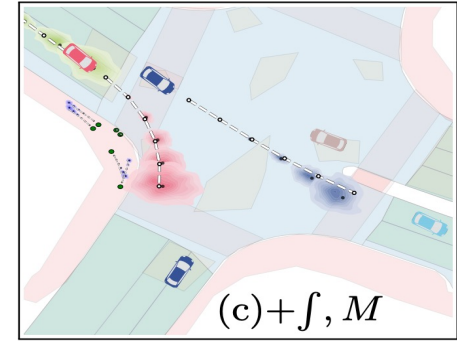
e.g. ICCV13, CVPR17, CVPR18, ICCV19, ICLR21, CVPR22

IMAGINE TO CREATE



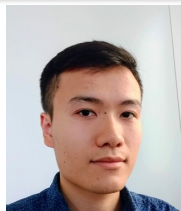
e.g. ICCV17, AACL18, ECCVW18, ICCV19, CVPR21, ICCV22, CVPR22, ICCV22

IMAGINE TO DRIVE



e.g. CVPR2020, ICLR2021, CoRL21, ECCV22, ICLR23

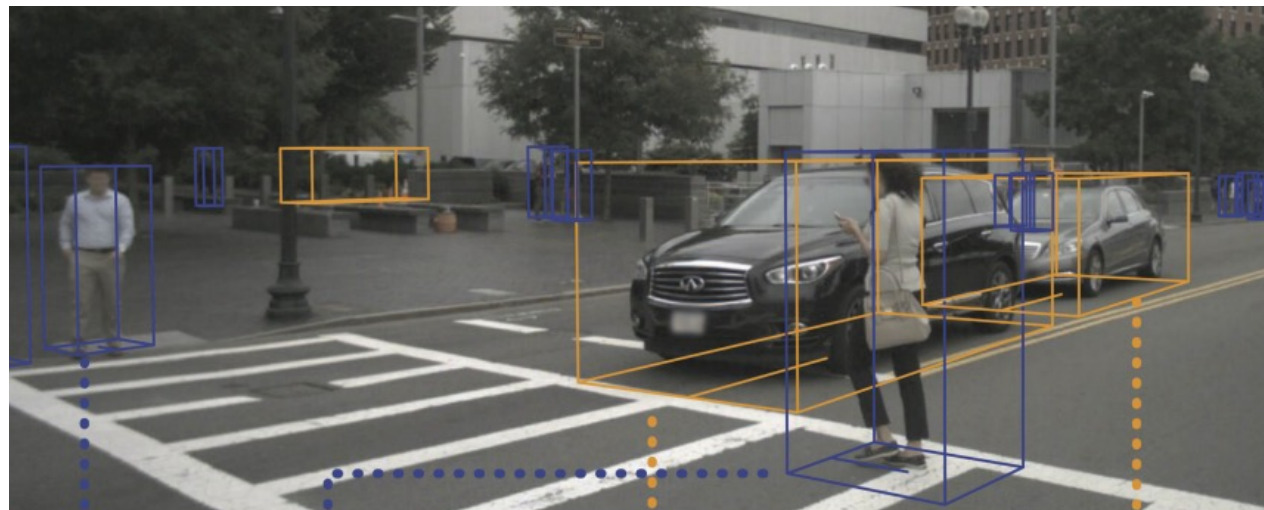
Imaginative Self Driving Tech , ICLR, 2021



Deyao Zhu



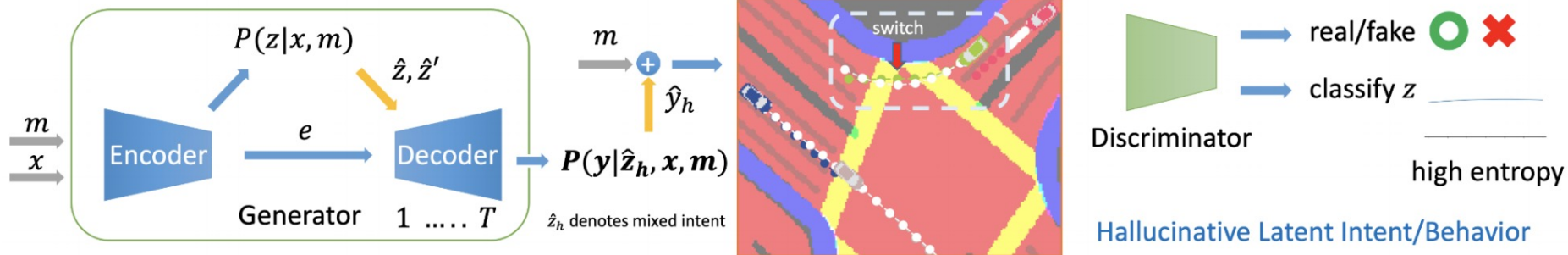
Mohamed Zahran



King Abdullah University of
Science and Technology

HalentNet: Multimodal Trajectory Forecasting with Hallucinative Intents,
Deyao Zhu, Mohamed Zahran, Li Erran Li, Mohamed Elhoseiny, *ICLR*, 2021

Hallucinative latent Intent Behavior



- Dynamically augmenting agent behaviors by generating trajectories with mixed intents
- Augment training data implicitly

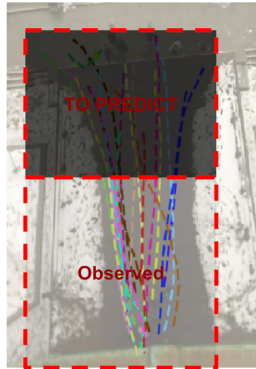
Quantitative Results, ICLR, 2021

Metric	Model	1s	2s	3s	4s	5s	6s	
FDE (Full)	Traj++	0.16	0.64	1.52	2.80	4.53	6.70	> 27% reduction
	Ours	0.09	0.52	1.21	2.17	3.41	4.93	
RB. Viol.	Traj++	0.24%	0.57%	2.55%	7.04%	12.95%	19.09%	> 50% reduction
	Ours	0.26%	0.45%	1.30%	3.21%	6.00%	9.22%	

- More than 50% reduction on Road Boundary violation (**RB. Viol.**)
- More than 27% reduction on the Final Displacement Error (**FDE**)

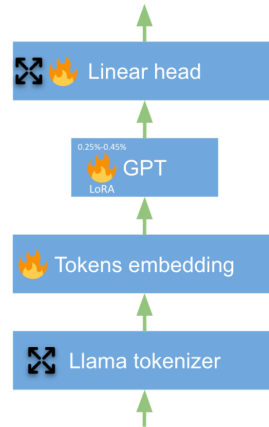
In progress, 2023

Fine-tuning Large Language Models on Motion Features

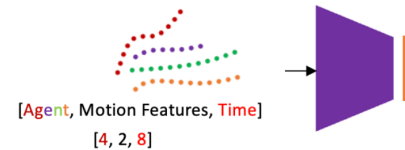
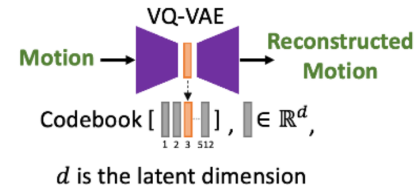


2D Trajectory motion of multiple agents

Output: Target agent motion symbols



Input prompt: Text + agents symbols + agents' motion symbols

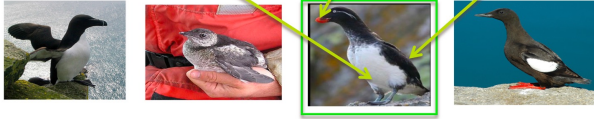


Encoding motion into discrete codes

Imagination to See and Creation

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short **orange** bill. The bird's plumage is **dark** above and **white** below.



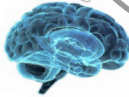
e.g, ICCV13, CVPR17, CVPR18, ICCV19, ICLR21, CVPR22, ECCV 2022

IMAGINE TO CREATE

Art



Fashion



e.g, ICCV17, AAAI18, ECCVW18, ICCV19, CVPR21, ICCV22, CVPR22, ICCV22

IMAGINE TO FEEL

Affective Visual Art

ArtEmis CVPR21, ArtEmis-2.0 CVPR22
Wofflin Generative analysis, ICCV, 2021



Makes me feel like birds are



The pale color palette of this watercolor painting is very relaxing. I can imagine myself sitting by the water *listening* to the birds.

Emotion and Cultural Aware Imaginative AI

ArtEmis (CVPR21)

450K additional utterances

ArtEmis: Affective Language for Visual Art



Panos Achlioptas¹

panos@cs.stanford.edu

Maks Ovsjanikov²

maks@lix.polytechnique.fr



Kilichbek Haydarov³

kilichbek.haydarov@kaust.edu.sa



Mohamed Elhoseiny^{3,1}

mohamed.elhoseiny@kaust.edu.sa

Leonidas Guibas¹

guibas@cs.stanford.edu



¹Stanford University

²LIX, Ecole Polytechnique, IP Paris

³King Abdullah University of Science and Technology (KAUST)

ArtEmis 2.0 (CVPR22)

260K additional utterances



It is Okay to Not Be Okay:

Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection

www.artemisdataset-v2.org



جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology



Youssef Mohamed



Faizan Khan



Kilichbek Haydarov





Mohamed Elhoseiny

ArtELingo (EMNLP 2022, long paper)

ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture

Youssef Mohamed, Shyma Yaser Alhuwaidar, Mohamed Abdelfattah, Feifan Li, Kenneth Ward Church, Xiangliang Zhang and Mohamed Elhoseiny

a)	 <p>شلال طبيعي جميل. مشاعر النمو والحياة والطاقة موجودة. Translation: Beautiful natural waterfall. Feelings of growth, vitality and energy.</p> <p>Excitement Arabic 😄</p>	<p>The water that's rushing downward looks like a bride's wedding veil.</p> <p>Awe English 😲</p>	<p>瀑布就像四蹄生风的白马如潮水涌来，非常的壮观 Translation: The waterfall is like a white horse and wind, it is spectacular.</p> <p>Contentment Chinese 😊</p>
b)	 <p>Translation: Girls sitting with their mother outside the house, exchanging love and affection, pigeons flying over a tree.</p> <p>Contentment Arabic 😊</p>	<p>The women relaxing while birds are flying about makes me feel relaxed and calm as well.</p> <p>Contentment English 😊</p>	<p>Translation: Three sisters lying on a bench and watching the birds fly comfortably.</p> <p>Contentment Chinese 😊</p>
c)	 <p>Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction.</p> <p>Contentment Arabic 😊</p>	<p>The trees are dead and exposing their roots due to erosion and lack of water.</p> <p>Sadness English 😞</p>	<p>Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed.</p> <p>Sadness Chinese 😞</p>

ArtELingo:

A Million Emotion Annotations of WikiArt
with Emphasis on Diversity over
Language and Culture



www.artelingo.org



Youssef Mohamed



Mohamed Abdelfattah



Shyma Alhuwaider



Feifan Li



Xiangliang Zhang



Kenneth Ward Church



Mohamed Elhoseiny

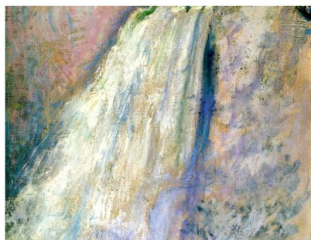


EMNLP
2022

Facts (COCO) → Emotions (ArtEmis) → Multilingual (ArtELingo)



	Facts	Emotions	Multilingual
	COCO	ArtEmis	ArtELingo
Image Source	Photos	WikiArt	WikiArt
#Images	328k	80k	80k
#Annotations	2.5M	0.45M	1.2M
#Annot/Image	7.6	5.68	15.3
Emotions	0	9	9
Languages	E	E	ACE



شلال طبيعي جميل. مشاعر النمو والحيوية والطاقة موجودة.

Translation: Beautiful natural waterfall. Feelings of growth, vitality and energy.

Excitement
Arabic



The water that's rushing downward looks like a bride's wedding veil.

Awe
English



瀑布就像四蹄生风的白马如潮水涌来，非常的壮观

Translation: The waterfall is like a white horse and wind, it is spectacular.

Contentment
Chinese



Translation: Girls sitting with their mother outside the house, exchanging love and affection, pigeons flying over a tree.

Contentment
Arabic



The women relaxing while birds are flying about makes me feel relaxed and calm as well.

Contentment
English



Translation: Three sisters lying on a bench and watching the birds fly comfortably.

Contentment
Chinese



Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction.

Contentment
Arabic



The trees are dead and exposing their roots due to erosion and lack of water.

Sadness
English

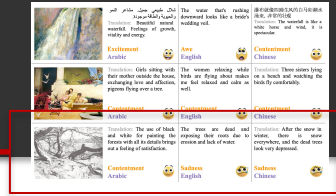


Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed.

Sadness
Chinese



Diversity over Language/Culture



Translation: The use of black and white for painting the forests with all its details brings out a feeling of satisfaction.

Contentment
Arabic



The trees are dead and exposing their roots due to erosion and lack of water.

Sadness
English



Translation: After the snow in winter, there is snow everywhere, and the dead trees look very depressed.

Sadness
Chinese



- An Arab Bedouin grew up in a dry climate → **Positive**
- English and Chinese → **Negative**

Multicultural Aware Emotion Classifier



[A]



Input Caption (Gloss)	Transformer Head		
	E	A	C
[A] A beautiful girl holding a Jesus cross stomping on the <u>devil</u>	Awe	Awe	<u>Fear</u>
[E] The man looks like he's <u>drunk</u> since his expression is so wired out	Amu.	<u>Sad</u>	Exc.
[C] Countless <u>babies</u> have descended into the world, giving life to the world and making <u>people</u> feel happy.	<u>Cont.</u>	<u>Cont.</u>	<u>Cont.</u>

ArtELingo:

A Million Emotion Annotations of WikiArt with
Emphasis on Diversity over
Language and Culture

- Download **ArtELingo** and **Baseline Models**
- Facts (COCO) → Emotions (ArtEmis) → Multilingual (ArtELingo)
- Embrace **Diversity**
- Two Tasks:
 - Emotion Prediction: Caption → Emotion Label
 - Caption Generation: Painting + Emotion Label + Language → Caption

Promise



www.artelingo.org

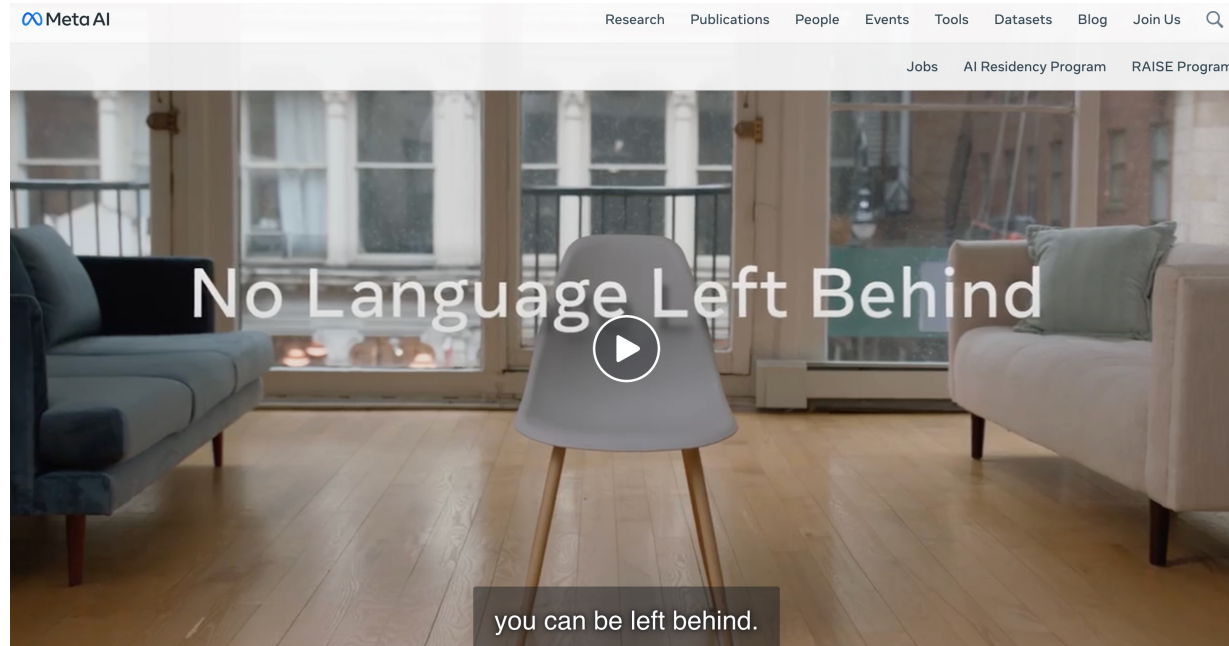
جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology



EMNLP
2022

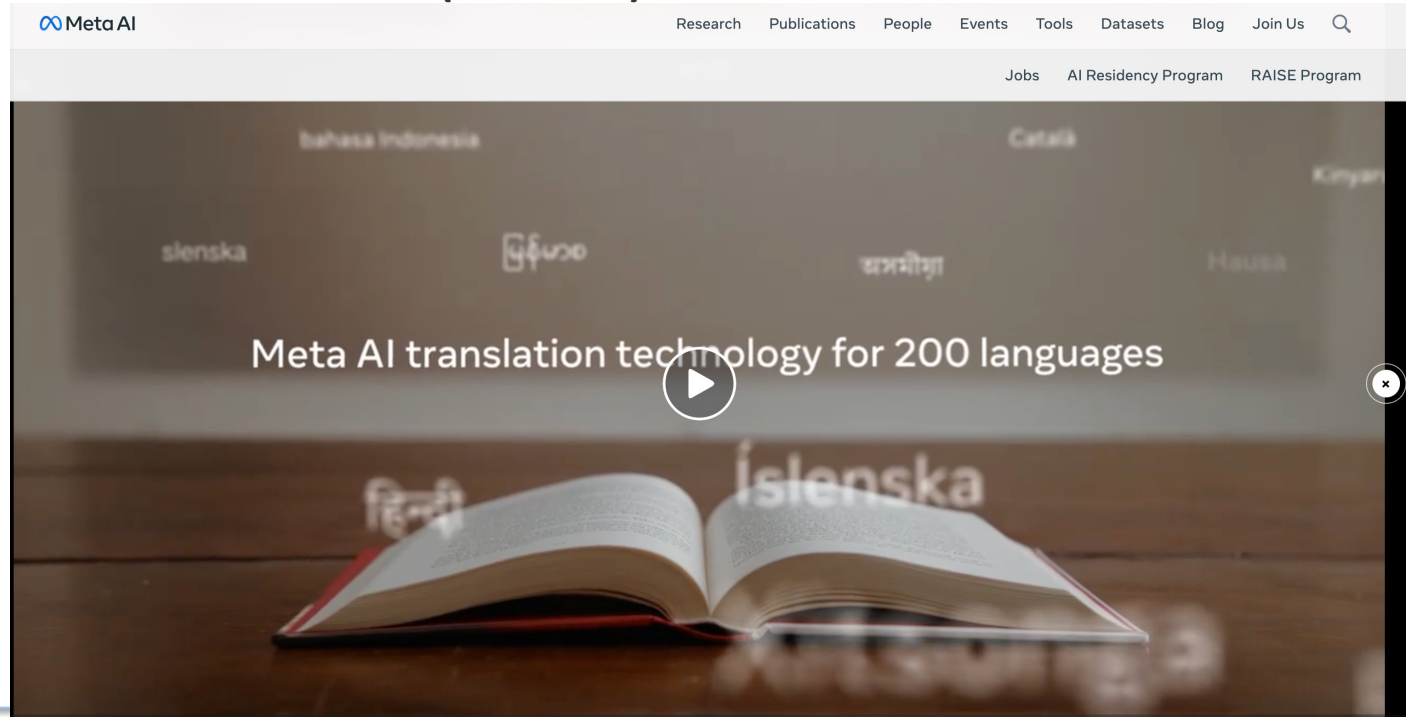
No Culture Left Behind Emotion Aware AI

NO LANGUAGE LEFT BEHIND (BY META). ONLY LANGUAGE BASED



No Culture Left Behind Emotion Aware AI

NO LANGUAGE LEFT BEHIND (BY META). ONLY LANGUAGE BASED



No Culture Left Behind Emotion-Aware AI

HOW NO CULTURE LEFT BEHIND EMOTION-AWARE AI?

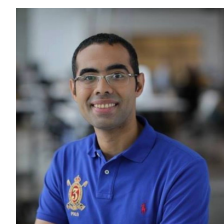
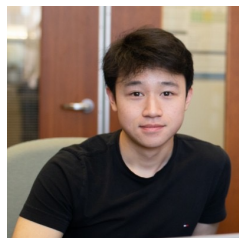
- DIFFERENT FROM NO LANGUAGE LEFT BEHIND, EMOTION DISTRIBUTION IS DIFFERENT FOR EACH CULTURE? VERY DIFFERENT AI SKILL COMPARED TO TRANSLATION AND MUCH HARDER.
- HOW TO EXTEND TO MANY MORE LANGUAGES TO COVER AS MANY CULTURES TO BUILDING INCLUSIVE PLATFORMS/ METAVERSE ? MUCH MORE CHALLENGING COMPARED TO TRANSLATION





VisualGPT: Data-efficient Adaption of Pretrained Language Models for Image Captioning

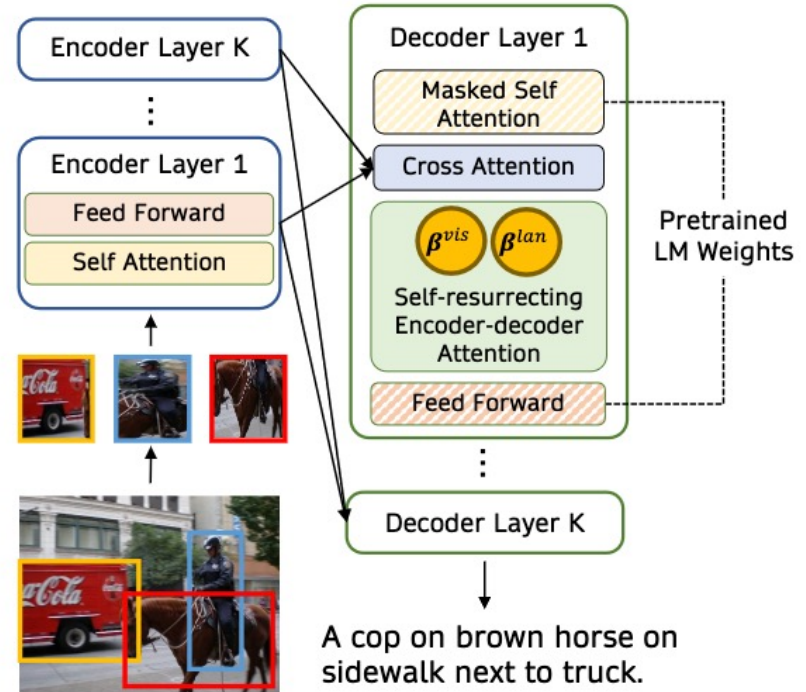
Jun Chen, Han Guo, Kai Yi, Boyang Li, Mohamed Elhoseiny



Visual GPT, CVPR 2022

Data-efficient Adaptation of Pretrained Language Models for Image Captioning

- Directly employs a GPT-2 model as the image captioning decoder
- a self-resurrecting encoder-decoder attention to balance the visual and linguistic knowledge.



MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models




Deyao Zhu* ([on job Market](#)), Jun Chen* ([on job Market](#)), Xiaoqian Shen,
Xiang Li, Mohamed Elhoseiny

King Abdullah University of Science and Technology

- GPT-4 achieves next-level vision-language abilities like
 - Explaining the funny part of an image

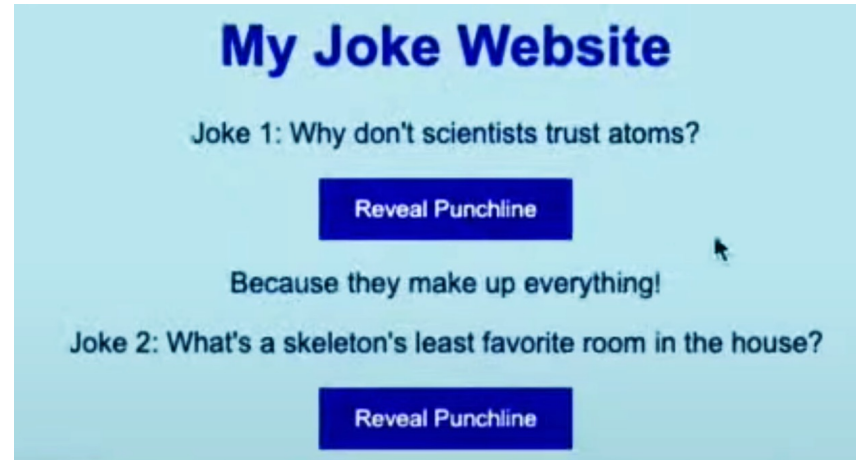
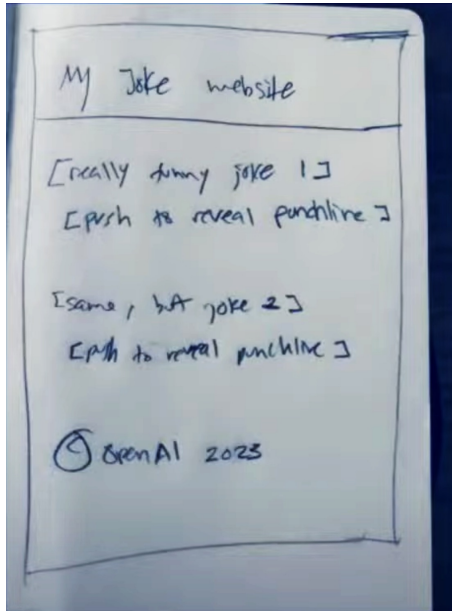


 **GPT-4 BOT** Today at 1:16 PM

The image is funny because it shows a squirrel holding a camera and taking a photo of a nut as if it were a professional photographer. It's a humorous situation because squirrels typically eat nuts, and we don't expect them to use a camera or act like humans.

😊 3 🍌 1 🐿️ 1

- GPT-4 achieves next-level vision-language abilities like
 - Creating a website from a draft



- Such abilities are never shown in previous SOTA methods like DeepMind's Flamingo^[1] or Salesforce's Blip-2^[2]
- **Nobody knows how they do it!**

GPT-4 Technical Report

This report focuses on the **capabilities, limitations, and safety properties** of GPT-4. GPT-4 is a Transformer-style model [33] pre-trained to predict the next token in a document, using both publicly available data (such as internet data) and data licensed from third-party providers. **The model was then fine-tuned using Reinforcement Learning from Human Feedback (RLHF)** [34]. Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains **no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.**

[1] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. NeurIPS 2022

[2] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint 2023

- **What is the secret of GPT-4's Vision-Language Abilities?**
 - Fancy large datasets with data like draft-to-website image pairs?
 - Secret model architectures?
 - Or just with an advanced large language model?





- It might be possible to simply aligning Blip-2's vision component with a better language model to achieve a much better vision-language instruction following ability.
- **Is it the secret of GPT-4's vision-language abilities?**

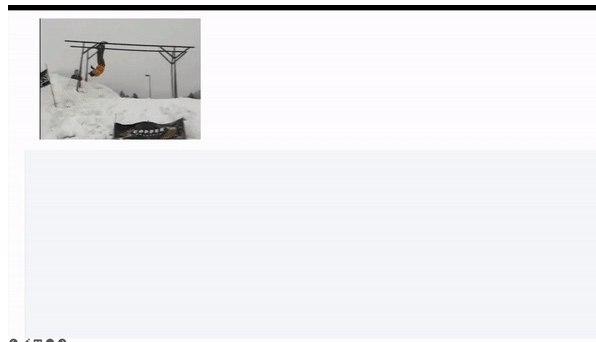
ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions

Deyao Zhu Jun Chen* Kilichbek Haydarov*
Xiaoqian Shen Wenxuan Zhang Mohamed Elhoseiny
King Abdullah University of Science and Technology

- The vision part of Blip-2 can provide rich information
- But the language part is not strong enough

Video ChatCaptioner: Towards Enriched Spatiotemporal Descriptions

Jun Chen Deyao Zhu Kilichbek Haydarov Xiang Li Mohamed Elhoseiny
King Abdullah University of Science and Technology

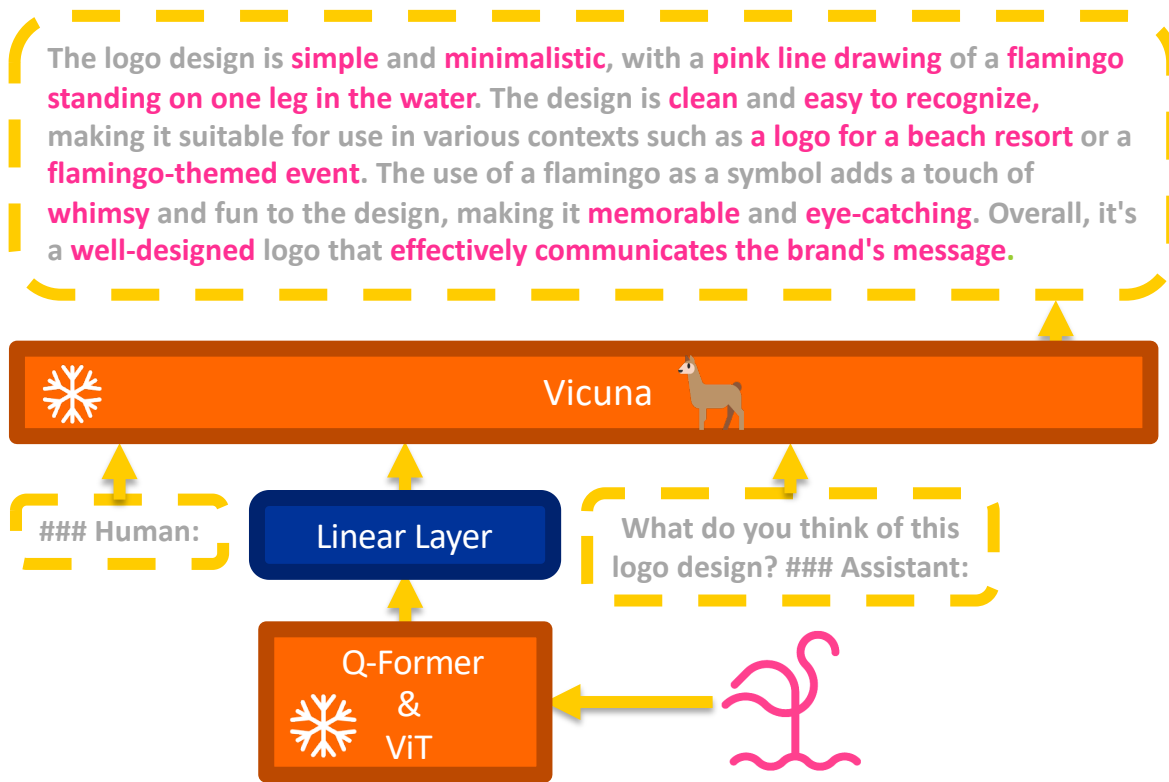


GitHub
300+★



Scan Me

- Take Blip-2's visual encoder. Freeze it.
- Take a powerful LLM Vicuna^[3]. Freeze it.
- Add a linear layer to connect modules.
- Enter also context texts.
- Generate response.
- **How to train it?**

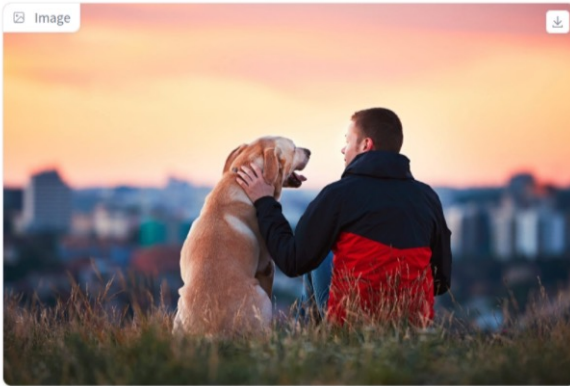


- **Traditional Alignment Method**
 - Training Dataset: Laion+CC+SBU
 - Input: Image Output: Caption
 - 10 hours training with 4 A100 GPUs

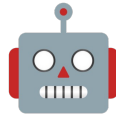


Stage 1 Issue

- MiniGPT-4 can see images now.
- But it forgets how to talk in a natural way!



Write me a poem for this image



A man sits with his dog on the

- Need a way to fix it

Stage 2 Dataset Preparation

- Create a small vision-language dataset with detailed, human-preferred descriptions.

Stage-1 Model
Generates
Descriptions



Polishing and
Filetering By ChatGPT



Further Polishing and
Filtering By Rules &
Human

Prompt

```
###Human: <img> </img>  
Describe this image in detail. Give  
as many details as possible. Say  
everything you see.  
###Assistant: <Description 1>  
###Human: Continue  
###Assistant: <Description 2>
```

Prompt

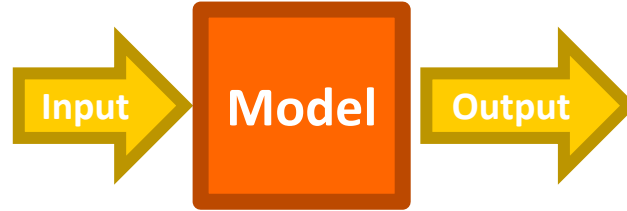
```
Fix the error in the given paragraph.  
Remove any repeating sentences,  
meaningless characters, not English  
sentences, and so on. Remove  
unnecessary repetition. Rewrite any  
incomplete sentences. Return directly  
the results without explanation. Return  
directly the input paragraph if it is  
already correct without explanation
```



This is an image of an elderly man with a beard and wearing a hat, standing in front of a mountainous backdrop with a village in the distance. The man is looking directly at the camera with a serious expression on his face.


Stage 2 Finetuning with Template

Stage 1



Young woman trekking through the forest

Stage 2

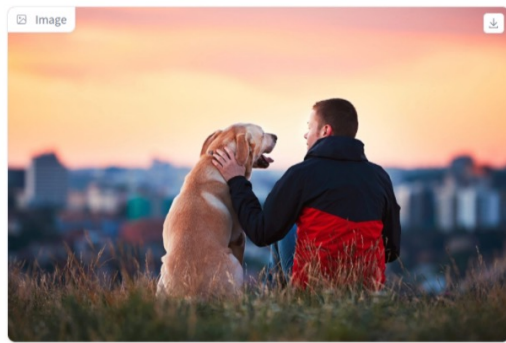
###Human:  Could you describe the contents of this image for me?

###Assistant:

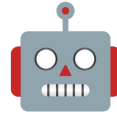


This is an image of an elderly man with a beard and wearing a hat, standing in front of a mountainous backdrop with a village in the distance. The man is looking directly at the camera with a serious expression on his face.

Stage 2 Improvement



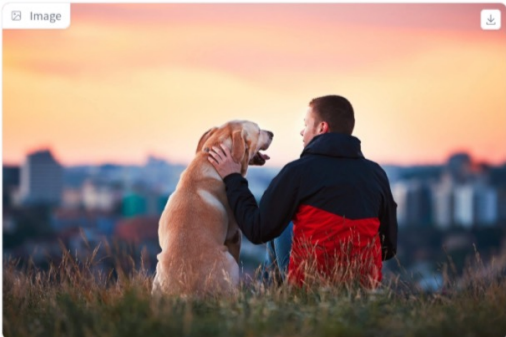
Write me a poem for this image



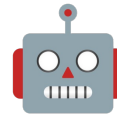
A man sits with his dog on the



Stage 2, **7 mins** in 1 A100
80G



Write me a poem for this image



A man sits on a hill, overlooking the city below

The sun sets in a blaze of orange and gold
The sky is painted with colors of light
The man holds his dog close, his face a smile
They sit in silence, watching the world go by

Demo: Describe an image in detail



Describe this image as detailed as possible.



The image shows a group of musicians performing on stage in front of a large audience. The musicians are playing electric guitars and are wearing black and white clothing. The audience is made up of people of different ages and ethnicities who are watching the performance and taking photos with their phones. The stage is lit up with bright lights and there is a large screen displaying the band's name and lyrics in Chinese characters. The atmosphere is lively and energetic, with the audience cheering and singing along to the music.

Demo: Explain why a meme is funny



Explain why this meme is funny.



This meme is funny because it shows a dog laying on the floor with its head resting on its paws, as if it is tired or sleepy. The caption, "monday just monday," adds to the humor by suggesting that the dog is feeling the same way as many people do on Mondays, which are often considered to be the most dreaded day of the week. The meme is relatable and humorous, making it a popular one among internet users.

Humorous Image captioning

ICCV23

Large Humorous Image
Captioning
Dataset

2.9 million image-text pairs

Collaboration with Oxford TVG

<https://torrvision.com/tvghic/>
<https://arxiv.org/pdf/2307.11636.pdf>

OxfordTVG-HIC: Can Machine Make Humorous Captions from Images?

[Runjia Li](#)^{1*} [Shuyang Sun](#)^{1*} [Mohamed Elhoseiny](#)² [Philip Torr](#)¹

¹University of Oxford ²KAUST *Equal Contribution

[\[Data\]](#) [\[Code\]](#) [\[PDF\]](#)

Demos

Demo of MiniGPT-4

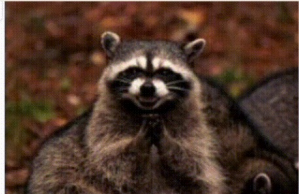
[NOTE] It is possible that you are waiting in a lengthy queue.
You can duplicate and use it with a paid private GPU.

[Duplicate this Space](#)

Alternatively, you can also use the demo on our [project page](#).

This is the demo of MiniGPT-4. Upload your images and start chatting!

[Project Page](#) [Github](#) [Code](#) [Paper](#) [PDF](#)

Image 

Start Chatting

Restart

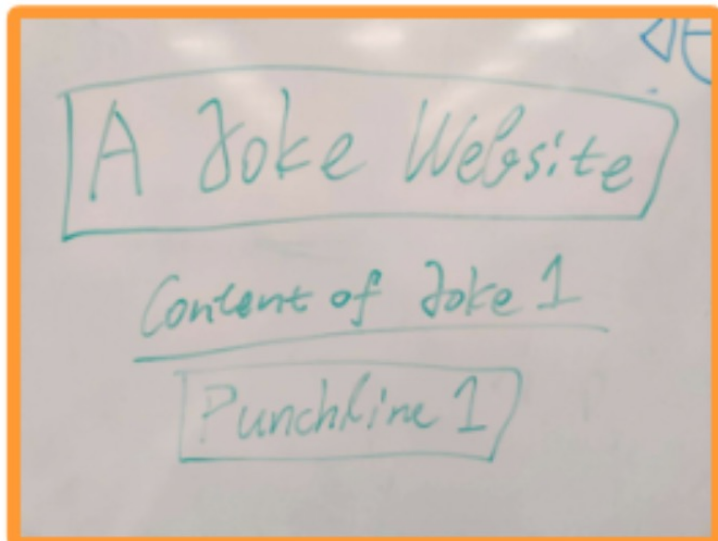
MiniGPT-4

can you generate a meme caption for this image?

Sure! Here's a meme caption for the image: "Raccoons are the best pets, they're so cute and cuddly!"

User

Type and press Enter



Write brief HTML/JS to turn this mock-up into a colorful website, where the jokes are replaced by a real joke



A Joke Website

Home About Contact

Today's Joke

Why did the tomato turn red?

Because it saw the salad dressing!

© 2023 A Joke Website

Demo: Problem solving



What is the issue with my plant? What should I do now?



The image shows a leaf with brown spots on it. The spots are likely caused by a fungal infection, which can be treated with a fungicide. It's important to identify the type of fungus causing the spots in order to choose the appropriate treatment.

To treat the plant, you can try the following steps:

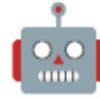


1. Identify the type of fungus causing the spots on the leaves.
2. Use a fungicide specifically labeled for the type of fungus you have identified.
3. Apply the fungicide according to the product's instructions, making sure to cover all affected leaves.
4. Repeat the treatment every 7 to 14 days, or as directed by the product's instructions, until the fungus is completely gone.
5. Water the plant well after treatment to help the fungicide penetrate the leaves.
6. Keep the plant in a well-ventilated area and away from direct sunlight to prevent overheating.
7. Monitor the plant's progress and repeat the treatment if new spots appear.

It's also important to keep the plant healthy by providing it with proper care, such as adequate watering, fertilization, and pruning. This will help the plant to recover and prevent future fungal infections.

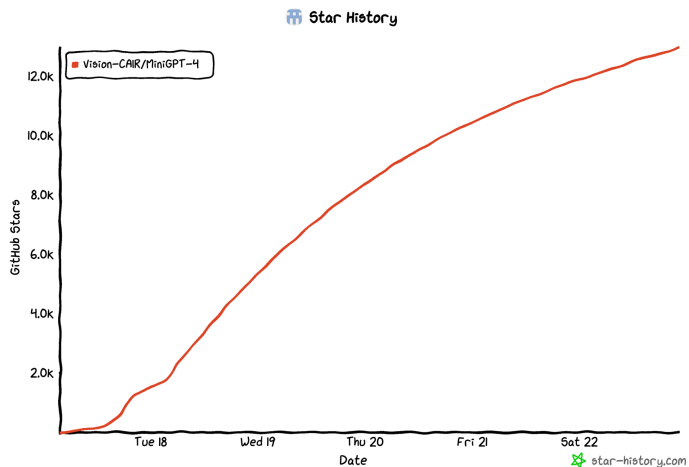


Help me write an advertisement for this.



Are you looking for a unique and stylish lighting solution for your home or office? Look no further than our brass toucan lamp! This beautifully crafted bird lamp is sure to add a touch of whimsy and charm to any room. The toucan's beak is designed to hold a light bulb, providing a warm and inviting glow. The base is made of wood, adding a natural touch to the overall design. The toucan lamp is a great addition to any room, whether you're looking to brighten up a dark corner or add a touch of personality to your space. Order yours today and bring a touch of the tropics to your home or office.

GitHub 10k 🌟 in 4 days!



Online Demo, Code, Pretrained Model, Dataset



Scan Me

Vision-CAIR / MiniGPT-4

Type to search

ode Issues 269 Pull requests 20 Discussions Actions Projects Wiki Security Insights Settings

MiniGPT-4 Public

Edit Pins Unwatch 212 Fork 2.8k Starred 23.8k



Table 1: Quantitative results on advanced vision-language tasks. MiniGPT-4 shows strong performance and successfully responses to 65% of the requests.

	Meme	Recipes	Ads	Poem	Avg.
BLIP-2	0/25	4/25	1/25	0/25	5/100
MiniGPT-4	8/25	18/25	19/25	20/25	65/100

Table 2: COCO caption evaluation. We use ChatGPT to judge if the generated caption covers all the visual objects and relations in the ground-truth caption.

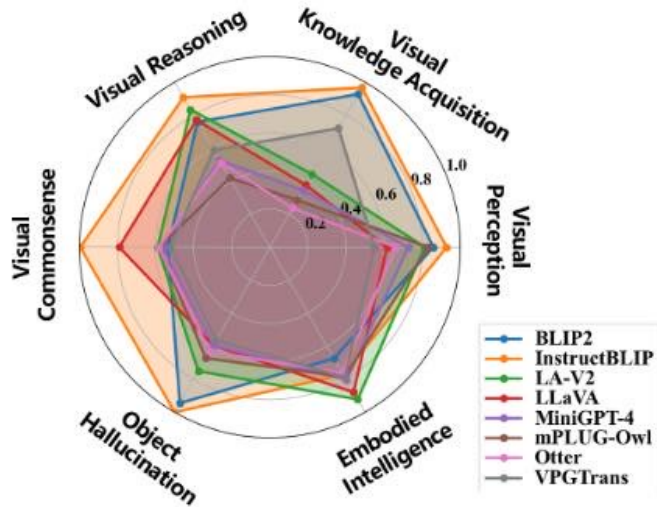
	BLIP-2	MiniGPT-4
Correctness	1376/5000	3310/5000
Percentage	27.5%	66.2%

Table 3: Failure rates of detailed caption and poem generation tasks before and after second-stage finetuning. The finetuning stage significantly reduces generation failures.

Failure rate	Detailed caption	Poem
Before stage-2	35%	32%
After stage-2	2%	1%

Experiments (in-domain. Vs out-of-domain)

In-domain



(a) Quantitative Capability Evaluation

Out-of-domain

Rank	Model	Score
1	mPLUG-Owl	1027.0
2	MiniGPT-4	1021.3
3	Otter	1013.2
4	LLaMA-Adapter V2	1010.2
5	LLaVA	1009.7
6	InstructBLIP	1003.7
7	VPGTrans*	974.3
8	BLIP2	949.4

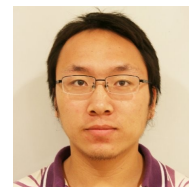
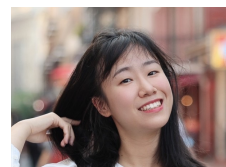
(b) LVLMs Arena Ranking up until June 3

Rank	Model	Score
1	MiniGPT-4	1022.9
2	LLaMA-Adapter V2	1021.2
3	VPGTrans	1017.0
4	mPLUG-Owl	1011.8
5	LLaVA	1008.9
6	Otter	1006.6
7	InstructBLIP	994.9
8	BLIP2	916.7

(c) LVLMs Arena Ranking up until June 13

MiniGPT-4-v2:

Large Language Model as a Unified Interface for Vision-Language Multi-task Learning

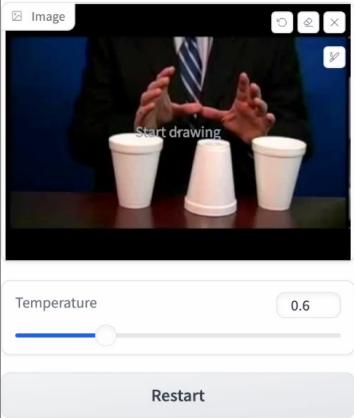


Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang,



Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong*, Mohamed Elhoseiny*

MiniGPT4-v2, September, 2023



Image

Start drawing

Temperature 0.6

Restart

For Abilities Involving Visual Grounding:

1. Grounding: CLICK **Send** to generate a grounded image description.
2. Refer: Input a referring object and CLICK **Send**.
3. Detection: Write a caption or phrase, and CLICK **Send**.
4. Identify: Draw the bounding box on the uploaded image window and CLICK **Send** to generate the bounding box. (CLICK "clear" button before re-drawing next time).
5. VQA: Input a visual question and CLICK **Send**.
6. No Tag: Input whatever you want and CLICK **Send** without any tagging

You can also simply chat in free form!

MiniGPT-v2

Task Shortcuts

No Tag Grounding Refer Detection Identify VQA

Hint: Type in a phrase about an object in the image and send the command

Upload your image and chat

Send

MINIGPT-V2: LARGE LANGUAGE MODEL AS A UNIFIED INTERFACE FOR VISION-LANGUAGE MULTI-TASK LEARNING



[detection] The image showcases a living room featuring a lamp, a spacious gray sofa, two wall paintings, and shoes resting on a blanket



a living room
a lamp
a spacious gray sofa
two wall paintings
shoes
a blanket

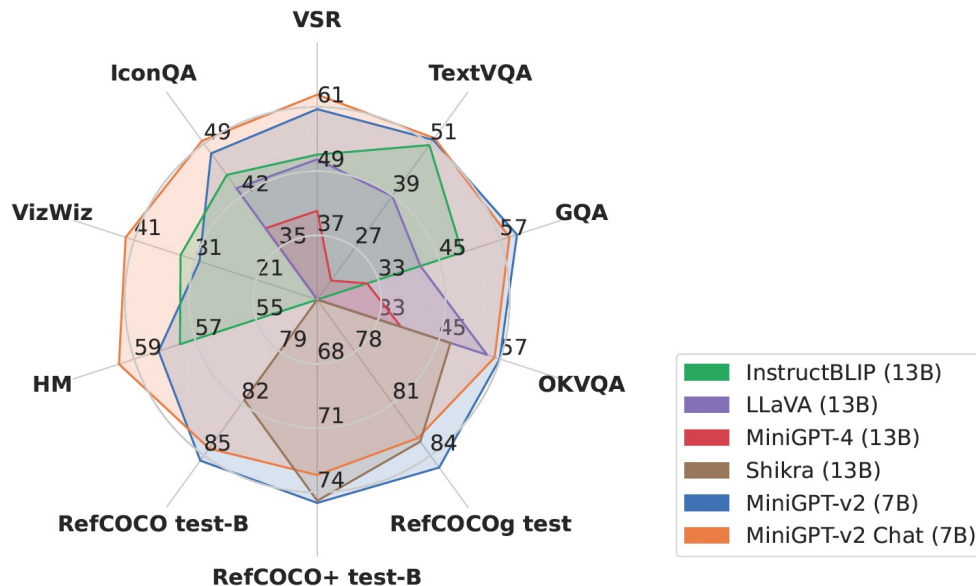
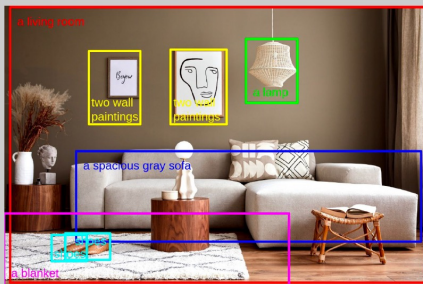


Figure 1: Our MiniGPT-v2 achieves state-of-the-art performances on a broad range of vision-language tasks compared with other generalist models.

Method	Grounding	OKVQA	GQA	VSR (zero-shot)	IconVQA (zero-shot)	VizWiz (zero-shot)	HM (zero-shot)
Flamingo-9B	✗	44.7	-	31.8	-	28.8	57.0
BLIP-2 (13B)	✗	45.9	41.0	50.9	40.6	19.6	53.7
InstructBLIP (13B)	✗	-	49.5	52.1	44.8	33.4	57.5
MiniGPT-4 (13B)	✗	37.5	30.8	41.6	37.6	-	-
LLaVA (13B)	✗	54.4	41.3	51.2	43.0	-	-
Shikra (13B)	✓	47.2	-	-	-	-	-
Ours (7B)	✓	56.9	60.3	60.6	47.7	32.9	58.2
Ours (7B)-chat	✓	57.8	59.9	60.4	49.8	53.6	58.8

Table 3: **Results on multiple VQA tasks.** We report top-1 accuracy for each task. Grounding column indicates whether the model incorporates visual localization capability. The best performance for each benchmark is indicated in **bold**.

Method	Model types	RefCOCO			RefCOCO+			RefCOCOg		Avg
		val	test-A	test-B	val	test-A	test-B	val	test	
UNINEXT	Specialist models	92.64	94.33	91.46	85.24	89.63	79.79	88.73	89.37	88.90
G-DINO-L		90.56	93.19	88.24	82.75	88.95	75.92	86.13	87.02	86.60
VisionLLM-H	Generalist models	-	86.70	-	-	-	-	-	-	-
OFA-L		79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58	72.65
Shikra (7B)		87.01	90.61	80.24	81.60	87.36	72.12	82.27	82.19	82.93
Shikra (13B)		87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16	83.96
Ours (7B)		88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66	84.29
Ours (7B)-chat		88.33	91.71	84.57	79.77	85.43	73.04	83.92	84.03	

Table 4: **Results on referring expression comprehension tasks.** Our MiniGPT-v2 outperforms many VL-generalist models including VisionLLM (Wang et al., 2023), OFA (Wang et al., 2022) and Shikra (Chen et al., 2023b) and reduces the accuracy gap comparing to specialist models including UNINEXT (Yan et al., 2023) and G-DINO (Liu et al., 2023c).

MiniGPT4-v2, code & examples

Edit Pins Unwatch 207 Fork 2.7k Starred 22.8k

Go to file Add file Code

About

Open-sourced codes for MiniGPT-4 and MiniGPT-v2

minigt-4.github.io

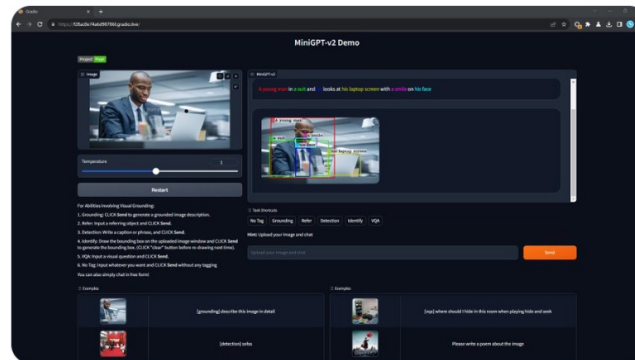
a2f0a44 4 hours ago 135 commits

2 months ago

camenduru @camenduru · 20h

MiniGPT-v2 Colab 🤖 Thanks to @garvinchen2 ❤️ @tikgiau ❤️ @xiaoqian_shen ❤️ @lix709 ❤️ @zechunliu ❤️ @PengchuanZ ❤️ Raghuraman Krishnamoorthi ❤️ @vikasc ❤️ @YoungXiong1 ❤️ @moElhoseiny ❤️

page: minigt-v2.github.io
colab: please try it 🤖 github.com/camenduru/Mini...



Jun (Garvin) Chen @garvinchen2 · Oct 13

Excited to share our MiniGPT-v2. It can perform many complex vision-language and visual grounding tasks with simple interaction

Demo & Project: minigt-v2.github.io...
[Show more](#)

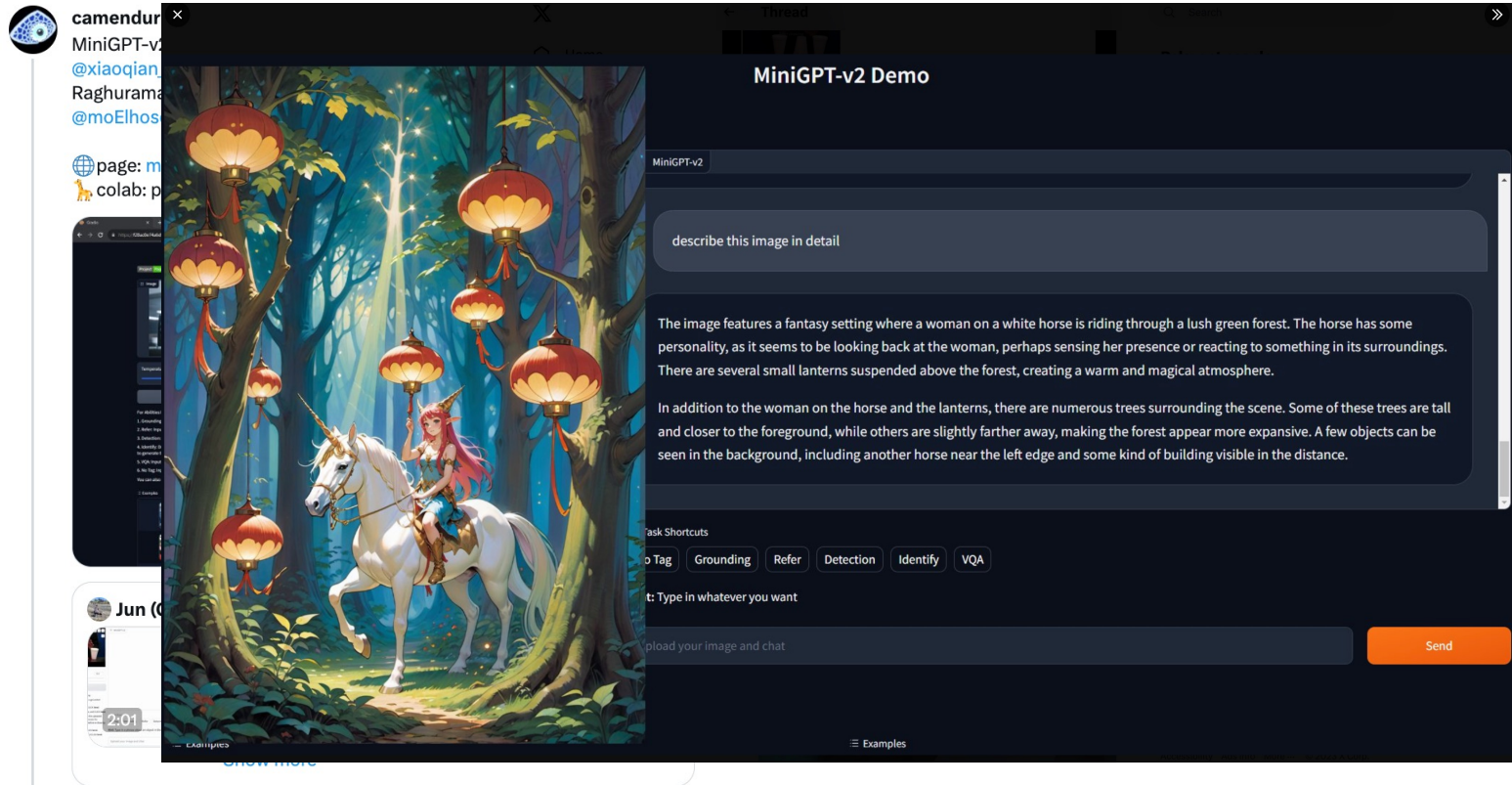
MiniGPT4-v2, Examples

The screenshot displays a social media post from user @camenduru. The post features a 'MiniGPT-v2 Demo' interface. On the left, there is a vibrant illustration of a woman sitting on a rock by a lake at sunset. The interface on the right has a dark theme and includes the following elements:

- Header:** 'MiniGPT-v2 Demo'
- Input:** A text box containing the prompt 'describe this image in detail'.
- Output:** A detailed AI-generated description: 'A woman is sitting on a rock near the water, surrounded by a serene landscape of hills and grassy fields. She appears to be enjoying her time in the quiet scenery. The water is glistening and reflecting the warm sunlight that illuminates the image. The woman is close to the shore, and a few trees can be seen nearby. In the background, there are two birds flying through the sky, adding a sense of movement and life to the tranquil scene. Overall, this illustration captures the beauty and peacefulness of a natural setting.'
- Task Shortcuts:** A row of buttons labeled 'No Tag', 'Grounding', 'Refer', 'Detection', 'Identify', and 'VQA'.
- Instructions:** A line of text stating 'Prompt: Send the command to generate a grounded image description'.
- Input Field:** A text box with the placeholder 'Upload your image and chat' and an orange 'Send' button.

The social media post also shows the user's profile, the artist's name 'ayaya_ayay@discord', the time '5:58 PM · Oct 15, 2023', and '982 Views'. There are 4 likes and 3 bookmarks visible.

MiniGPT4-v2, Examples



camenduru
MiniGPT-v2
@xiaoqian
Raghurama
@moElhos

page: m
colab: p

MiniGPT-v2 Demo

describe this image in detail

The image features a fantasy setting where a woman on a white horse is riding through a lush green forest. The horse has some personality, as it seems to be looking back at the woman, perhaps sensing her presence or reacting to something in its surroundings. There are several small lanterns suspended above the forest, creating a warm and magical atmosphere.

In addition to the woman on the horse and the lanterns, there are numerous trees surrounding the scene. Some of these trees are tall and closer to the foreground, while others are slightly farther away, making the forest appear more expansive. A few objects can be seen in the background, including another horse near the left edge and some kind of building visible in the distance.

Task Shortcuts
Tag Grounding Refer Detection Identify VQA

t: Type in whatever you want

upload your image and chat

Send

Examples



camenduru
@camenduru

Artist: ayaya_ayay@discord

5:58 PM · Oct 15, 2023 · 1,040 Views

2 8 3

Post your reply Reply

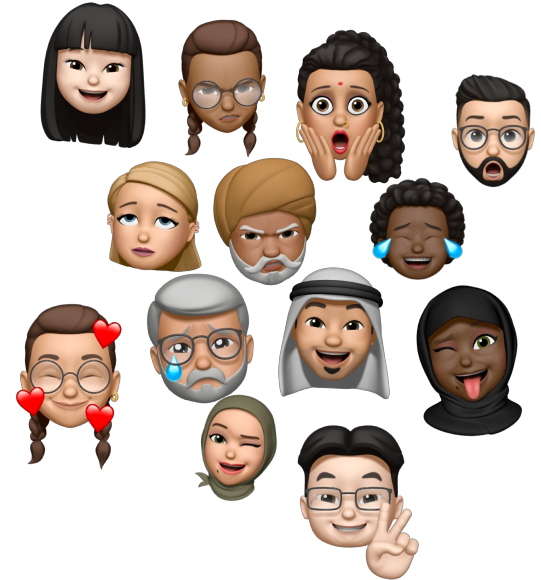
Jun (Garvin) C... @garvinch... · 19h ...
those are great examples!
3 75

Show more replies

No Culture Left Behind Emotion-Aware AI

HOW NO CULTURE LEFT BEHIND EMOTION-AWARE AI?

- DIFFERENT FROM NO LANGUAGE LEFT BEHIND, EMOTION DISTRIBUTION IS DIFFERENT FOR EACH CULTURE? VERY DIFFERENT AI SKILL COMPARED TO TRANSLATION AND MUCH HARDER.
- HOW TO EXTEND TO MANY MORE LANGUAGES TO COVER AS MANY CULTURES TO BUILDING INCLUSIVE PLATFORMS/ METAVERSE ? MUCH MORE CHALLENGING COMPARED TO TRANSLATION



ArtElingo-23: Towards No Culture Left Behind

- SO FAR 23 LANGUAGES, 100K+ ANNOTATIONS



Youssef Mohamed



Kenneth Ward Chui



Mohamed Elhoseiny

꽃차례가 없어서 무언가를 바라 보고 있는 여자의 얼굴이 만족스러워

Translation: The face of a woman sitting upright looking at something is satisfying

Korean

Morris wa mosetsana yo ga wa nna sentle, ka mokgwa o o ntseng ka teng o a tshegisla, e kete o ne a fofa

Translation: the hair is not well organized, the manner in which it's organized is funny, as if the girl was lying

Setswana

isthombhe sibhe ingemava limnyama nentokazi ehle enamehlo agqamile izinwele ezlungiswe kahle futhi igogoke kahle inika umuzwa wokwaneliseka

Translation: This is a beautiful picture, with a dark background, the beautiful lady with bright eyes, has well-groomed hair and is well-dressed, this gives me a sense of satisfaction.

IsiZulu

Upande wa nyuma, kuna picha ya wenzi wa ndoa. Ingawa sasa mwanamke huyu ni mjane, anatabasamu akimfikiria mume wa ujana wake

Translation: Although this woman is now a widow, she smiles thinking of the husband of her youth.

Swahili

คุณยายใจดีมาก เธอต้องสวยแน่ๆ ตอนเธอยังเด็ก

Translation: Grandma looks very kind. She must be beautiful, when she was young

Thai

यह बुढ़िबौरत बहुत ही ज्यादा खुश नजर आ रही है

Translation: The old woman looks very happy

Hindi

ဆံပင်တိုစဉ် မျက်စိအဖျဉ်းကလေးများ တင့်တယ်သော်လည်း ရင်ဘတ်ကလေးက နေထွေးသော အဝတ်အစားက သရုပ်ပျက်နေသည်။

Translation: The short hair and facial expression are beautiful, but the clothes that reveal the big chest are decadent.

Burmese

gaun yang dipakai terlahu terdedah menampakkan bahagian lurahnya

Translation: The dress worn too exposed showing her cleavage

Malay

Umukecuru wicaye iruhande rw'idirishya iruhande rwaryo harambitse igikombe kirimo amazi n'indabo

Translation: An old woman is sitting by the window next to which lies a cup of water and flowers.

Kinyarwanda

Bà lão tóc bạc, da nhăn nheo nhón miêng cười khi ngắm nhìn những nụ hoa.

Translation: The old woman with silver hair and wrinkled skin smiled as she looked at the flower buds.

Vietnamese

حکومتی افسر دفتر میں موجود اپنی کرسی پر بیٹھا ہے جبکہ دائیں جانب کچھ میز اور کرسیاں بھی نظر آتی ہیں

Translation: A government official is sitting in his office chair while the business city is visible in the window on the right

Urdu

Resimdeki beyaz, kırmızı ve siyah tonları arasında canlı bir uyum var.

Translation: There is a vibrant harmony between the white, red and black tones in the picture.

Turkish

Kitang kita ang nasusunog na bahagi sa gubat at nakakatatat maisip na maari itong kumalat sa gubat.

Translation: I can see the burning part in the forest and it is scary to think that it can spread in the forest.

Tagalog

A na-egosi ulo na-agba oku n'ime abali, ndi bi na ya na-agbanahukwa ya

Translation: A house is shown burning at night, and its occupants are fleeing

Igbo

رجل جالس في بئر هائل شي حاجة ف نديه ومخزرف

Translation: A man is sitting in his fancy chair holding an expensive pen. He seems filthy rich.

Darija

Shaxsan men ushbu ijodkor shaxsdan judaxam mamnunman uning shu yoshidagi qilyotgan ijodini ko'rib ruxlandim.

Translation: Personally, I am very pleased with this creative person, I am inspired by his work at this age.

Uzbek

Mutumun na zaune kan kujera akwai kuma tebur inda ya ajiyye littafan sa, ga kuma wata kan kamar tana bada sadaka, jama'a sun gero.

Translation: The man was sitting on a chair and there was a table where he kept his books, and there was a woman there as if she was giving alms, the people were shocked.

Hausa

ஒரு இரவு நேரத்தில் காட்டுக்கு நடுவில் தீ கொழுந்து விட்டு எரிகிறது

Translation: One night a fire burns in the middle of the forest

Tamil

Iná ñ jó pápá o si lèè fa janba fun àgbègbè naa

Translation: The fire is burning and it can cause disaster for the area

Yoruba

Seseorang meninggalkan api unggun tanpa mematikannya. Akibatnya sebagian hutan terbakar habis.

Translation: Someone left the campfire without putting it out. As a result, part of the forest burned down.

Indonesian

Language	# Ann.	Language	# Ann.
Urdu	12	Uzbek	6
Indonesian	11	Korean	6
Hindi	10	Malay	6
Turkish	9	Hausa	5
Tamil	9	IsiZulu	5
Burmese	7	Darija	5
Vietnamese	7	Kinyarwanda	5
Thai	7	Swahili	5
Tagalog	7	Yoruba	5
Igbo	6	Setswana	4

Multilingual MiniGPT-4, 2023



Youssef Mohamed



Kenneth Ward Church



Mohamed Elhoseiny



Seen Language

خزانة صغيرة على طاولة خشبية تحمل مزهرية بها أزهار حمراء وبيضاء وزرقاء.
Translation: A small dresser on a wooden table holds a vase with red, white, and blue flowers.

Arabic

There are a bunch of flowers in a yellow vase on a table. It looks like something from a restaurant. The table has a yellow cloth on it.

English

花瓶里白色的百合和绿色的小花搭配着，让人感到很美。
Translation: The combination of white lilies and small green flowers in the vase looks nice.

Chinese



Seen Language

تصور اللوحة امرأة وطفل في مشهد عن قرب. المرأة تمسك ذراع الطفل في ذراعيها.
Translation: The painting depicts a woman and child in a close-up view. The woman is holding the child's arm in hers.

Arabic

The baby in the picture looks so calm with the mother closing her eyes and feeling peaceful and content.

English

图片描绘了母亲抚摸着她的孩子，他穿着白色短裤。母亲看起来很温柔。
Translation: The picture depicts a mother stroking her baby, who is wearing white shorts. Mother looks very gentle.

Chinese



Unseen Language

Ce tableau représente une femme vêtue d'une robe orange, assise dans le dos d'un enfant.
Translation: This painting shows a woman in an orange dress, seated behind the back of a child.

French

Esta pintura es una fotografía de un niño durmiendo con su madre.
Translation: This painting is a photograph of a boy sleeping with his mother.

Spanish

इस पेंटिंग में एक महिला मैनीक्योर किया हुआ चश्मा पहने हुए और गले में कपड़ा लपेटे हुए अपने बच्चे को देख रही है
Translation: In this painting, a woman is looking down at her child wearing manicured glasses, holding a cloth wrapped around her neck

Hindi

Multilingual MiniGPT-4, 2023

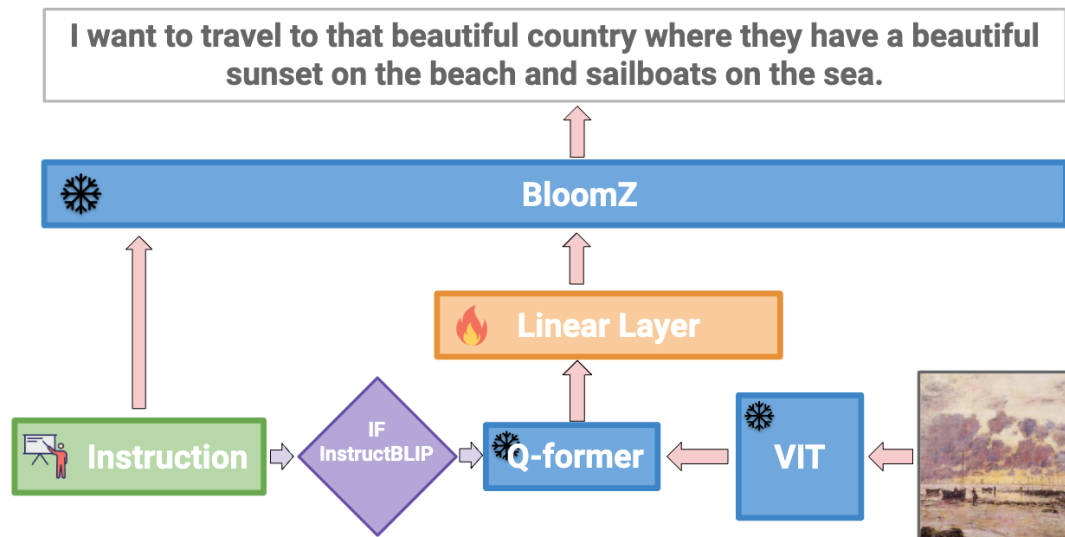


Figure 3: **Our model** is based on MiniGPT4. We train a linear projection layer to align the visual features from a Q-former with the input space of the language model. We use the multilingual Bloomz as our language model. In the version based on InstructBLIP we feed the caption to the Q-former as well.

Multilingual MiniGPT-4, 2023

CAPTIONING RESULTS

		MiniGPT4																										
Pretrain	LAION-English												LAION-2B-multi															
Finetune	CC-Align				ArtELingo				ArtELingo Pairs				CC-Align				ArtELingo				ArtELingo Pairs							
Test Set	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE
B_4	1.0	0.0	0.0	0.3	2.7	0.8	9.0	4.1	2.6	0.8	9.1	4.2	1.1	0.0	0.0	0.4	2.8	0.7	9.1	4.3	2.7	0.7	9.4	4.3				
M	21.6	0.3	0.2	7.4	20.5	7.8	24.4	17.6	20.6	7.9	24.8	17.8	21.3	0.2	0.2	7.2	20.4	7.8	24.6	17.6	20.6	7.7	25.0	17.8				
R	18.4	0.0	0.2	6.2	27.4	13.5	33.9	25.0	27.5	13.7	34.1	25.1	17.0	0.0	0.2	5.7	27.6	13.6	34.1	25.1	27.7	13.5	34.2	25.1				
C	0.2	0.0	0.0	0.0	4.6	4.0	8.7	5.7	4.4	4.1	8.7	5.7	0.1	0.0	0.0	0.0	4.6	4.0	9.0	5.9	4.7	3.9	8.9	5.8				
B_{Google}	3.1	0.0	0.0	1.1	7.4	3.8	11.5	7.6	7.3	3.8	11.6	6.9	3.0	0.0	0.0	1.0	7.3	3.8	11.6	7.6	7.4	3.8	11.8	7.6				

		InstructBLIP																										
Pretrain	LAION-English												LAION-2B-multi															
Finetune	CC-Align				ArtELingo				ArtELingo Pairs				CC-Align				ArtELingo				ArtELingo Pairs							
Test Set	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE	E	A	C	ACE
B_4	1.2	0.0	0.0	0.4	2.8	0.5	9.3	4.2	2.5	0.7	9.3	4.2	1.1	0.0	0.0	0.3	3.0	0.9	9.4	4.4	2.8	0.8	9.9	4.5				
M	22.7	0.3	0.1	7.7	21.6	8.4	25.6	18.5	22.0	8.2	25.3	18.5	22.4	0.2	0.1	7.6	21.3	8.5	25.9	18.5	22.7	8.7	26.0	19.2				
R	18.5	0.0	0.1	6.2	27.9	12.7	34.7	25.1	27.6	13.4	34.5	25.2	17.2	0.0	0.1	5.8	28.2	14.2	34.8	25.7	28.1	13.9	35.2	25.8				
C	0.2	0.0	0.0	0.1	4.4	3.0	8.6	5.3	4.2	4.0	9.1	5.8	0.2	0.0	0.0	0.1	4.8	4.7	9.8	6.4	4.6	4.5	9.9	6.3				
B_{Google}	3.4	0.0	0.0	1.1	7.6	3.1	12.0	7.6	7.5	3.7	11.8	7.7	2.9	0.0	0.0	1.0	7.6	4.1	12.0	7.9	7.7	3.9	12.2	8.0				

Table 1: **Multilingual Captioning Baseline.** MiniGPT4 and InstructBLIP are initially pre-trained using either LAION-English (specifically, LAION-Synthetic-115M) or LAION-2B-multi datasets. After this pre-training phase, they are further fine-tuned using one of three datasets: Conceptual Captions Align (CC-Align), ArtELingo, or ArtELingo Pairs. Notably, when the models are fine-tuned with ArtELingo Pairs, they are trained to produce output in two languages simultaneously, which facilitates the alignment of understanding across different languages. The trained models are evaluated on a test set from each language as well as a combined test set. For metrics, we use BLEU-4 (B_4), METEOR (M), ROUGE (R), CIDEr (C), and Google-BLEU (B_{Google}).

Multilingual MiniGPT-4, 2023

PERFORMANCE ON UNSEEN LANGUAGES

Train set	×		E		EA	EC
	A	C	A	C	C	A
B_4	0.0	0.0	0.0	0.4	0.4	0.0
M	0.0	0.0	1.0	4.9	5.6	1.6
R	0.0	0.0	2.3	6.5	8.6	2.9
C	0.0	0.0	0.5	0.8	1.2	0.5
B_{Google}	0.0	0.0	0.5	1.0	1.2	1.7

Table 2: **Unseen Language Performance.** Our MLLMs are pre-trained on LAION-2B-multi and fine-tuned on ArtELingo but with only part of the available languages. The trained models are evaluated on a test set made up of languages not appearing in the training set. For metrics, we use BLEU-4 (B_4), METEOR (M), ROUGE (R), CIDEr (C), and Google-BLEU (B_{Google}). Each column corresponds to a combination of train and test sets designed specifically for evaluating performance in an unseen language.

Multilingual MiniGPT-4, 2023



batu barabaka ndo ya ndimo yaki nyumba zinayo
za nyumba zikavu na kabaana na zija

Translation: Batu Barabaka is the land where the
houses have dry houses and cabanas. **Swahili**

розрастные сторон, немногие картники,
встроенных детей

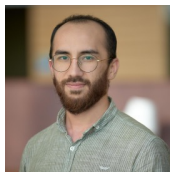
Translation: Big parties, few cards, built-in
children **Russian**

Figure 5: Qualitative Affective and Failed Samples

Affective Visual Dialog, 2023

Affective Visual Dialog (Arxiv23)

50,000 dialogs, 27,180 working hours : thanks Jack Urbanek, Mephisto team)



Kilichbek Haydarov, Xiaoqian shen, Avinash Madasu



Jia Li



Gamal Elsayed



Q Questioner

A Answerer



Pos: The men seems to be happy and enjoying the time together here, undisturbed.

Neg: The man is looking over his shoulder because he sees a threat

😊 contentment

😨 fear

Q1: How many people do you see in the image?

A1: I seen two people in the image

Q2: What are they doing in the image

A2: They are sitting on the boat, and they both are looking some where.

Q10: What is the background of the image?

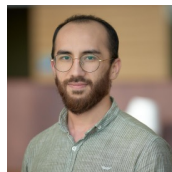
A10: Trees, clouds are presented by the background of the image.

<https://affective-visual-dialog.github.io/>

Affective Visual Dialog, 2023

Affective Visual Dialog (Arxiv23)

50,000 dialogs, 27,180 working hours : thanks Jack Urbanek, Mephisto team)



Kilichbek Haydarov, Xiaoqian shen, Avinash Madasu



Jia Li



Gamal Elsayed

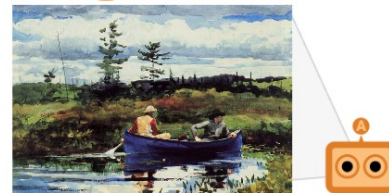


Q Questioner



Pos: The men seems to be happy and enjoying the time together here. undisturbed.

A Answerer



Neg: The man is looking over his shoulder because he sees a threat

Affective emotion explanations:



The two men are thinking about how to go home, I feel scared because the clouds are coming and it looks like the rain will fall and storms will come into the lake.



fear



No, It is not that I'm thinking it is cool weather and it seems they doing boating in the lake and enjoying their time



contentment



contentment

The men are sitting in the boat and seeing somewhere very interesting I think they are waiting for their friend to sail in the lake in the beautiful whether alone with the lake is surrounded by a lot of trees which looks very excellent so it gives me a good feeling



<https://affective-visual-dialog.github.io/>

Affective Chatbot : Emotion Guidance



Affective Chatbot : Emotion Guided Image Editing

sadness 🥹 → contentment 😊

- 🗨️ what else is there which is seeking more attention in the picture?
- ❌ womans eyes and the way is she holding her elbow with other hand 🤖
- ✅ there is a mountain in the picture 🤖

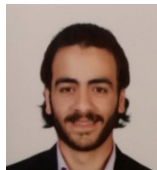
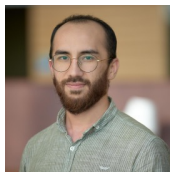


Figure 8: Example of altering answers to evoke opposite emotions, and make "edit" in the original image.

Introduced Datasets (Affective Vision and Language)

AI+Mental Health (2023)

Work in progress



Kilichbek Haydarov, Youssef Mohamed



Jia Li



Emilio Goldenhersch



MindCo Health



<https://affective-visual-dialog.github.io/>

Large Language Models as Consistent Story Visualizers



Xiaoqian shen



Mohamed Elhoseiny

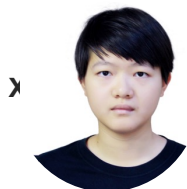
- Fred is standing in the living room while holding the phone and talking.
- He is in a room. He picks up the phone and then speaks into the phone.
- He stands next to a small table in the room. He holds the receiver for a phone while talking to someone. He then hangs up the phone when he finishes the call.
- Fred and Barney are standing in a room. There is a telephone next to Fred. Barney is talking with something in his hand.

- Barney is in the dining room at the table. He is holding a stack of papers and talking.
- He stands in the room, laughing at a newspaper.
- He opens a box while holding papers in a room. Then he hold the papers with both hands and laughs.
- Betty is sitting on a chair in the living room.

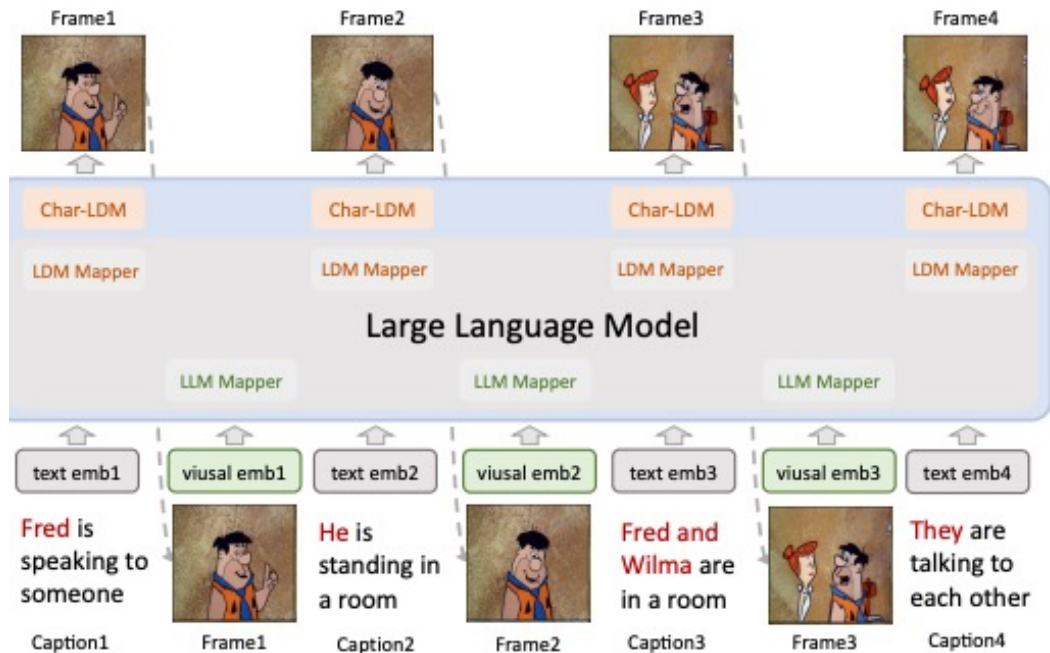


Large Language Models as Consistent Story Visualizers

StoryGPT-V



Mohamed Elhoseiny



Large Language Models as Consistent Story Visualizers

StoryGPT-V: can extend story in both vision and language



Fred is looking over the food on the table in the dining room.

Wilma is speaking to Fred in the dining room.

Fred is in the kitchen. He talks while looking at a giant pile on the table.

He is in the dining room. He puts his hands on his hips as he talks.

Wilma says (excitedly) Oh boy, look at all the food!

Wilma looks at Fred in disbelief.

Wilma rolls her eyes and mutters under her breath.

Fred looks at Wilma with a mischievous grin on his face.



Mohamed Elhoseiny



Xiaoqian shen

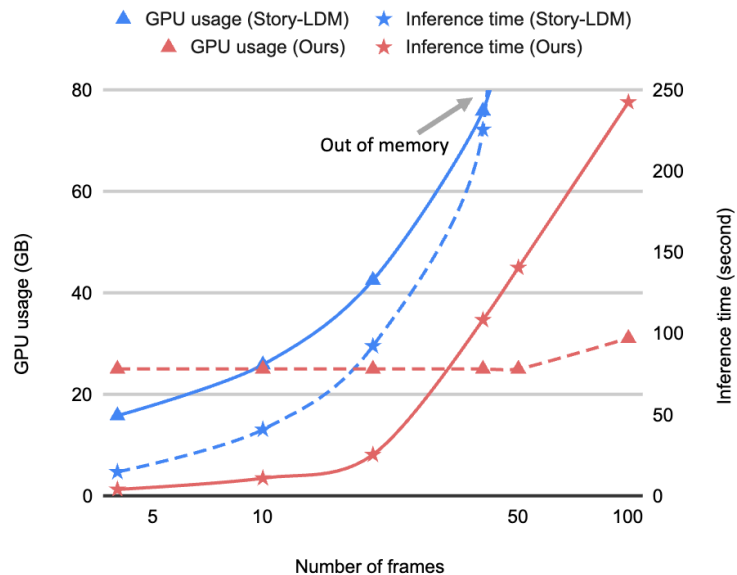


Figure 4. Compare inference speed and GPU memory consumption between our method and Story-LDM [14].

Large Language Models as Consistent Story Visualizers



Xiaoqian shen

Models	Char-Acc (↑)	Char-F1 (↑)	FID (↓)	BLEU4 (↑)	CIDEr (↑)
StoryDALL-E [†] [28]	21.03	50.56	40.39	0.2295	0.3666
LDM [41]	27.81	57.02	28.98	0.2560	0.5122
Story-LDM [38]	29.14	57.56	26.64	0.2420	0.4581
StoryGPT-V (Ours)	36.06	62.70	19.56	0.2586	0.5279

Table 2. Performance comparison on PororoSV [22] with co-referenced descriptions. [†]StoryDALL-E [28] takes the source frame as additional input.

Models	Ref text	Char-Acc (↑)	Char-F1 (↑)	BG-Acc (↑)	BG-F1 (↑)	FID (↓)	BLEU4 (↑)	CIDEr (↑)
StoryDALL-E [†] [28]		69.49	83.35	48.46	55.24	44.24	0.4666	1.4473
LDM [41]		85.66	93.41	54.85	62.04	32.05	0.5230	1.8048
Story-LDM [38]	×	82.43	91.86	55.3	61.58	36.29	0.4656	1.4335
Char-LDM (Ours)		90.36	95.76	58.36	63.92	21.13	0.5260	1.8361
StoryDALL-E [†] [28]		61.83	78.36	48.10	54.92	44.66	0.4460	1.3373
LDM [41]		75.37	87.54	52.57	58.41	32.36	0.4911	1.5103
Story-LDM [38]	✓	77.23	88.26	54.97	60.99	36.34	0.4585	1.4004
StoryGPT-V (Ours)		87.96	94.17	56.01	61.07	21.71	0.5070	1.6607

Table 1. Main experiments on FlintStonesSV [12]. The top portion is evaluated on the dataset w/o extended referential text. The bottom half displays the results on the extended dataset with co-reference. [†]StoryDALL-E [28] takes the source frame as additional input.

Large Language Models as Consistent Story Visualizers



Xiaoqian shen

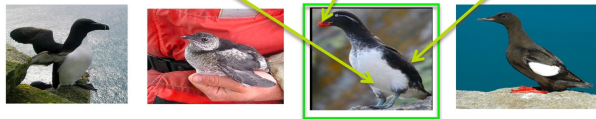
- Poby is seated beside a canvas. He holds a red pencil in his hand. There are many pictures on the wall.
- He is seated beside a canvas. He holds a red pencil in his hand. He lowers down his arm and makes a big smile. There are many pictures on the wall.
- Harry is in a house. Harry is seated on a green bed.
- He comes out of the house. He looks around the room. In the middle of the room, there is a wooden table. There is an apple on the table.



Imaginative Vision Language Models

IMAGINE TO SEE

Parakeet Auklet is a small bird that has an short orange bill. The bird's plumage is dark above and white below.



e.g. ICCV13, CVPR17, CVPR18, ICCV19, ICLR21, CVPR22, ECCV 2022



Makes me feel like birds are flying over my head.

IMAGINE TO CREATE



e.g. ICCV17, AAAI18, ECCV18, ICCV19, CVPR21, ICCV22, CVPR22, ICCV22

IMAGINE TO FEEL

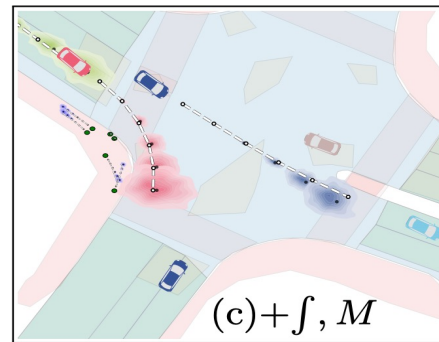
Affective Visual Art

ArtEmis CVPR21, ArtEmis-2.0 CVPR22
Wofflin Generative analysis, ICCV, 2021



The pale color palette of this watercolor painting is very relaxing. I can imagine myself sitting by the water *listening* to the birds.

IMAGINE TO DRIVE



e.g. CVPR2020, ICLR2021, CoRL21, ECCV22, ICLR23