Satellite Branch

Location Branch

Street View Branch

Text Generation

Range Query

Retrieve

Text Generation

Large Multimodal Model

(lat, lon)

(lat, lon)

...

(lat, lon)

Large Multimodal Model

Satellite Text

Satellite Imagery

Street View Gallery

Street View Text

Tokenize

Text Refinement

Patchify

Patchify

Text Calibration

Tokenize

The region contains building ...

Skyscrapers stand in the city ...

Linear Projection

[SOS]

[CLS]

Linear Projection

GeoCLIP f_ϕ

Linear Projection

[CLS]

[SOS]

Linear Projection

L ... 3 2 1 0*

0* 1 2 3 ... J

Location Encoder

J ... 3 2 1 0*

0* 1 2 3 ... L

ST Text Encoder f_β

ST Image Encoder f_α

SV Image Encoder f_λ

SV Text Encoder f_γ

Aggregation

Aggregation

Aggregation

[EOS]

[EOS]

Fusion

Text Tokens

	T_1^{st}	T_2^{st}	...	T_N^{st}
I_1^{st}				
I_2^{st}				
...				
I_N^{st}				

	I_1^{st}	I_2^{st}	...	I_N^{st}
T_1^{st}				
T_2^{st}				
...				
T_N^{st}				

Image-to-Text Global Contrastive

Text-to-Image Global Contrastive

I_1^{st}	I_2^{st}	I_3^{st}	I_4^{st}

| T_1^{st} | T_2^{st} | T_3^{st} | T_4^{st} |

Batch Size N

❄️ **Frozen**
🔥 **Tuning**

Image Tokens

Token-wise Similarity

1.max
2.mean
2.mean
1.max

	T_1^{sv}	T_2^{sv}	...	T_N^{sv}
I_1^{sv}				
I_2^{sv}				
...				
I_N^{sv}				

	I_1^{sv}	I_2^{sv}	...	I_N^{sv}
T_1^{sv}				
T_2^{sv}				
...				
T_N^{sv}				

Image-to-Text Local Contrastive

Text-to-Image Local Contrastive

Stage1: Pretraining

Stage2: Fine-Tuning

Range Query

Retrieve

(lat, lon)

...

GeoCLIP f_ϕ

Location Encoder

ST Image Encoder f_α

SV Image Encoder f_λ

Predictor

Urban Insight

Population | GDP | Carbon | Night Light | House Price | POI