KOREA UNIVERSITY 1905

WOORI BANK

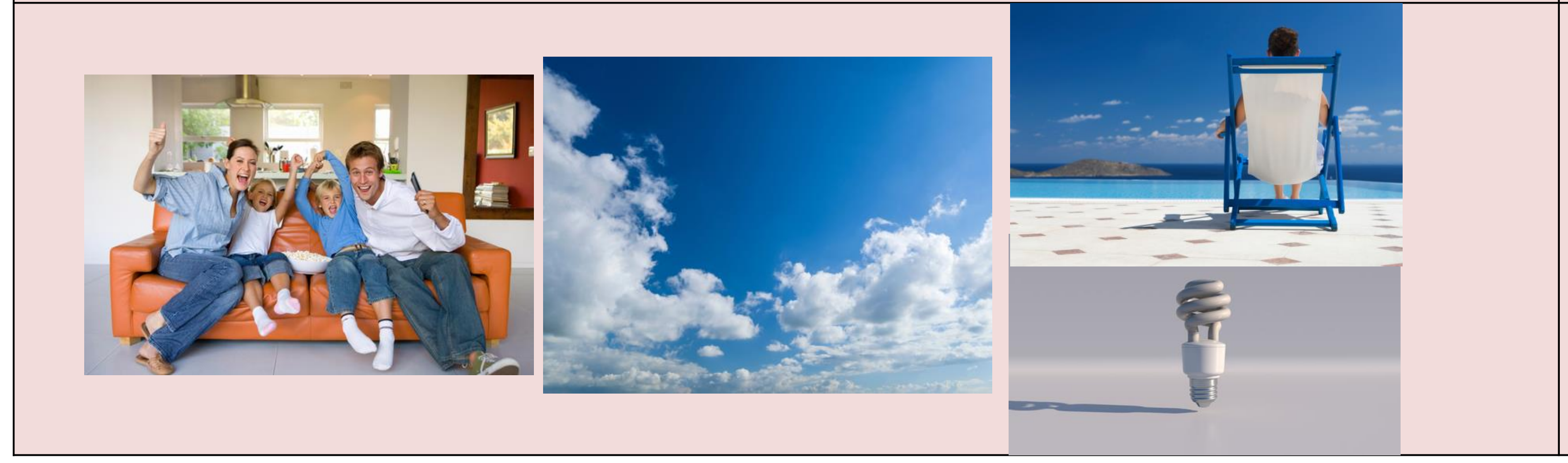
Levels: Generating Image Captions Using Levels of Hints Based on Vector Similarity

Solgil Oh

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Introduction of the NICE Dataset

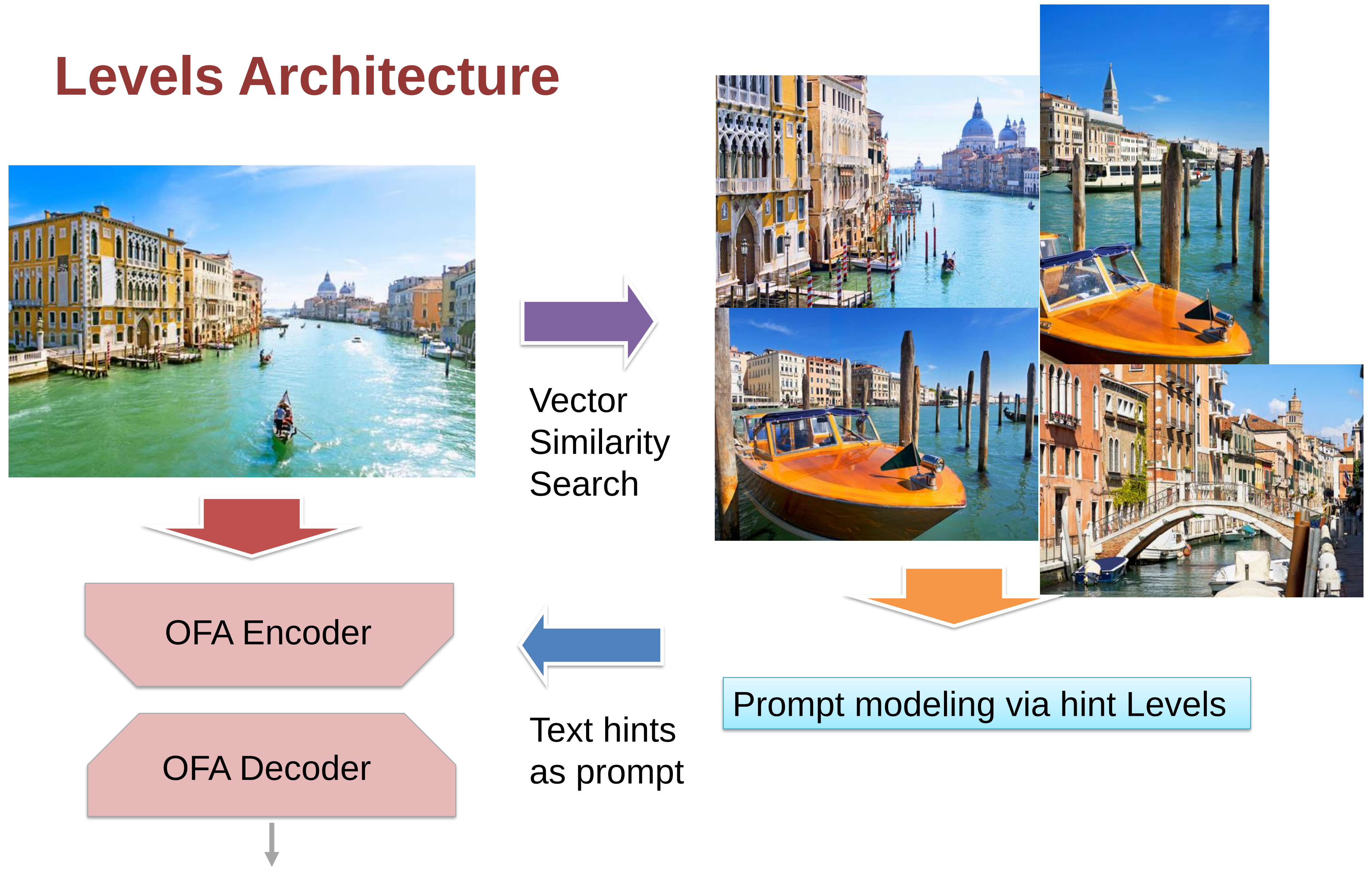
- 5000 validation photos with ground-truth caption
- 21377 test photos for challenge
- A caption with unique tone that describe each photo

Photo Examples	Common word
	Vertical shot of
	Germany
	No shot style No location

Caption Characteristics

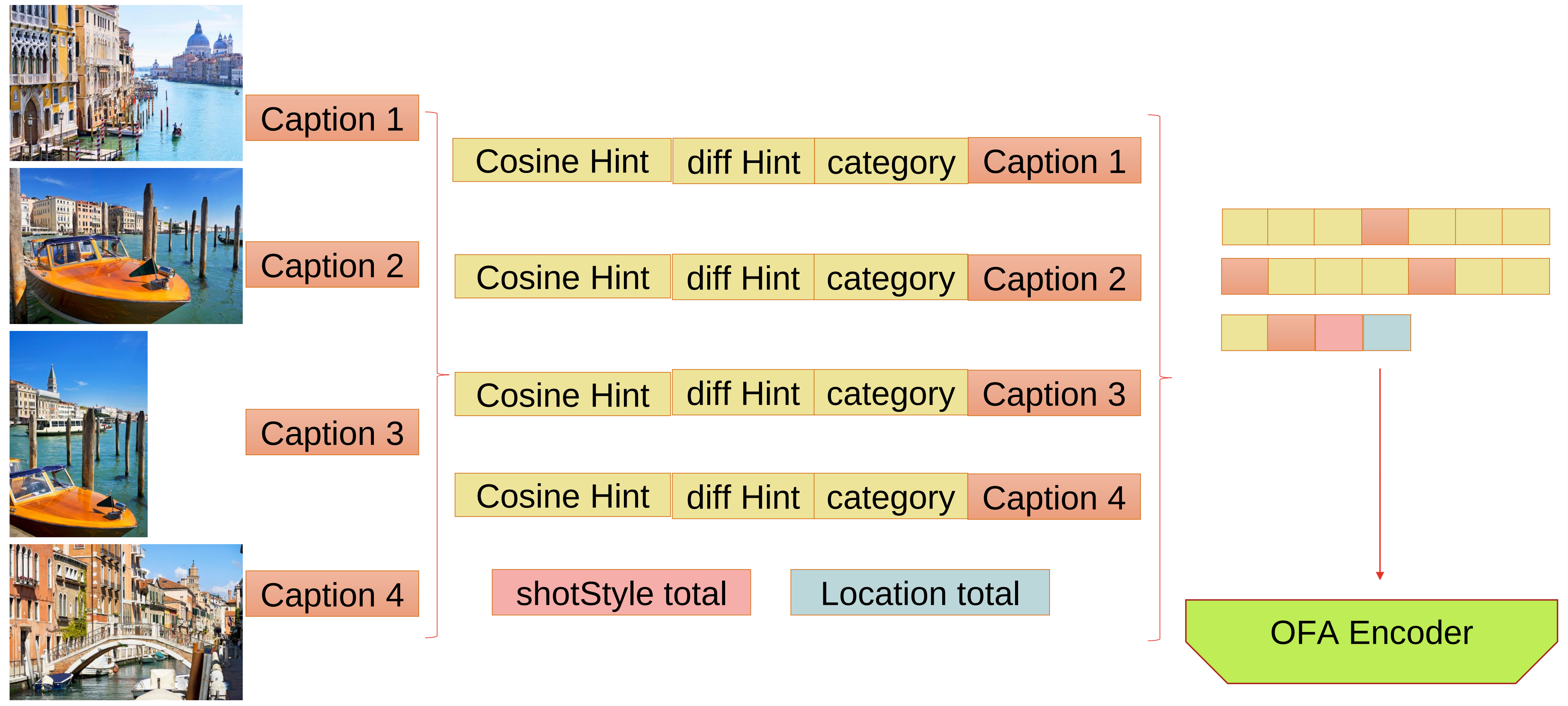
Characteristic caption	Examples
Photo style (shot style)	View of, horizontal view of, vertical shot of, detail view of, cutout of, portrait of etc.
Locations	Balloon Festival Albuquerque New Mexico USA, Cathedral of Palma de Mallorca at night Mallorca Spain, Carinthia Austria etc.

Levels Architecture



```
Gondoliers paddling tourists in gondola among architectural
buildings in the sunny Grand Canal in Venice Italy
```

Input Prompt Modeling Architecture



Hint Level Token Threshold Details

Hint Levels	Degree of hint effect	Criterion
[cosHint lv4]	Strong hints for nearly identical photos	CS*>0.40
[cosHint lv3]	Same topic but expected different captions	CS>0.32
[cosHint lv2]	Similar photo but different caption	CS>0.29
[cosHint lv1]	Irrelevant photo	CS≤0.29
[diffHint lv3]	ID difference value is very small	IDD*<100
[diffHint lv2]	ID difference value is small	IDD<10000
[diffHint lv1]	ID difference value is large	IDD≥10000

CS: Cosine Similarity, IDD: ID difference

Results

- CIDEr (287.69) Track2 (2nd) Total (4th)
- Five-checkpoints ensemble to control convergence rate

```
View of a
colorful hot
air balloon
against blue
sky Balloon
Festival
Albuquerque
New Mexico USA
```

```
Young man
sitting on a
railing and
using a digital
tablet with a
stop sign in
the background
```

Reflection on Results

- Difficulty adjusting convergence rate of text prompts/images
- Further research exploration based on the vector database

Model and code are publicly available: