MetaCLUE: Towards Comprehensive Visual Metaphors Research

(project page: https://metaclue.github.io)

Arjun Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, Varun Jampani

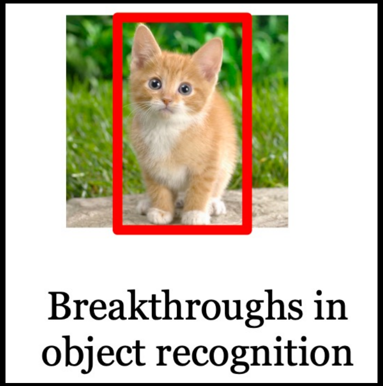
CVPR 2023

Poster Session: THU-PM-248

Google

Much of the computer vision is literal

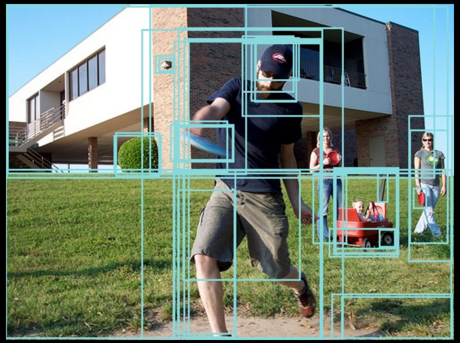
- Most CV datasets and tasks (Classification, Detection, Segmentation) for semantic understanding focus on literal semantics.



IMAGENET

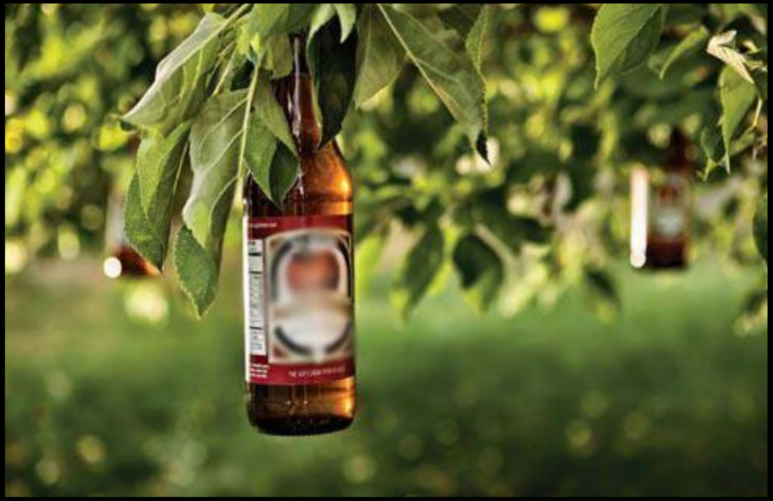
YouTube-VOS

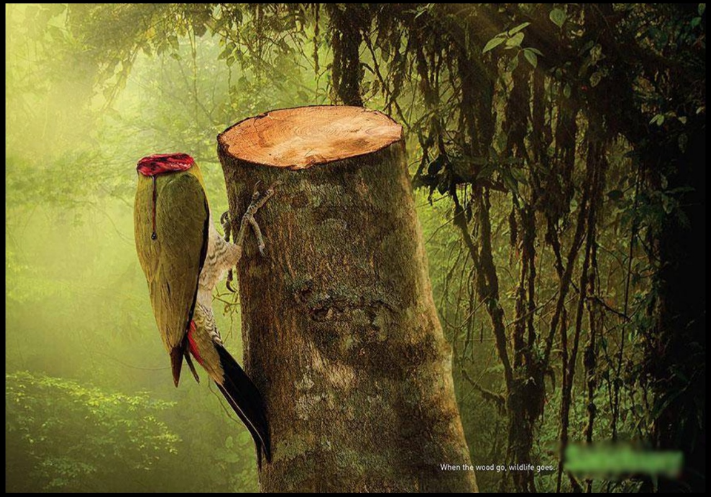
VQA

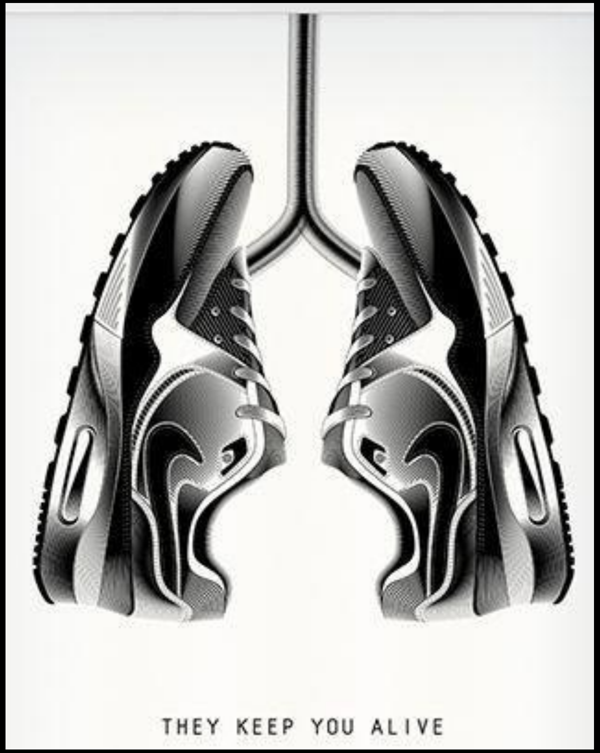


Visual Genome

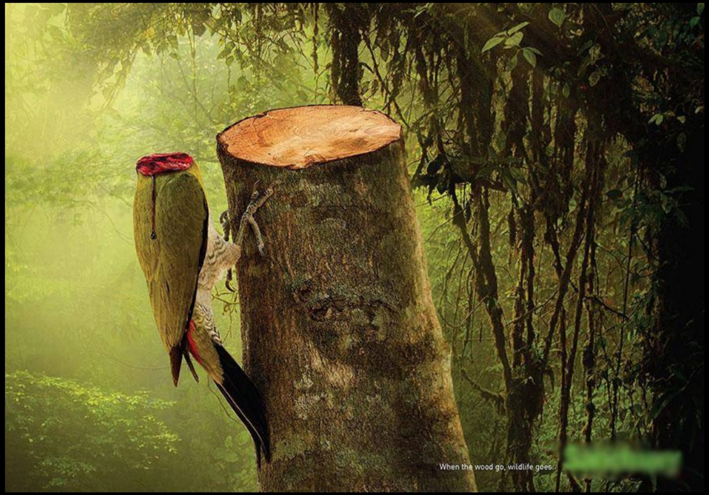
How about metaphorical images like these?







Introducing MetaCLUE



Metaphor: Killing trees is as harmful as killing wildlife

Classification

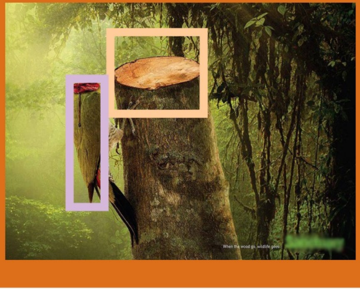
Is this a visual metaphor?

- YES
- NO

Localization

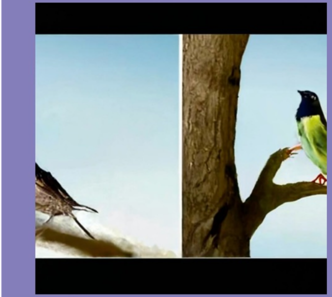
Detect image regions that invoke the concepts:

- Killing trees
- killing wildlife

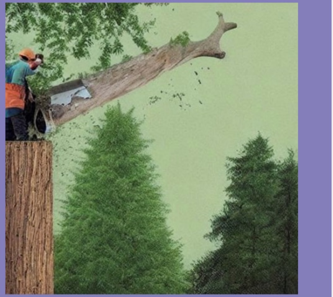


gEneration

Prompt: "An advertisement where killing trees is as harmful as killing wildlife."



Stable Diffusion



Imagen

Understanding

Retrieval

Pick the right one:

(a) Killing the forest is as deadly as killing the animals too.

(b) Birds is as much a part of our world as used cans.

Captioning

Sample predictions:

1. Deforestation is as damaging as killing wildlife.
2. Deforestation is as bad as ending the death penalty.

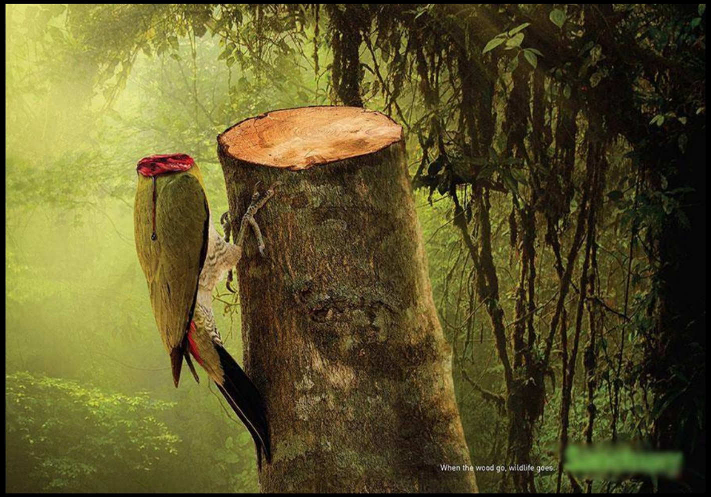
Visual Question Answering

Sample Questions:

Q. What is as harmful as killing wildlife?

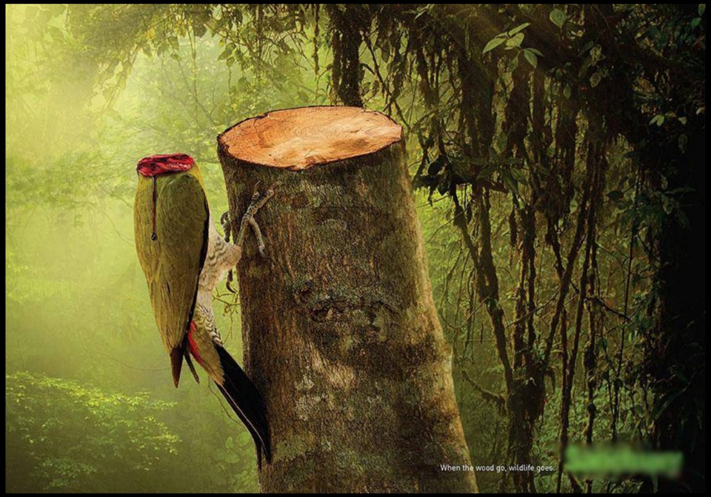
Q. What is compared to killing wildlife?

What is Visual Metaphor?



Metaphor Image

What is Visual Metaphor?

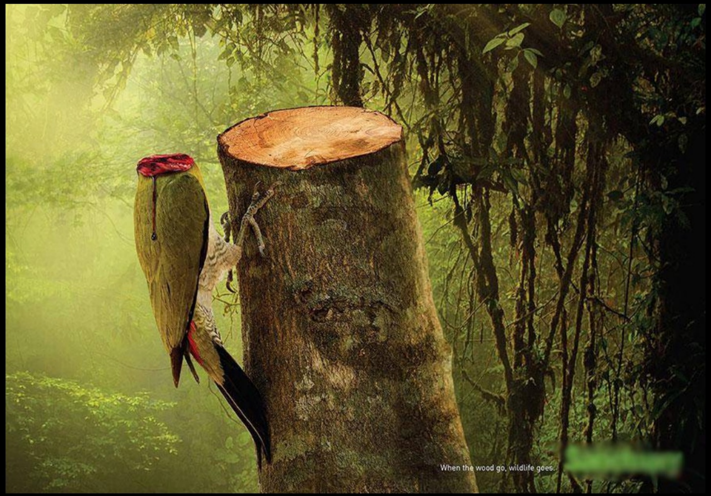


Metaphor Image

Interpretation of this Image?

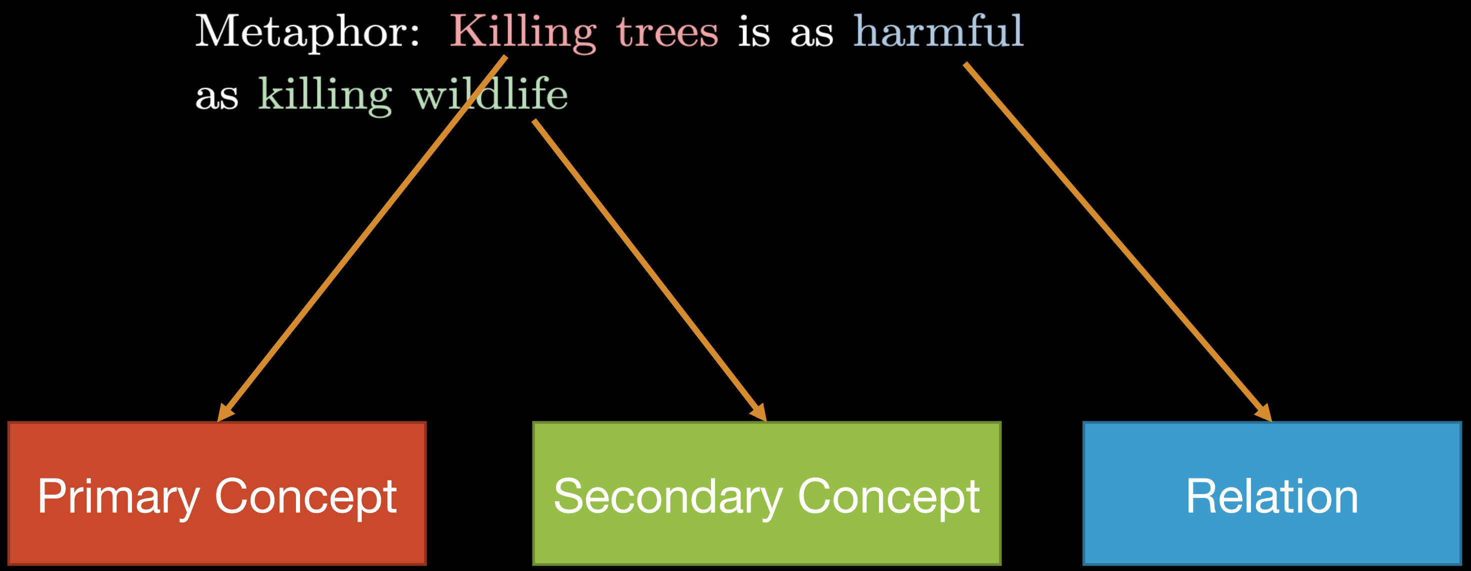
Metaphor: Killing trees is as harmful as killing wildlife

What is Visual Metaphor?



Metaphor Image

Interpretation of this Image?



Metaphor: Killing trees is as harmful as killing wildlife

Primary Concept

Secondary Concept

Relation

George Lakoff and Mark Johnson. Metaphors we live by. University of Chicago press, 2008.

Why we need Visual Metaphors?

Metaphors provide a sophisticated tool for nuanced human communication.

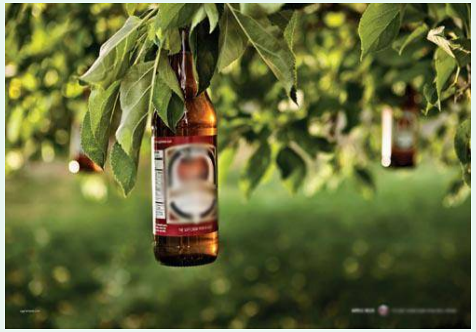
Metaphorical advertisements help in highlighting product attributes.

Examples:

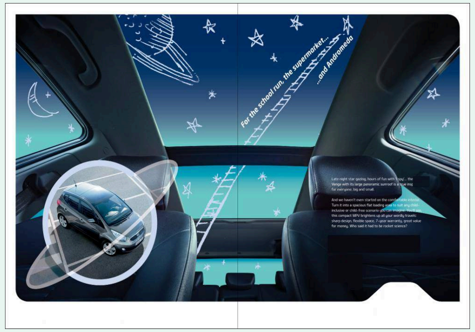
1) This car is *fast*
2) This phone is *futuristic*
3) This washer is *smart*

Types of Metaphor

Contextual



This beer is as tasty as a real apple.

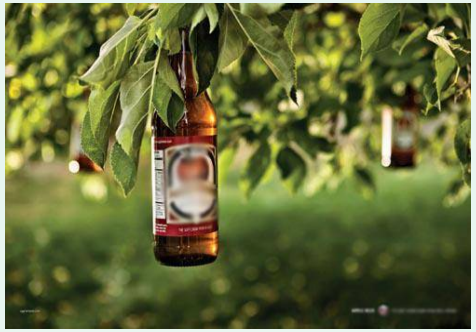


This car is as adventurous as a space ship.

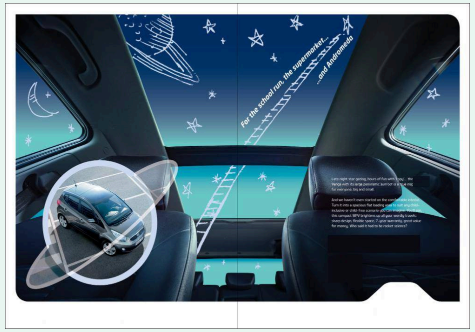
Charles Forceville. Pictorial metaphor in advertising. Routledge, 1996.

Types of Metaphor

Contextual

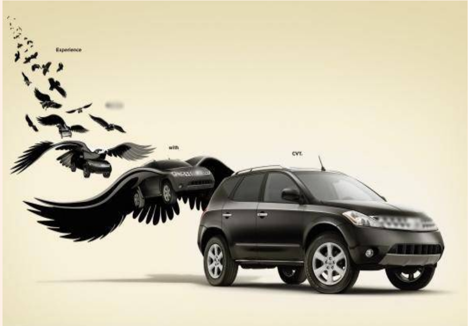


This beer is as tasty as a real apple.

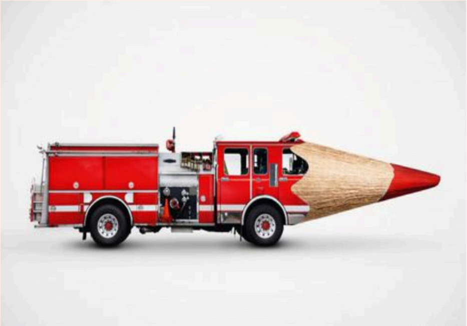


This car is as adventurous as a space ship.

Hybrid



Driving this SUV is as smooth as birds flying in the sky.

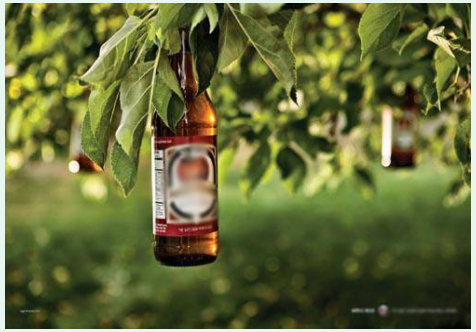


This pencil is as red as a fire truck.

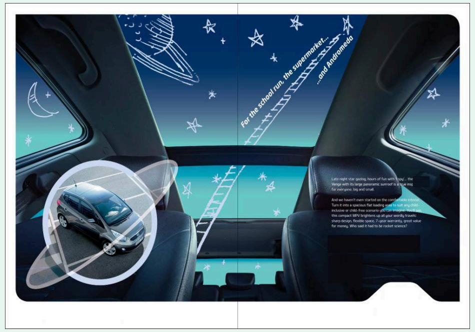
Charles Forceville. Pictorial metaphor in advertising. Routledge, 1996.

Types of Metaphor

Contextual

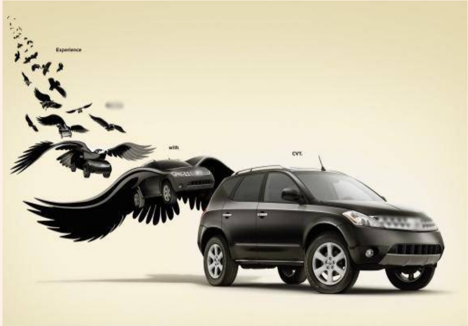


This beer is as tasty as a real apple.

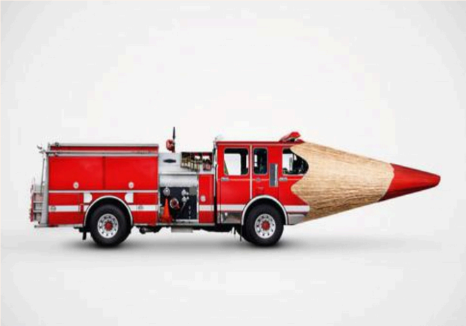


This car is as adventurous as a space ship.

Hybrid

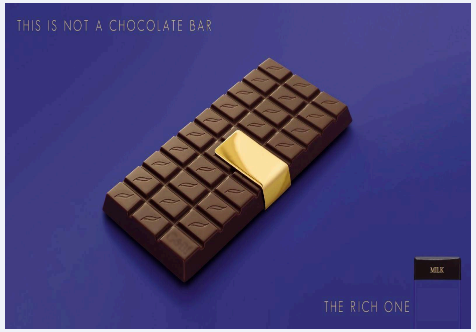


Driving this SUV is as smooth as birds flying in the sky.

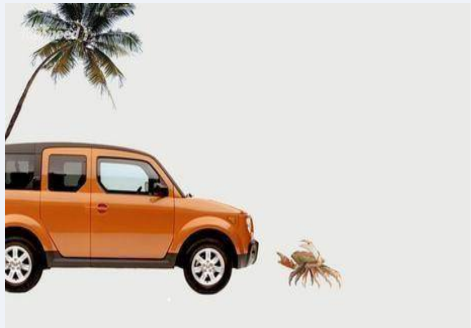


This pencil is as red as a fire truck.

Juxtaposition



This chocolate bar is as rich as gold.

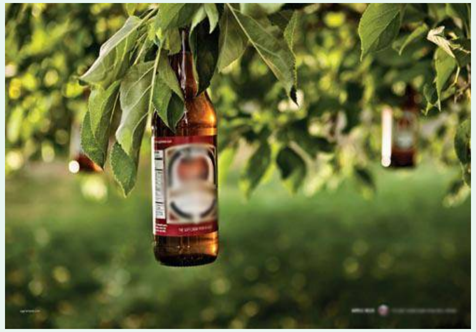


This car is as made for the beach as a crab.

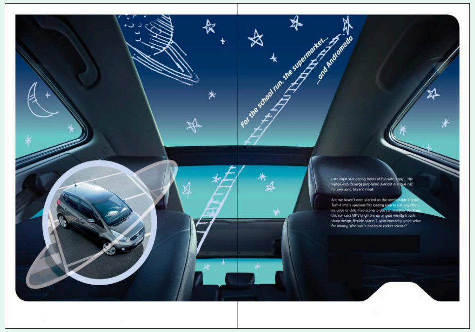
Charles Forceville. Pictorial metaphor in advertising. Routledge, 1996.

Types of Metaphor

Contextual

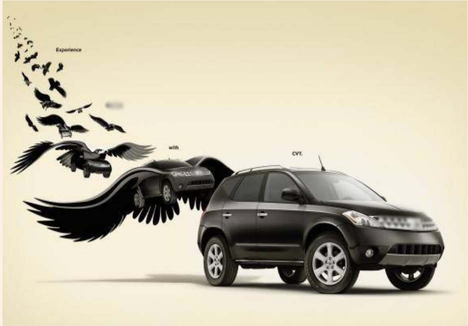


This beer is as tasty as a real apple.

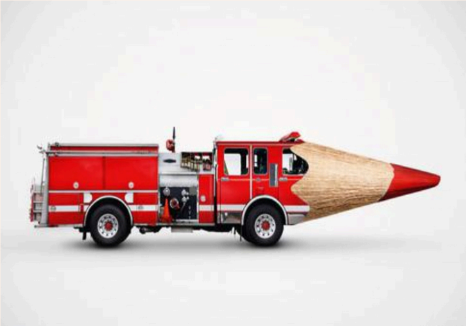


This car is as adventurous as a space ship.

Hybrid

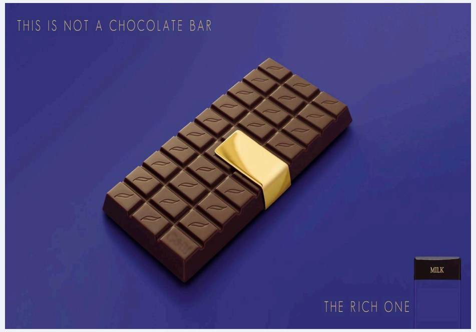


Driving this SUV is as smooth as birds flying in the sky.

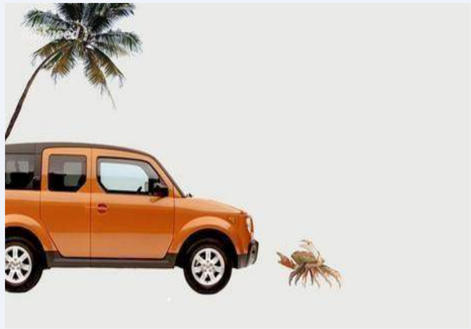


This pencil is as red as a fire truck.

Juxtaposition

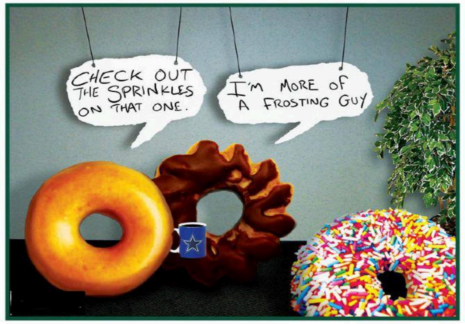


This chocolate bar is as rich as gold.

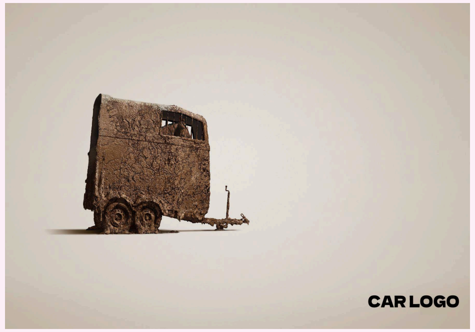


This car is as made for the beach as a crab.

Multimodal


CHECK OUT THE SPRINKLES ON THAT ONE.
I'M MORE OF A FROSTING GUY

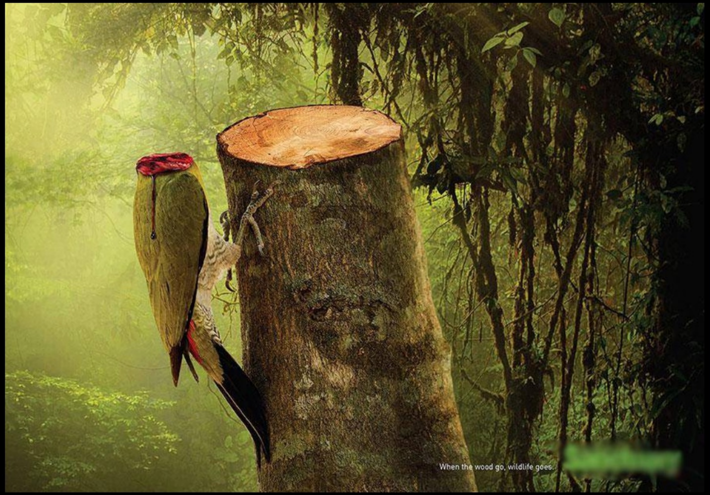
These donuts are as unique as as talking people.



The car is as rugged as this muddy trailer.

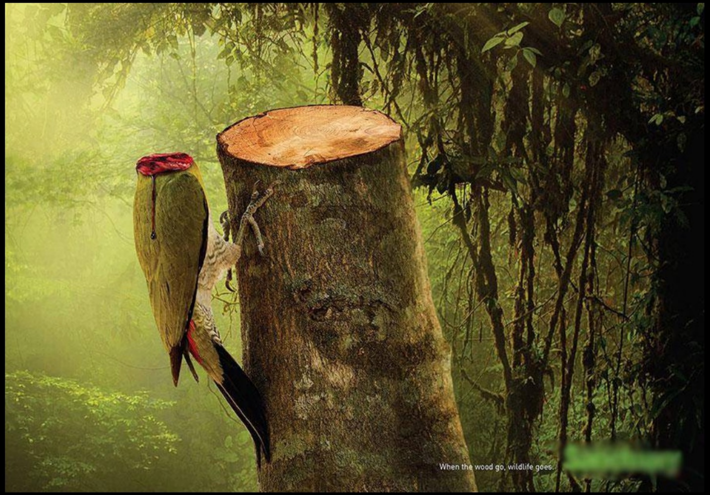
Charles Forceville. Pictorial metaphor in advertising. Routledge, 1996.

Introducing MetaCLUE



Metaphor: Killing trees is as harmful as killing wildlife

Introducing MetaCLUE



Metaphor: Killing trees is as harmful as killing wildlife

Classification

Is this a visual metaphor?

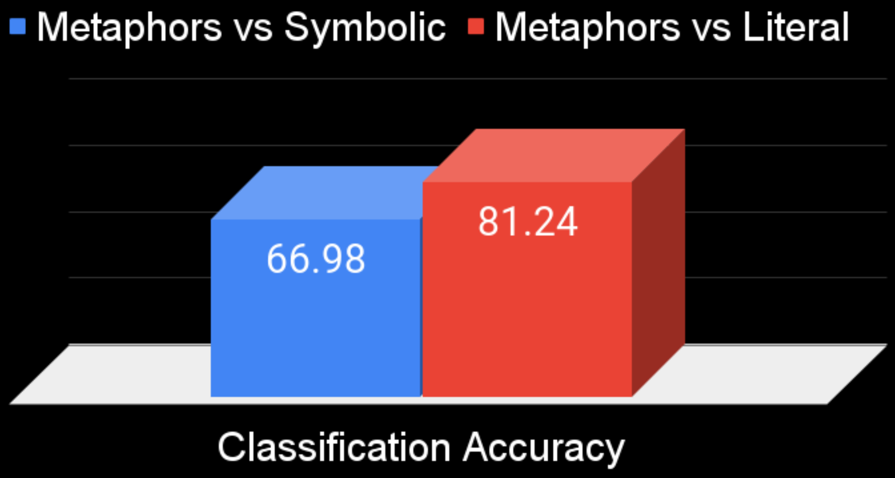
- YES
- NO

MetaCLUE Classification

Annotated Dataset

5061 metaphor Images
3.5k non-metaphor but symbolic images
2k non-metaphor, non-symbolic, literal images

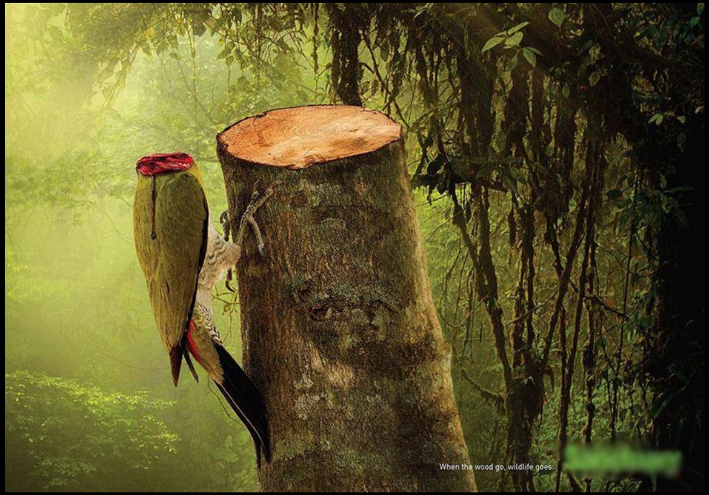
ViT-L/16 [1]



Models find it relatively easier to distinguish metaphor images from literal images, with more than 81% accuracy.

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Introducing MetaCLUE



Metaphor: Killing trees is as harmful as killing wildlife

Classification

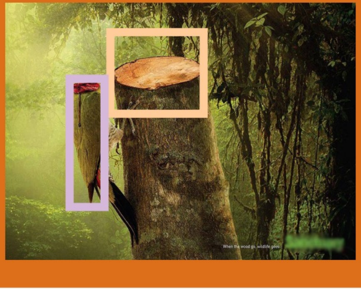
Is this a visual metaphor?

- YES
- NO

Localization

Detect image regions that invoke the concepts:

- Killing trees
- killing wildlife



MetaCLUE Localization

Two key differences compared to standard localization in literal images:

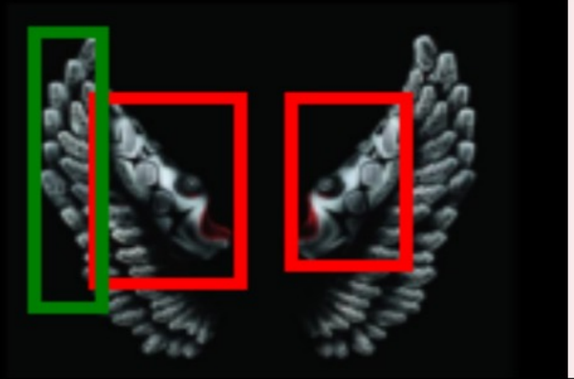
1. Explicit, Contextual boxes
2. Text, Logo boxes

Annotated Dataset

30k bounding box annotations

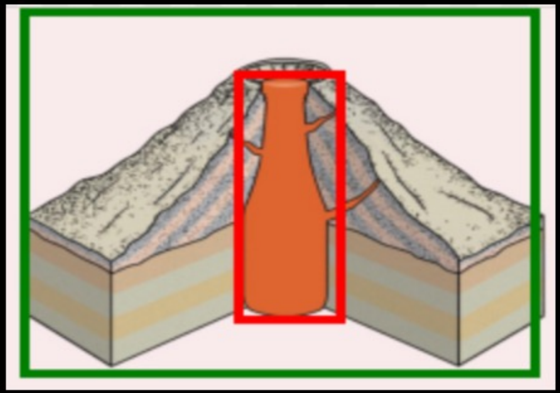
MetaCLUE Localization

Few Examples from our Annotations



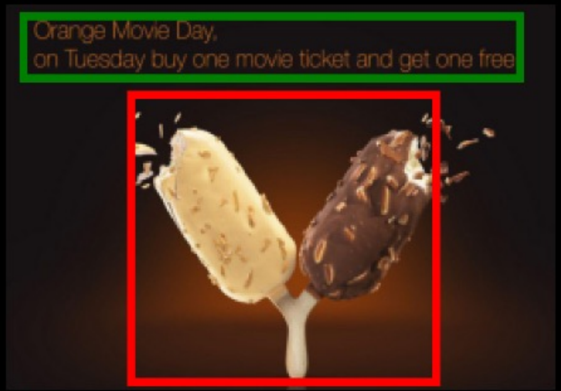
The shoe is as light as feathers

Explicit



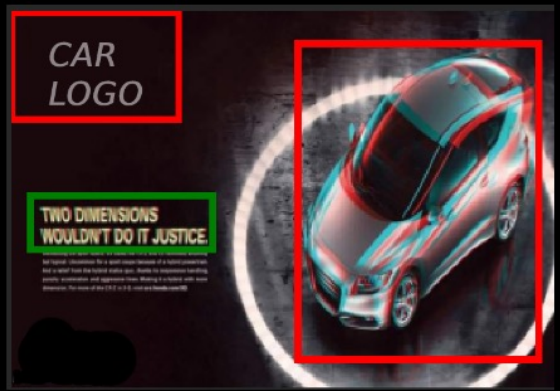
Ketchup is as hot as a volcano

Contextual



Getting two ice creams is as exciting as getting two movie tickets for one

Text



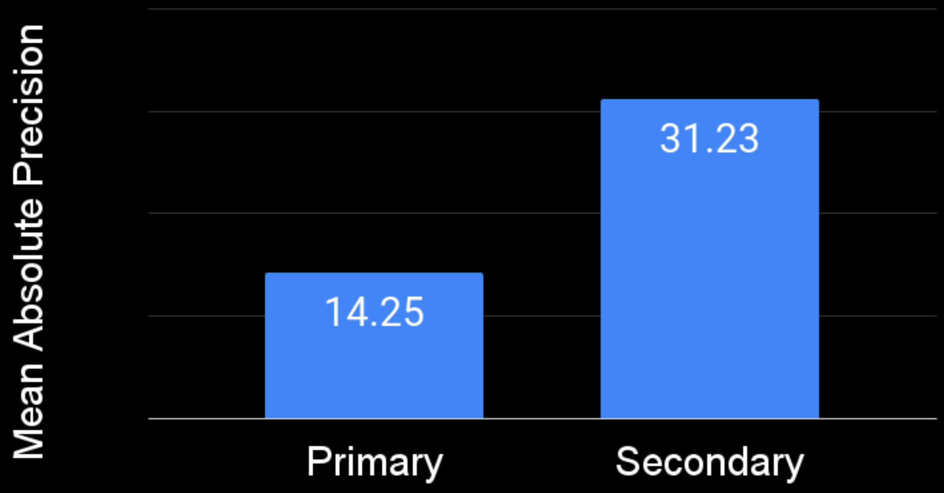
This car is as out of this world as three dimensional

Logo

MetaCLUE Localization

Results

CLIP based Phrase Grounding [1]

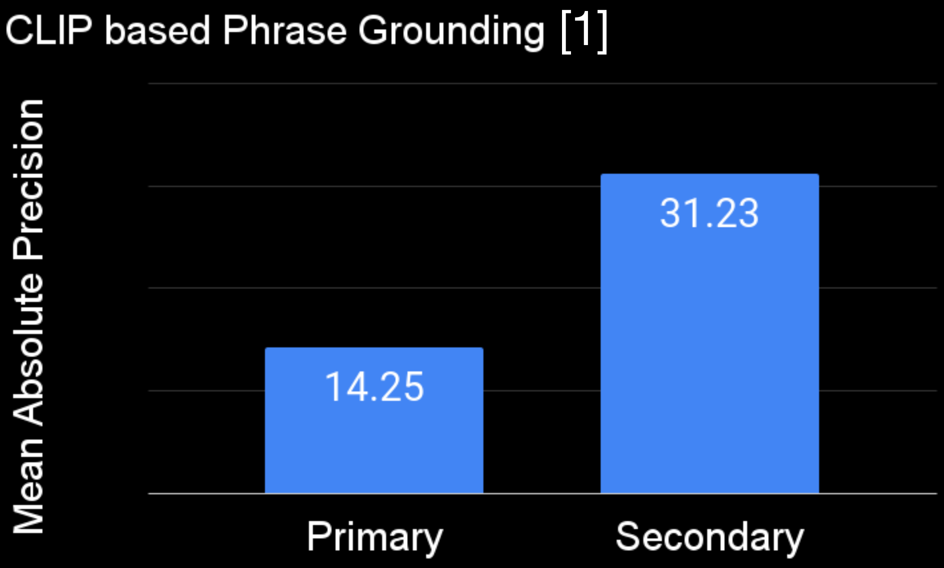


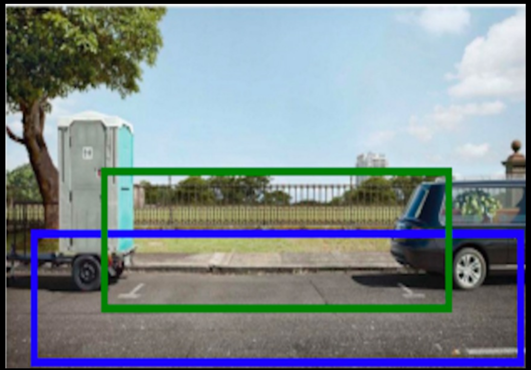
We find relatively better performance in localizing secondary objects compared to primary objects

[1] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. arXiv preprint arXiv:2204.03647, 2022.

MetaCLUE Localization

Results





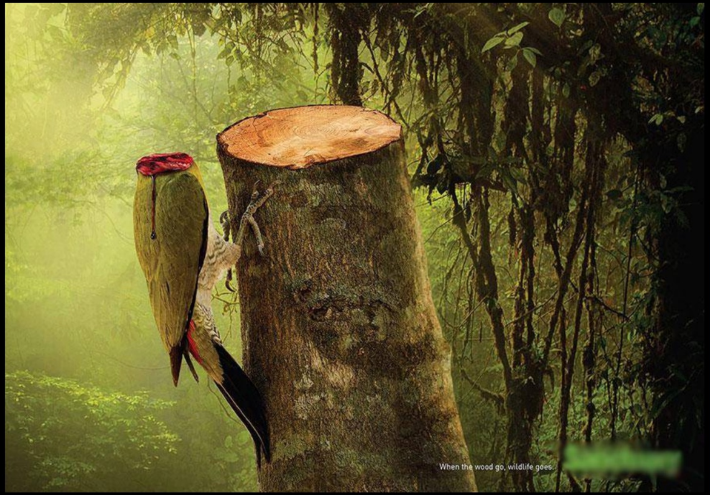
Parking assist is as safe as parking far from objects

This chocolate is as fun as going to a concert

We find relatively better performance in localizing secondary objects compared to primary objects

[1] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. arXiv preprint arXiv:2204.03647, 2022.

Introducing MetaCLUE



Metaphor: Killing trees is as harmful as killing wildlife

Classification

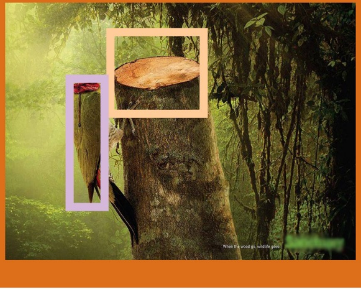
Is this a visual metaphor?

- YES
- NO

Localization

Detect image regions that invoke the concepts:

- Killing trees
- killing wildlife



Understanding

Retrieval

Pick the right one:

(a) Killing the forest is as deadly as killing the animals too.

(b) Birds is as much a part of our world as used cans.

Captioning

Sample predictions:

1. Deforestation is as damaging as killing wildlife.
2. Deforestation is as bad as ending the death penalty.

Visual Question Answering

Sample Questions:

Q. What is as harmful as killing wildlife?

Q. What is compared to killing wildlife?

MetaCLUE Retrieval

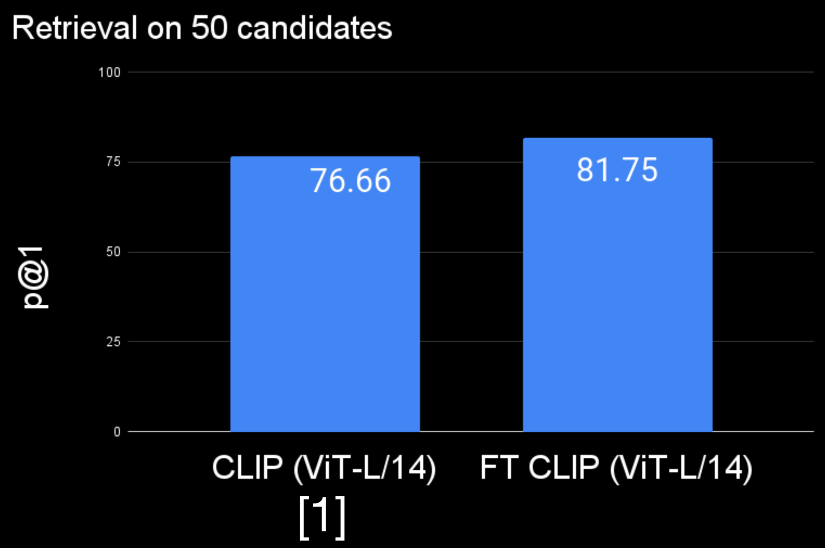
Annotated Dataset

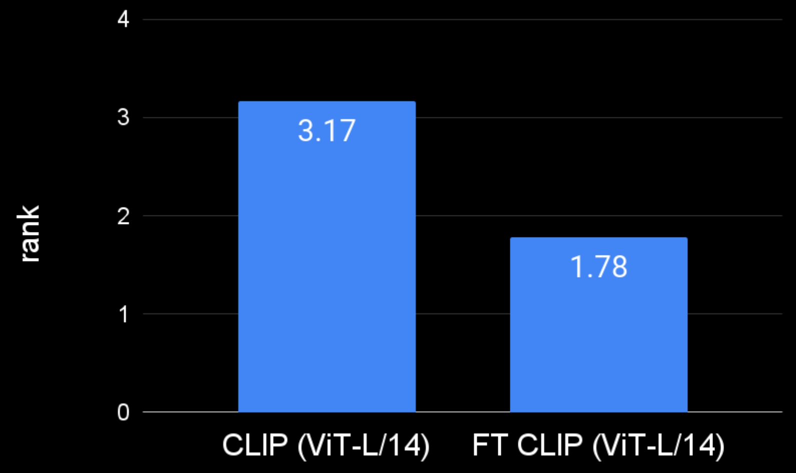
26k metaphor Annotations
Pilot studies
Multi-stage filtering

Validation Phase:

Is the grammar correct?
Are primary and secondary concepts correct?
Is the relationship correct?.

MetaCLUE Retrieval





Key Findings:

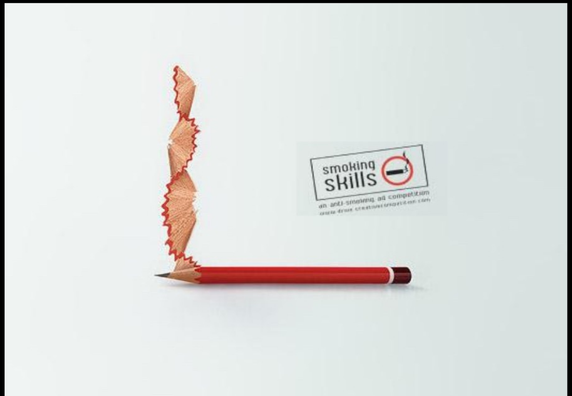
1. Models tend to rely more on primary object than secondary object
2. Fine-tuning helps
3. Performance drops greatly by increasing number of negatives (K)
4. Ample room for improvement

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021.

MetaCLUE Captioning and VQA

Results with PaLI-17B [1]

	PaLI-17B
Captioning	CIDEr: 1.076
VQA	Accuracy: 19.9%



Metaphor: Smoking cigarettes is as life-shortening as sharpening a pencil

Captioning Result

Prediction: *Smoking is as dangerous as burning a pencil.*

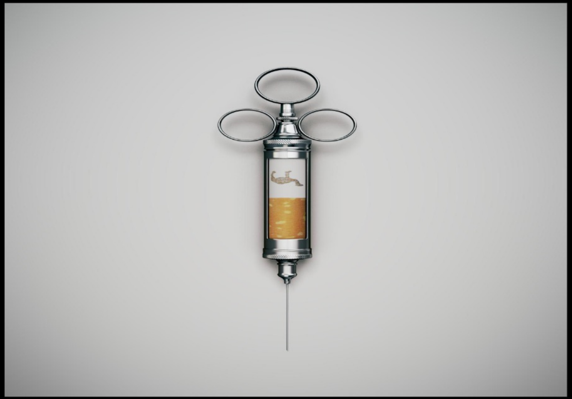
Visual Question Answering Result

Q1: What is a smoker's life compared to a sharpened pencil?

Pred: *Sharp* GT: Shortened

Q2: What is used as a visual metaphor for a sharpened pencil?

Pred: *Not smoking* GT: Smoker's life



Metaphor: Smoking is as deadly as injecting drugs

Captioning Result

Prediction: *Drinking and driving is as bad as injecting drugs.*

Visual Question Answering Result

Q1: What is a cigarette as compared to injecting drugs?

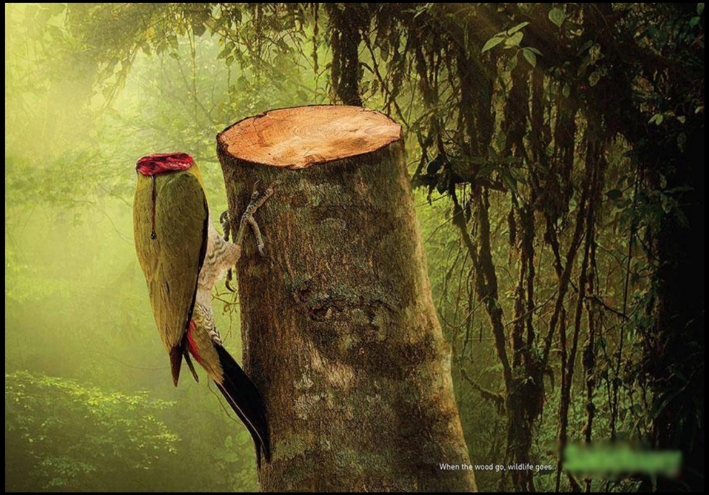
Pred: *Scary* GT: Deadly

Q2: What is used as a visual metaphor for injecting drugs?

Pred: *Drinking* GT: Cigarette

1. Xi Chen, Xiao Wang, Soravit Changpinyo, et al., PaLI: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794, 2022

Introducing MetaCLUE



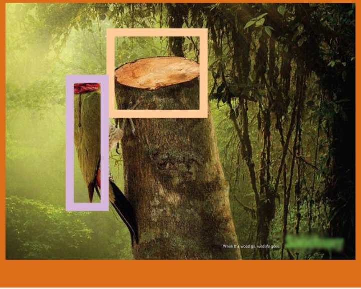
Metaphor: Killing trees is as harmful as killing wildlife

Classification

Is this a visual metaphor?
- YES
- NO

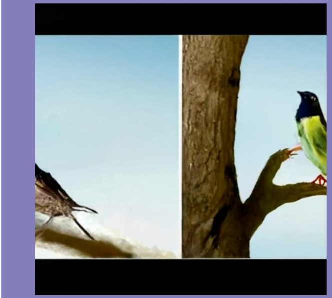
Localization

Detect image regions that invoke the concepts:
- Killing trees
- killing wildlife

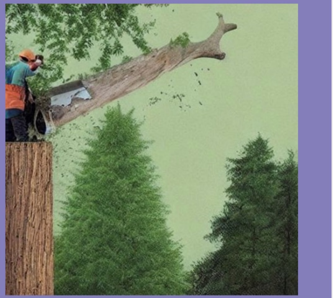


gEneration

Prompt: "An advertisement where killing trees is as harmful as killing wildlife."



Stable Diffusion



Imagen

Understanding

Retrieval

Pick the right one:

(a) Killing the forest is as deadly as killing the animals too.

(b) Birds is as much a part of our world as used cans.

Captioning

Sample predictions:

1. Deforestation is as damaging as killing wildlife.
2. Deforestation is as bad as ending the death penalty.

Visual Question Answering

Sample Questions:

Q. What is as harmful as killing wildlife?

Q. What is compared to killing wildlife?

MetaCLUE Generation

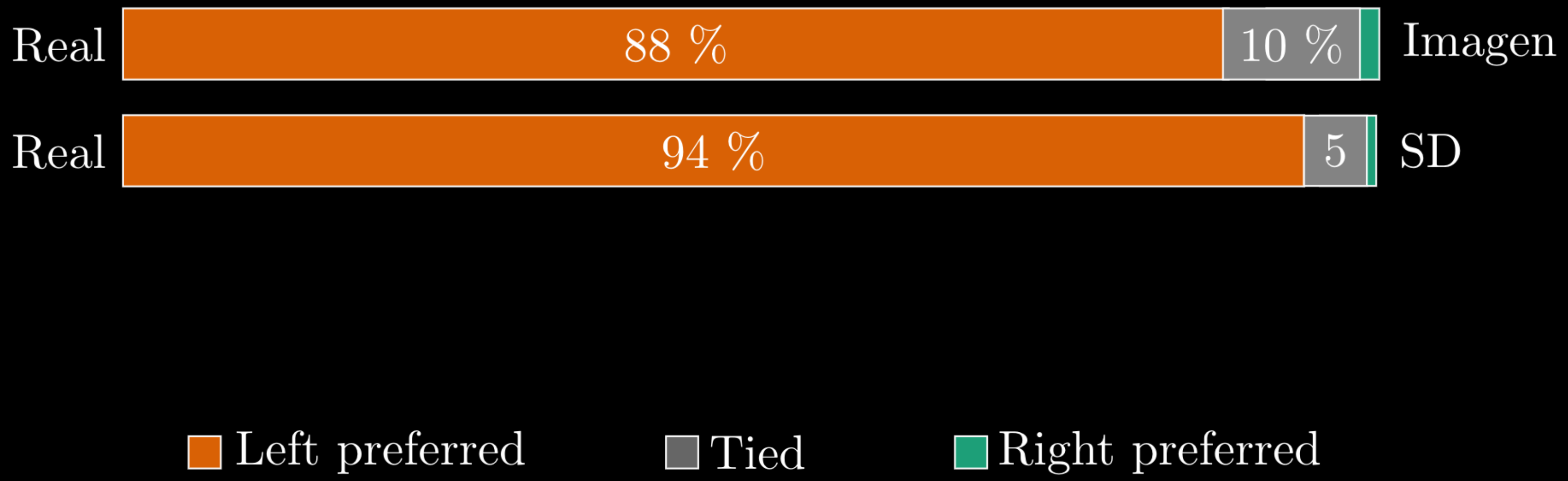
Can recent text-to-image models also work well in metaphorical image generation?

	FID	CLIP Similarity
Imagen [1]	153.1	32.1
Stable Diffusion [2]	161.6	30.8
Stable Diffusion - FT	154.3	32

1. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022.
2. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

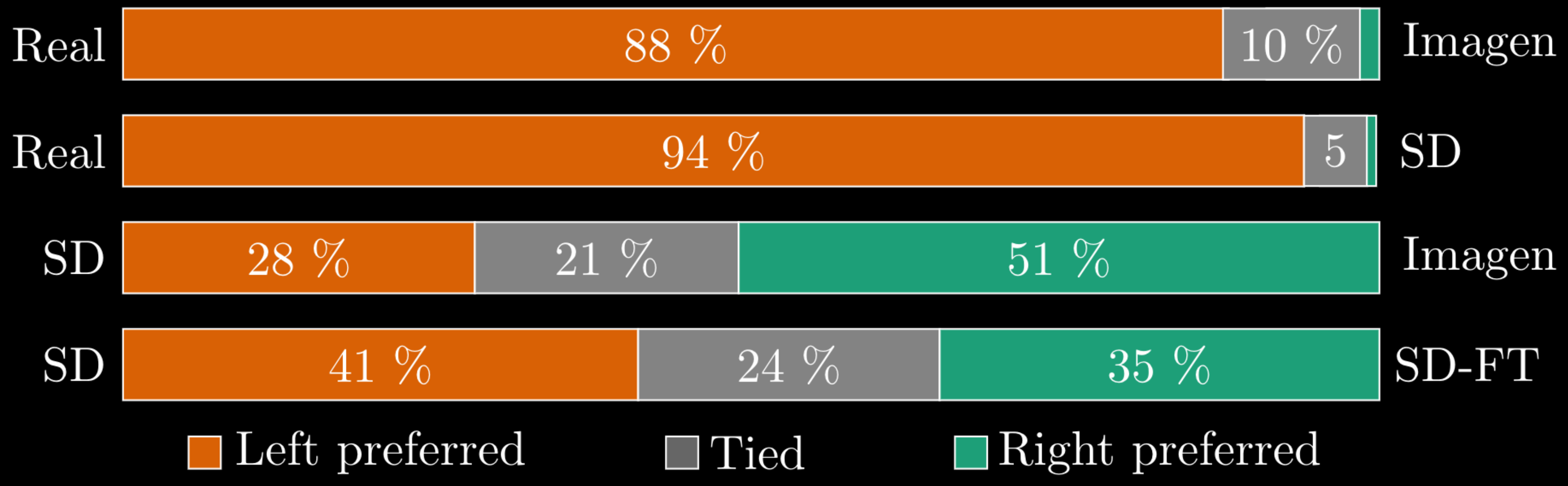
MetaCLUE Generation

User-studies



MetaCLUE Generation

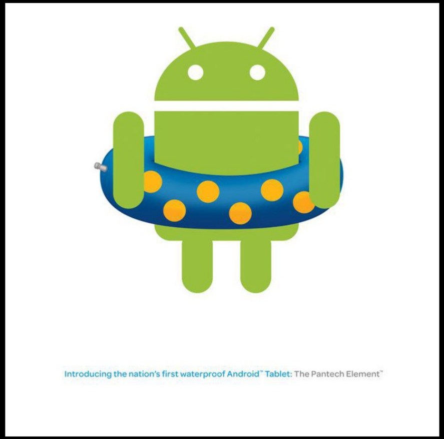
User-studies



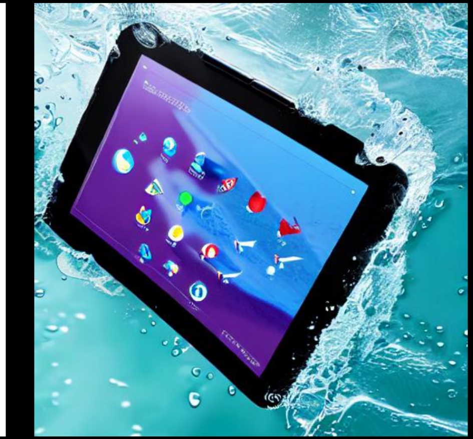
MetaCLUE Generation

Qualitative Results

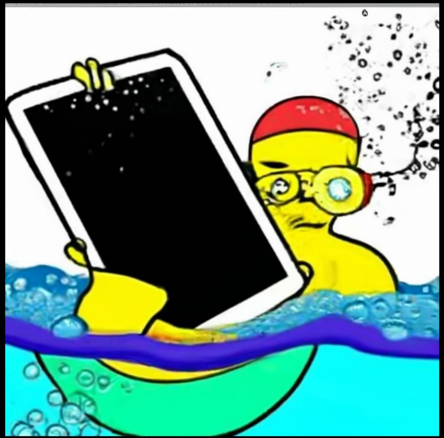
Metaphor: This android tablet is as waterproof as someone in a swimtube


Introducing the nation's first waterproof Android™ Tablet: The Pantech Element™

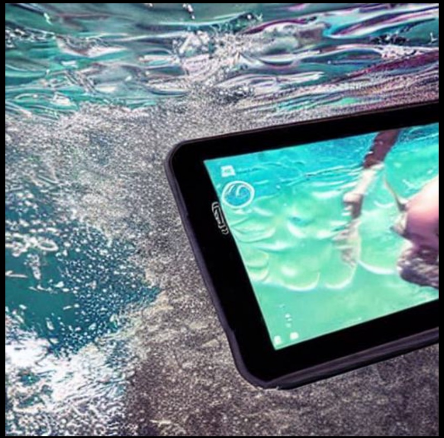
Real



Imagen



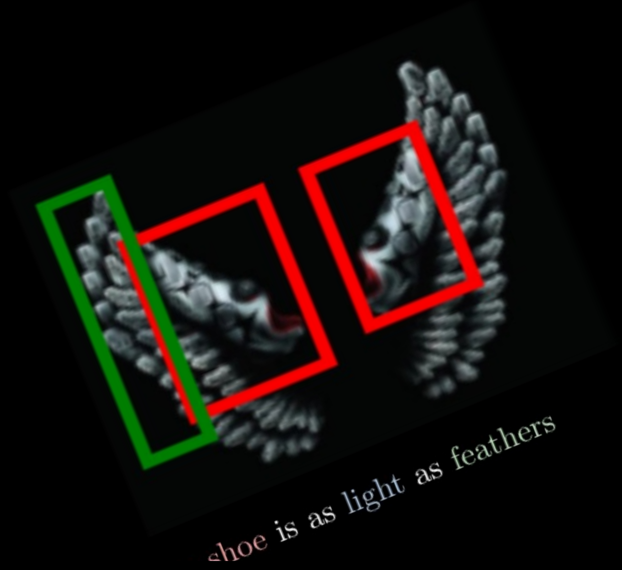
Stable Diffusion



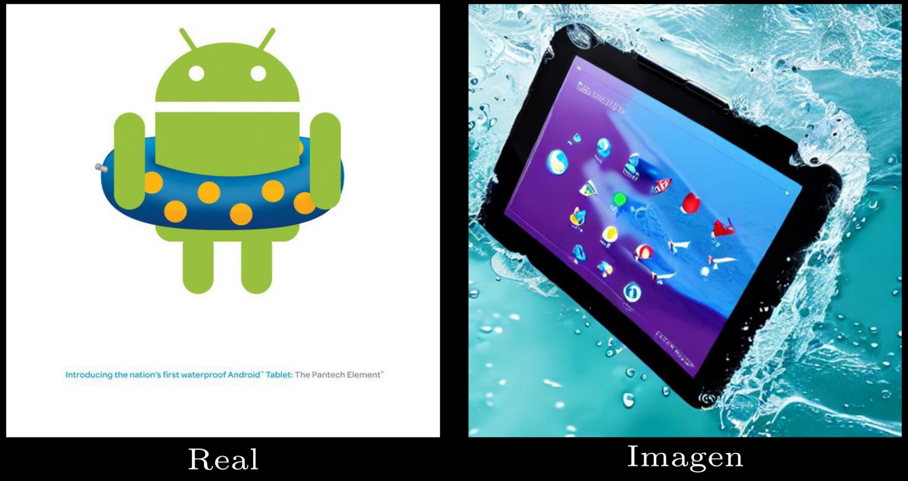
Stable Diffusion - FT

Summary – MetaCLUE

- Comprehensive and Measurable progress
- High quality and rich annotations
- Collection of tasks
    - Classification
    - Understanding
    - Localization
    - Generation
- Existing methods demonstrate poor results
- Concrete first step towards further AI research on Visual Metaphors



Metaphor: This android tablet is as waterproof as someone in a swimtube


Introducing the nation's first waterproof Android Tablet: The Pantech Element
Real
Imagen

MetaCLUE: Towards Comprehensive Visual Metaphors Research

(project page: https://metaclue.github.io)

Thank you

CVPR 2023

Poster Session: THU-PM-248

Google