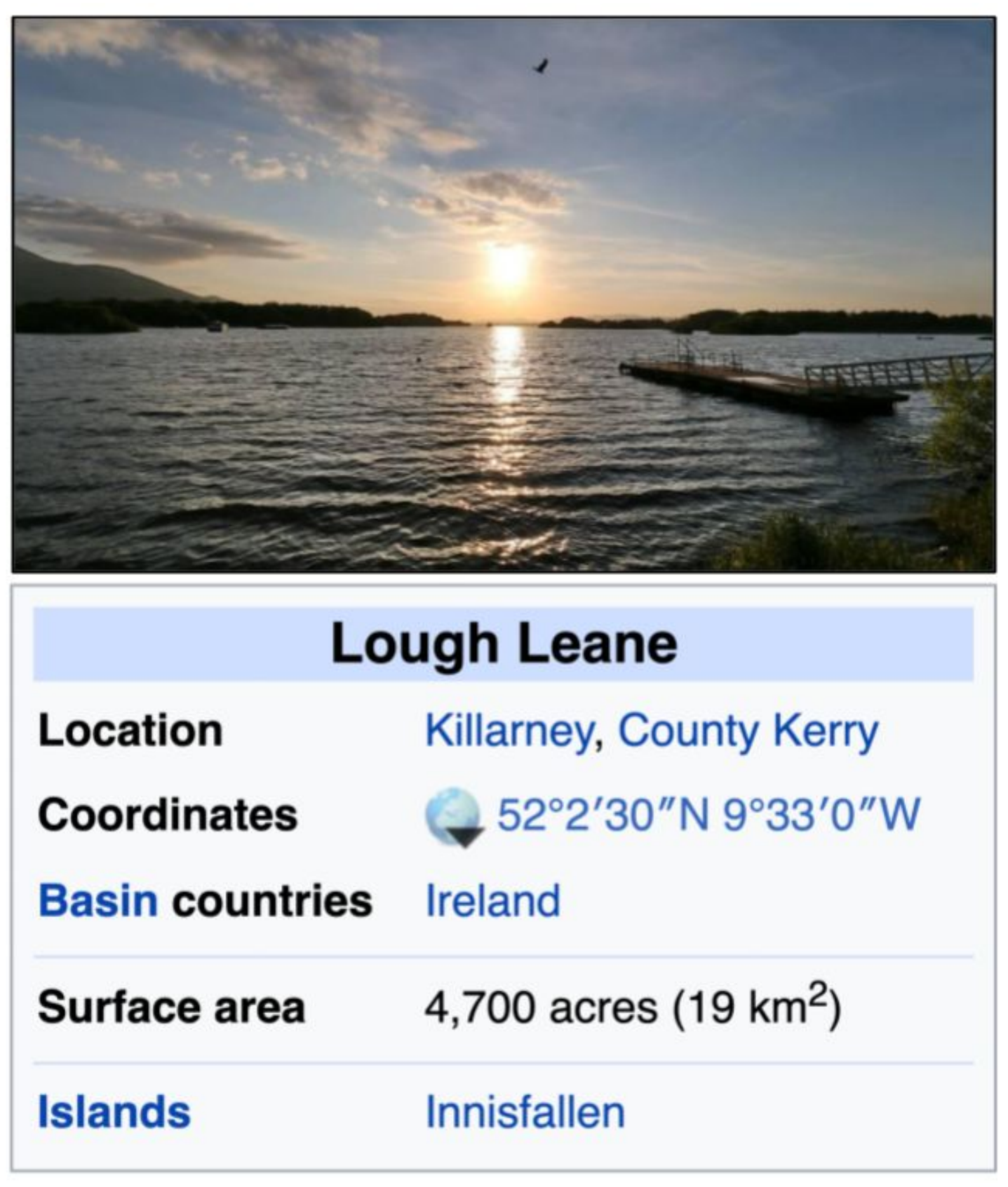
VisToT: Vision-Augmented Table-to-Text Generation

Prajwal Gatti[1], Anand Mishra[1], Manish Gupta[2], Mithun Das Gupta[2]

[1]Indian Institute of Technology Jodhpur, [2]Microsoft

Looking at the Table-to-Text problem from a multimodal lens

VisToT Task



Lough Leane	
Location	Killarney, County Kerry
Coordinates	52°2′30″N 9°33′0″W
Basin countries	Ireland
Surface area	4,700 acres (19 km^2)
Islands	Innisfallen

⇨ "Lough Leane is a large lake in Killarney, County Kerry, Ireland."

Tables contain a structured list of facts, images are a rich source of unstructured visual information.

VisToT proposes use of information from both modalities to generate a meaningful text description.

Given a table **T** describing an entity **E** and an associated image **I**, the **goal** is to generate a sentence description **S** such that it accurately describes E using the source context of T and I.

VisToT can be applicable in domains such as tourism, healthcare and e-commerce.

WikiLandmarks Dataset

Name	Amitabha Drukpa
Country	Nepal
Location	Kathmandu
Dedicated To	Amitabha

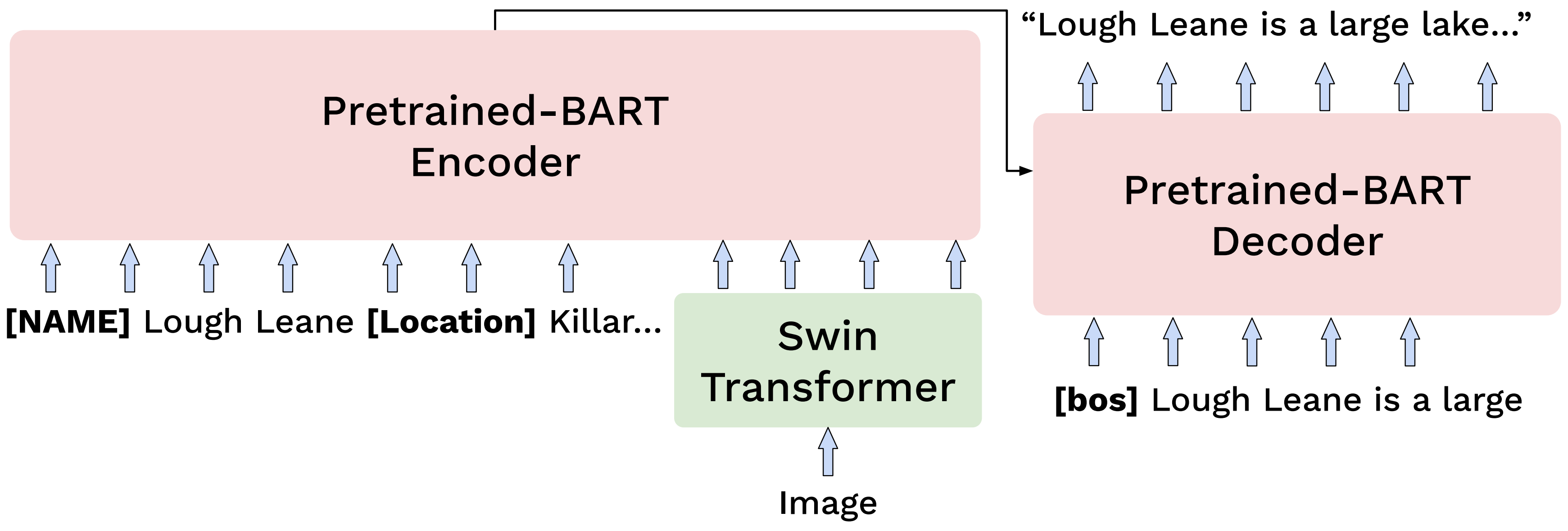
*"Amitabha Monastery is a **Tibetan Buddhist Monastery** in Nepal"*

Name	Michigan Stadium
Location	1201 South Main Street Ann Arbor, Michigan
Owner	University of Michigan
Nickname	The Big House

*"Michigan Stadium, nicknamed The Big House, is **the football stadium** for the University of Michigan in Ann Arbor, Michigan"*

- Tables and Images for **73K+ world landmarks**.
- Each sample contains a *table*, *image*, and a *text summary.*
- Table and text summaries are obtained from Wikipedia.
- Images contain visually inferable facts –
  - **Type of landmark** (e.g., Church, Castle)
  - **Architecture** (e.g., Ancient Roman, Mughal),
  - **Composition** (e.g., White Marble, Bronze), and many more.

VT3: Vision-Tabular Data to Text Transformer


"Lough Leane is a large lake..."
Pretrained-BART Encoder
Pretrained-BART Decoder
[NAME] Lough Leane [Location] Killar...
Swin Transformer
Image
[bos] Lough Leane is a large

We also propose three pre-training objectives:

a. Image-Table Matching (ITabM),
b. Masked Value Modeling (MVM), and
c. Image Captioning (IC).

Experiments

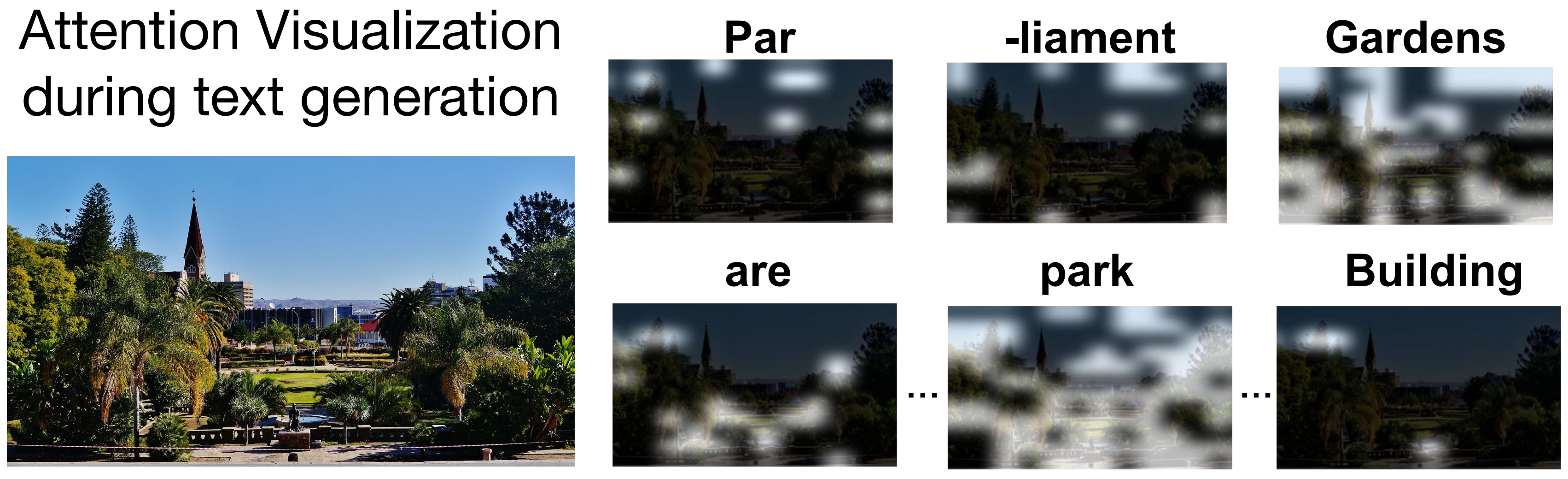
Method	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT
Image captioning-based						
PureT	6.4	26.1	33.2	12.8	31.1	0.40
Table-to-Text						
Pointer-Generator	17.8	39.2	51.6	31.7	49.2	0.50
BERT-to-BERT	22.1	43.9	55.3	35.6	53.1	0.50
T5	25.8	48.1	58.8	38.8	57.0	0.54
PlanGen	8.6	20.6	32.5	20.2	31.9	0.49
Visual-Tabular Data-to-text						
LSTM+ResNet50	6.5	19.8	31.0	19.1	30.3	0.39
VisualBERT+BERT	26.1	49.0	60.4	39.2	58.8	0.54
VT3	**30.2**	**53.5**	**62.9**	**43.4**	**60.8**	**0.56**

Table 1: Performance comparison on the WikiLandmarks test set.

Metric	FRCNN	CLIP-ViT	ViT	Swin
BLEU	27.4	28.2	29.6	**30.2**
METEOR	50.8	51.6	52.9	**53.5**
ROUGE-1	59.9	60.3	61.7	**62.9**
ROUGE-2	42.3	43.0	42.7	**43.4**
ROUGE-L	58.2	58.9	59.5	**60.8**

Table 2: Ablation for VT3 model with different Visual Encoders.

Attention Visualization during text generation



Summary

We propose the task of VisToT, a vision-augmented extension to the table-to-text problem.

We introduce WikiLandmarks dataset to study VisToT task.

We present VT3, a multimodal transformer for solving VisToT.

EMNLP 2022

We thank Microsoft for supporting this work through Microsoft Academic Partnership Grant (MAPG) 2021

Paper, code, and dataset available here: https://vl2g.github.io/projects/vistot/