

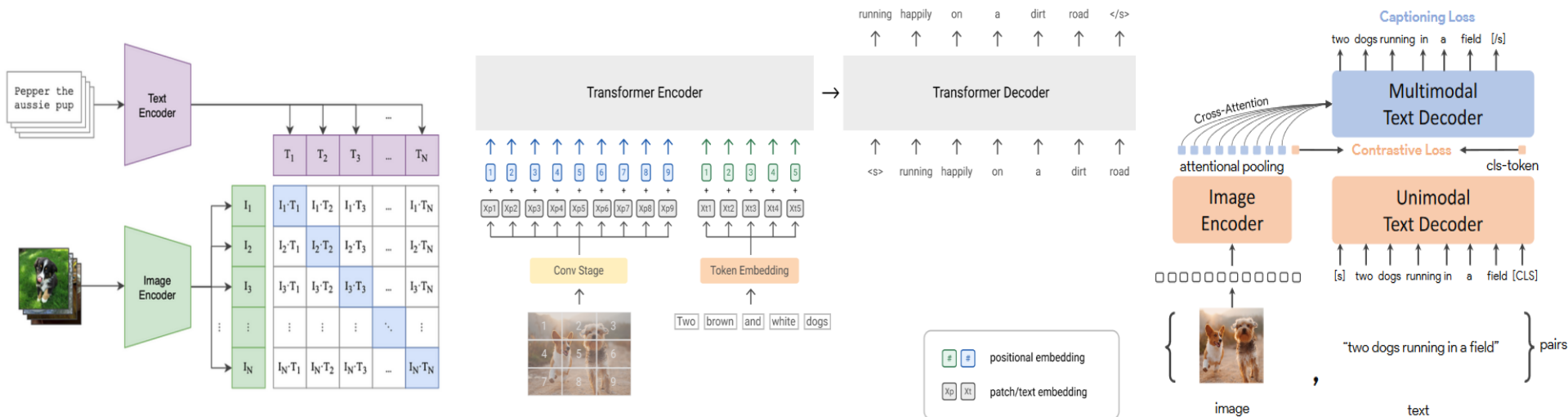
MaMMUT: A Simple Architecture for Joint Learning for MultiModal Tasks

Transactions on Machine Learning Research, August 2023
(12 Citations)

Weicheng Kuo, AJ Piergiovanni, Dahun Kim, Xiyang Luo, Ben Cain, Wei Li, Abhijit Ogale,
Luowei Zhou, Andrew Dai, Zhifeng Chen, Claire Cui, Anelia Angelova

Presented By:
Adrian Mauricio-Gonzalez, Cesar Hernandez, Ehtesamul Azim, Jatin Bharati

Background and Motivation



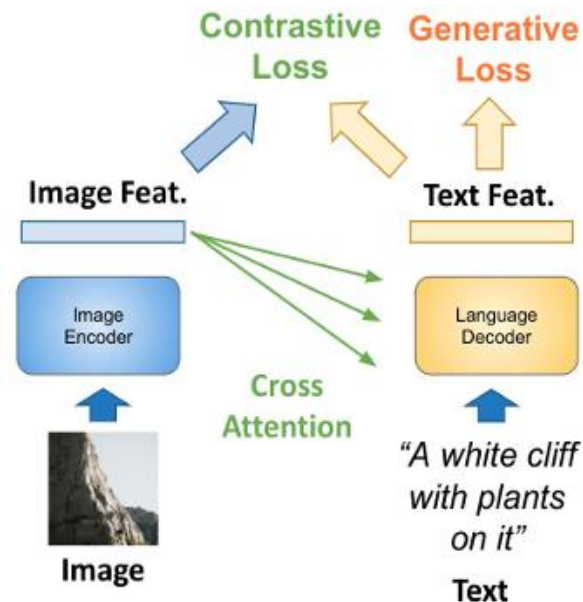
- Successful image understanding
- Cross-modal retrieval tasks

- Caption generation
- Visual Question Answering

- Needs multiple training stages or components
- Needs special recipe for video tasks

Proposed Solution

- Authors introduce **MaMMUT**
- Contains one visual encoder and one text decoder
- **Shared weights** in two-pass learning method.
- **Tested** against other vision-language models
 - Zero-shot Image Retrieval
 - Visual/Video Question Answering
 - Open-Vocabulary Object Detection



MaMMUT In Action

Image-Text and Text-Image Retrieval

Query

Top Retrievals



1. A girl with a guitar and a guy with an umbrella sitting in front of a gate on the sidewalk
2. A man and a woman are chatting while sitting next to a black gate
3. A man and a woman are sitting on the base of fence having a conversation

Query

Climbers with hiking boots and blue helmets ascend a snow covered mountain



Open-Vocabulary Detection



Novel classes:

- Cooking utensil
- Hotplate
- Saucepan
- Cooker
- Kitchen table

MaMMUT In Action(Cont'd)

VQA



Question: what is the number on the red bus ?

Predicted Answer:
15

Groundtruth:
15

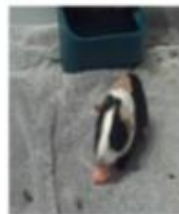
VideoQA



Question: What wants to jump in the water?

Predicted:
dog

Ground truth:
dog



Question: What is eating a carrot?

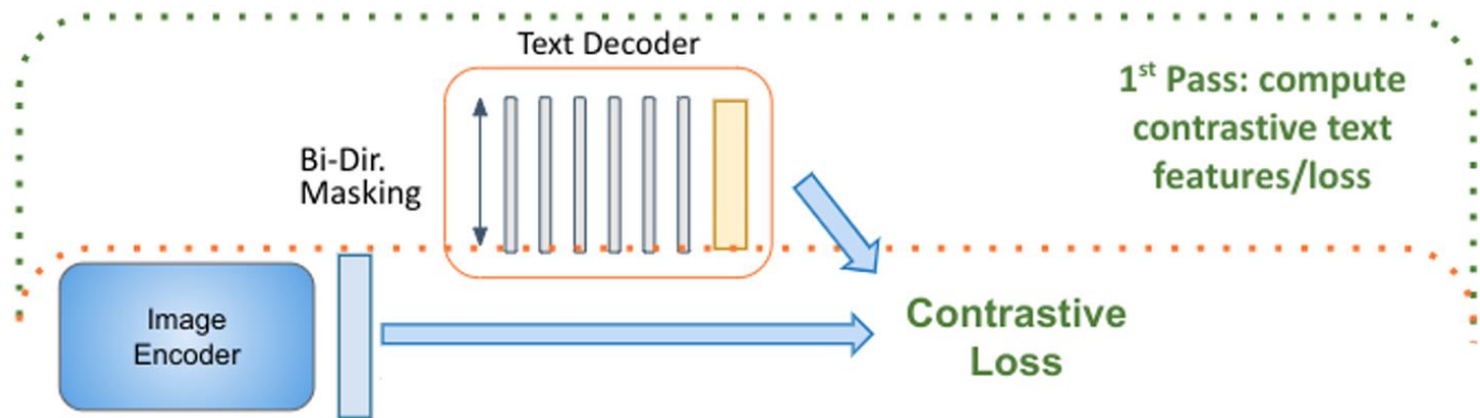
Predicted:
guinea

Ground truth:
gunia

Methodology

- ViT-based vision encoder and a single text decoder
 - M cross attention layers into N text decoder layers
 - No task-specific head required
- Uses a two-pass learning strategy approach to unify
 - contrastive learning
 - autoregressive captioning
 - localization awareness(with **cropped positional embedding**)
- Allows maximal weight sharing for generative and contrastive tasks
- Uses a noisy **web alt-text dataset**

1st Pass: Contrastive Loss



Focal Contrastive Loss

- Contrastive loss requires larger batch size
- Goal is to learn from the more challenging and informative examples
- Solution: **FOCAL LOSS!**
- Provides additional sensitivity to objects

$$p_i = \begin{cases} \sigma(v_i l_j / \tau) & \text{if } i = j \\ 1 - \sigma(v_i l_j / \tau) & \text{if } i \neq j \end{cases}$$

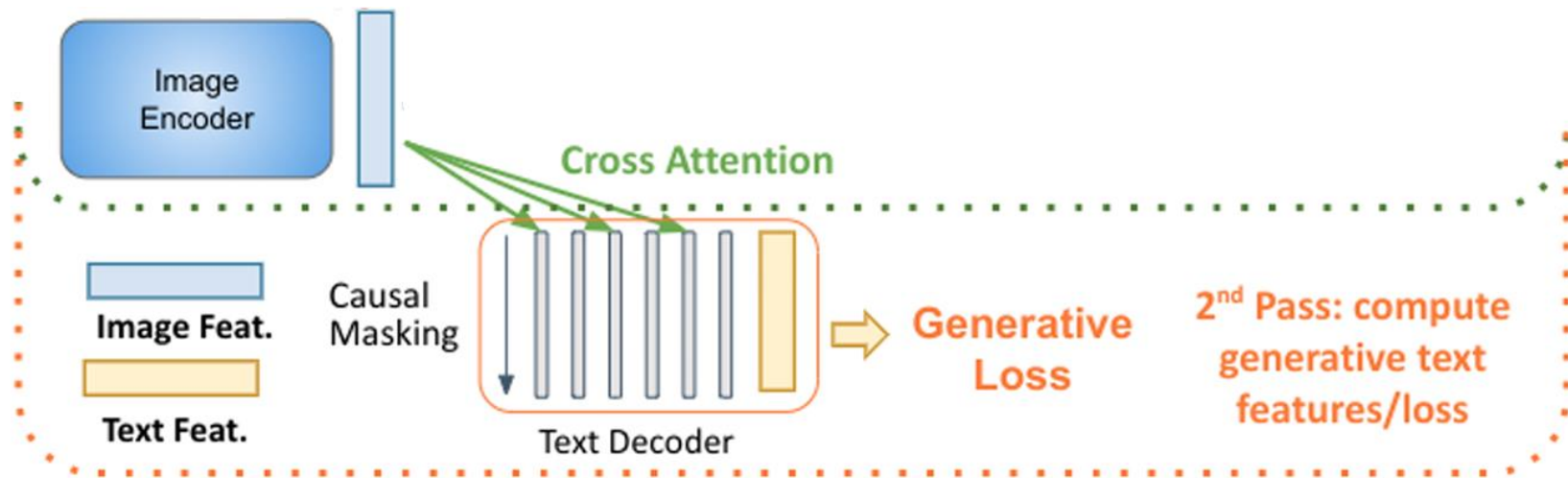
v_i and l_j represents normalized image and text embeddings

τ is the temperature hyperparameter

$$L_{\text{focal_contrastive}} = -\frac{1}{B} \sum_{i=1}^B \sum_{j=1}^B (1 - p_i)^\gamma \log(p_i),$$

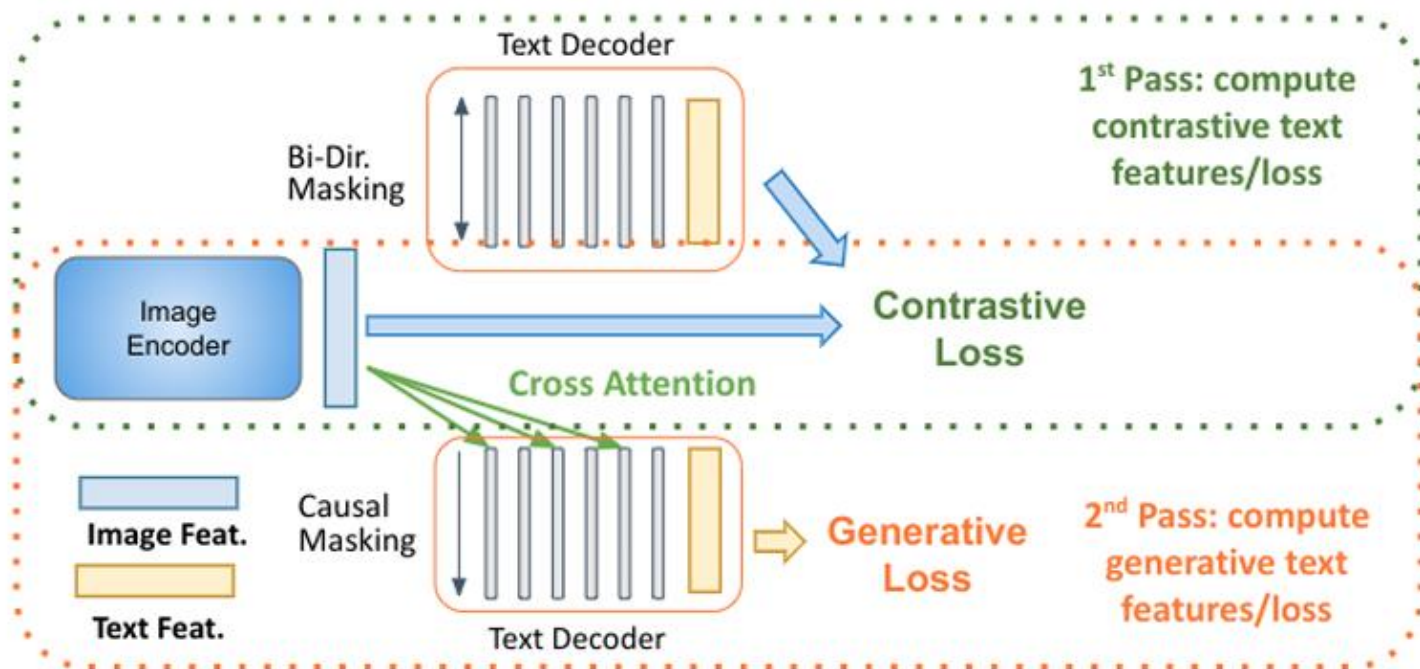
$$L_{\text{contrastive}} = L_{I2T} + L_{T2I}$$

2nd Pass: Generative Loss



$$L_{captioning} = - \sum_{t=1}^T \log P_{\theta}(y_t | y_{1,2,\dots,t-1}, x)$$

Architecture Overview

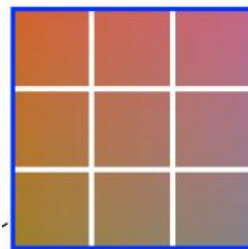


$$L_{total} = \lambda_{cap} L_{captioning} + \lambda_{focal} L_{focal_contrastive}$$

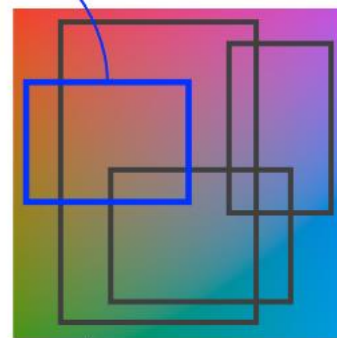
Learning Positional Embeddings for Localization Awareness

- VLMs use full-image positional embeddings
 - Works well for image classification
 - Does not work for detection at region level
- Solution: **Cropped Positional Embeddings!**

randomly crop
and resize



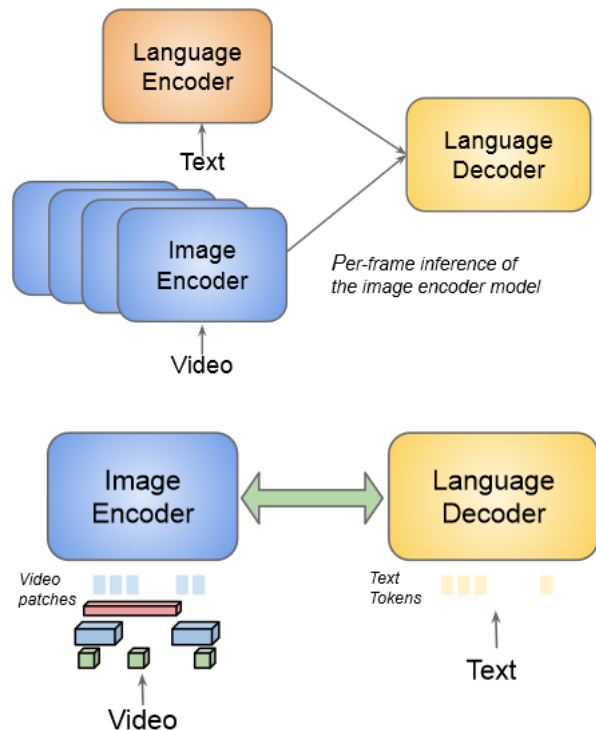
positional
embeddings



upsample

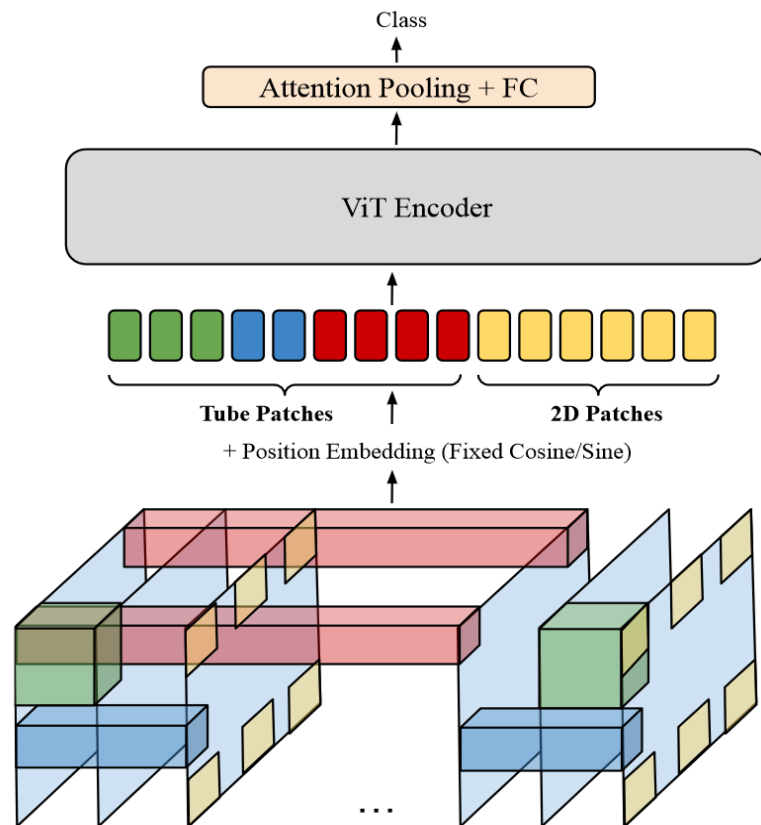
MaMMUT for Video Tasks

- Previous approaches include frame-by-frame processing
 - Just captures spatial information
- Motivation from TubeViT paper
 - 2D patches to process frames
 - 3D tubes to process multiple frames
 - **Sparse temporal stride** for 2D patches



MaMMUT for Video Tasks(Cont'd)

- Modifications to TubeViT approach
 - TubeViT uses fixed positional embeddings
 - MaMMUT combines learnt embeddings from encoder with weighted connections to fixed embeddings
- No additional pre-training on video data



MaMMUT for Video Tasks



Implementation Details

- ViT-Huge image encoder with **650M parameters**
- Transformer text decoder with **1B parameters**
 - Cross attention layers applied every two decoder layers
- AdamW optimizer with **0.01 weight decay value**
- Generative and contrastive loss weights set to **1.0**
- Pre-training images resized to **272x272**
 - Later cropped to **224x224**

Implementation Details

- Web alt-text dataset with **1.8B image-text pairs**
 - Used for contrastive and generative pre-training
- Fine-tuned with **Cropped Positional Embedding** for downstream detection.
- Used for ablation studies:
 - **ViT-Base** image encoder (**86M params**)
 - Text decoder (**128M params**)

Results: Zero-Shot Image-Text Retrieval

Method	image model size	MS COCO (5K test set)						Flickr30K (1K test set)					
		image-to-text			text-to-image			image-to-text			text-to-image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP (Radford et al., 2021)	302M	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN (Jia et al., 2021)	408M	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	99.7	75.7	93.8	96.8
FLAVA (Singh et al., 2022)	86M	42.7	76.8	-	38.4	67.5	-	67.7	94.0	-	65.2	89.4	-
FILIP (Yao et al., 2021)	302M	61.3	84.3	90.4	45.9	70.6	79.3	89.8	99.2	99.8	75.0	93.4	96.3
Florence (Yuan et al., 2021)	637M	64.7	85.9	-	47.2	71.4	-	90.9	99.1	-	76.7	93.6	-
CoCa-L (Yu et al., 2022)	303M	65.4	85.6	91.4	50.1	73.8	81.8	91.4	99.2	99.9	79.0	95.1	97.4
CoCa (Yu et al., 2022)	1B	66.3	86.2	91.8	51.2	74.2	82.0	92.5	99.5	99.9	80.4	95.7	97.7
MaMMUT (ours)	630M	70.7	89.1	93.7	54.1	76.8	84.2	94.9	99.5	99.9	82.5	96.0	98.0

Results: Visual Question-Answering

Method	Test-Dev	Test-Std
FLAVA (Singh et al., 2022)	72.8	-
METER (Dou et al., 2022)	77.7	77.6
Unified-IO (Lu et al., 2022)	77.9	-
OmniVL (Wang et al., 2022c)	78.3	78.4
Florence (Yuan et al., 2021)	80.2	80.4
SimVLM (Yu et al., 2022)	80.0	80.3
OFA (Wang et al., 2022d)	82.0	82.0
CoCa (Yu et al., 2022)	82.3	82.3
BEiT-3 (Yu et al., 2022)	84.2	84.0
ALBEF (Li et al., 2021)	75.8	76.0
AnswerMe (Piergiovanni et al., 2022a)	73.6	-
BLIP (Li et al., 2022a)	78.2	78.3
GIT (Wang et al., 2022b)	78.6	78.8
Flamingo-80B (Alayrac et al., 2022)	82.0	82.1
BLIP-2-7B (Li et al., 2023)	82.3	-
PaLI-3B (Chen et al., 2022)	79.3	-
PaLI-15B (Chen et al., 2022)	80.8	-
PaLI-17B (Chen et al., 2022)	84.3	84.3
MaMMUT (2B)	80.7	80.8

Overall	Yes/No	Number	Other
80.84	93.41	63.89	73.78

Results: Video QA and Video Captioning

Video QA Results

Method	MSRVTT-QA	MSVD-QA
Just Ask (Yang et al., 2021)	41.5	46.3
MERLOT (Zellers et al., 2021)	43.1	-
OmniVL (Wang et al., 2022c)	44.1	51.0
VindLU (Cheng et al., 2022)	44.6	-
Iterative Co-Tok (Piergiovanni et al., 2022b)	45.7	48.8
All-in-one (Wang et al., 2022a)	46.8	48.3
Video-Coca (Yan et al., 2022)	46.3	56.9
VIOLET (Fu et al., 2021)	43.9	47.9
VIOLETv2 (Fu* et al., 2023)	44.5	54.7
Dynamic Pretr. (Piergiovanni et al., 2023b)	45.1	47.1
GIT (Wang et al., 2022b)	43.2	56.8
GIT2 (Wang et al., 2022b)	45.6	58.2
InternVideo (Wang et al., 2022f)	47.1	55.5
Flamingo (Alayrac et al., 2022)	47.4	-
MaMMUT (ours)	49.5	60.2

Video Captioning Results

Method	MSRVTT	MSVD
ORG-TRL (Zhang et al., 2020)	50.9	95.2
OpenBook (Zhang et al., 2021b)	52.9	-
SWINBert (Lin et al., 2022)	53.8	120.6
VIOLETv2 (Fu* et al., 2023)	58.0	130.2
MV-GPT (Seo et al., 2022)	60.0	-
Vid2Seq (Yang et al., 2023)	64.6	146.2
Video-Coca (Yan et al., 2022)	73.2	-
GIT (Wang et al., 2022b)	73.9	180.2
GIT2 (Wang et al., 2022b)	75.9	185.2
MaMMUT (ours)	73.6	195.6

Results: Open-Vocabulary Detection

Method	APr	AP
DetPro-Cascade (Du et al., 2022)	20.0	27.0
Detic-CN2 (Zhou et al., 2022)	24.6	32.4
RegionCLIP (Zhong et al., 2022)	22.0	32.3
ViLD-Ensemble (Gu et al., 2022)	21.7	29.6
ViLD-Ensemble (Gu et al., 2022)	26.3	29.3
VL-PLM (Zhao et al., 2022)	17.2	27.0
Rasheed et al. (Rasheed et al., 2022)	21.1	25.9
OWL-ViT (Minderer et al., 2022)	23.3	35.3
OWL-ViT (Minderer et al., 2022)	25.6	34.7
MaMMUT (ours)	31.0	32.8

Ablation Studies

- **Cross-task Benefits**

Contrastive	Generative	MS COCO		Flickr30K		VQA
		I2T	T2I	I2T	T2I	Acc.
✓		54.8	38.2	82.6	67.1	63.5
	✓	-	-	-	-	69.9
✓	✓	54.3	38.7	80.6	67.5	71.7

- **Balancing Losses**

weights	MS COCO		Flickr30K		VQA
	I2T	T2I	I2T	T2I	Acc.
(2.0, 0.5)	56.71	40.77	82.52	67.77	70.08
(2.0, 1.0)	56.7	40.27	81.84	67.25	70.84
(1.0, 1.0)	56.25	39.73	81.74	67.50	71.48
(1.0, 2.0)	55.39	39.09	81.54	65.64	72.27
(0.5, 2.0)	52.05	37.32	78.32	62.73	71.79

Ablation Studies

- **Cross-attention Design**

# Cross-Att.	MS COCO		Flickr30K		VQA
	I2T	T2I	I2T	T2I	Acc.
1	55.3	39.6	81.7	67.2	68.7
2	56.6	40.1	81.9	67.6	70.8
4	55.7	39.9	82.2	67.3	71.5

- **Video Adaptation Experiments**

	MSRVTT-QA	MSVD-QA
MaMMUT- Full Model	42.1	45.8
No Gated Connection	41.8	45.5
No Fixed Embeddings	41.5	45.1
No Tubes	40.3	42.6

Ablation Studies

- **Projections and attention pooling**

Att. Pool/ Proj	image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10
N/N	53.32	80.57	86.91	39.86	67.7	77.46
Y/N	52.44	78.71	86.33	39.3	66.62	76.91
N/Y	51.37	78.81	86.13	39.49	66.97	77.05
Y/Y	50.78	76.95	85.74	38.59	66.54	76.07

Total Train Computage

- MaMMUT performance **from scratch**:
 - **3.4x** cheaper than PaLI (relies on pretrained image encoder)
 - **5.5x** than CoCa
 - **10.3x** than Flamingo (relies on pretrained image encoder)
 - **41.2x** than GIT-2

Limitations

- MaMMUT relies on web alt-text data for pre-training
 - The model is subject to text generative risks from biased data
 - Further investigation is needed
- The model relies on a single text decoder for joint learning
 - Introduces conflict through trade-offs
 - Weights assigned to contrastive and generative loss
 - Number of cross-attention layers
 - Performance on image-to-text retrieval

Conclusion

- The MaMMUT model consists of a vision encoder and text decoder
- Two-pass learning allows the model to train for retrieval and text generation using shared weights
- The model is capable of handling a diverse set of tasks
 - Image-text / Text-image retrieval
 - Open vocabulary object detection
 - VQA
 - VideoQA
 - Video Captioning