A. FUSECAP Fused Captions

In this section, we present supplementary details about the newly proposed fused dataset presented in Section 3. Appendix A.2 presents examples of this enriched data, adding to the ones shown in Figure 1.

A.1. Visual Experts Implementation

We utilize the Visual Experts discussed in Section 3.1, in the following manner,

- **Object detection** We consider objects as valid detections if they surpass a pre-determined detection confidence threshold of 0.7.
- **Attribute Detection** We incorporate attributes to each valid predicted object if the attribute confidence surpasses a 0.2 threshold.
- **OCR**: We use CRAFT and Parseq with default inference parameters [5, 8]. The text recognized is attributed to the object that has the smallest bounding box encompassing it.

A.2. Training Set Generation for LLM Fuser with ChatGPT.

As outlined in Section 3.2, we harness the zero-shot capabilities of ChatGPT to generate a compact dataset encompassing 20,000 examples. This data is subsequently used to fine-tune an open-source Large Language Model (LLM). The prompt provided to ChatGPT is as follows:

"A caption of an image is given: *original caption*.
The following objects are detected in the image from left to right:
A a_1^1, ..., a_1^{k-1} and a_1^k o_1 [with the following text: t_1].
$\vdots$
A a_n^1, ..., $a_n^{k_n-1}$ and $a_n^{k_n}$ o_n [with the following text: t_n].
Write a comprehensive and concise caption of the scene using the objects detected."

We denote $\{o_i\}_{i=1}^N$ as the set of objects detected, $\{a_i^j\}_{i=1}^N$ as the attributes related to each object, and $\{t_i\}_{i=1}^N$ as texts related to each object.

B. Training Settings

B.1. LLM Fuser

Our LLM Fuser is a fine-tuned FlanT5-XL [19] model. We used the huggingface library to fine-tune it on a single NVIDIA A40 GPU. We trained the model for $4,000$ optimization steps, with batch size of 32, a learning rate of $5 \cdot 10^{-5}$, and a linear scheduling strategy. We limit the source and target length to 100 and 200 tokens, respectively.

B.2. Caption Generator

Figure 5 offers further examples, supplementing those found in Figure 3 and providing additional instances of the model's outputs. These examples further emphasize the ability of the captioning model to generate semantically rich captions.

Training. The pre-training and subsequent fine-tuning of the captioning model, described in Section 5, was performed on eight NVIDIA A100 GPUs. For setting the pre-training hyperparameters, we followed the approach outlined in the original BLIP implementation [38], except for maximum caption length, batch size, and initial learning rate. We deviated from the original model's batch size to accommodate the increased token length used in our implementation. To maintain stability during pre-training with a smaller batch size, we reduced the initial learning rate. The batch size used for pre-training was 400 (50 per GPU). The initial learning rate used was $6 \cdot 10^{-5}$.

Maximum Caption Length. The original BLIP model imposes a training and inference limit of 30 tokens per caption. With our enrichment process, however, caption lengths tend to be longer and beyond this original threshold. Accordingly, we have increased the maximum caption length to 60 tokens. This limit is maintained through both the pre-training and fine-tuning stages of the caption generator, as well as during the actual generation of captions.

Original: A man preparing to catch a frisbee in front of some houses.

Ours: A man in white and blue shorts prepares to catch a white frisbee in front of a stone wall and a black metal fence, with a brown and red roof in the background, under a blue sky.

Original: The Bug Club | Membership | Romney Hythe & Dymchurch Railway

Ours: A black and orange train, belonging to The Bug Club, sits on the tracks of Romney Hythe & Dymchurch Railway.

Original: A hard black case sits outside a door.

Ours: A hard black case sits outside a wood door with the text "DESPARPAJO" and a pink handle.

Original: rainbow lorikeet bird in tree

Ours: A rainbow lorikeet bird perches on a tree amidst green leaves and blue sky, with its striking red body, blue head, and vibrant blue and orange beak.

Original: Hard times teach us valuable lessons. Handwriting on a napkin with a cup of coffee stock photos

Ours: A cup of coffee sits on a red saucer, accompanied by a white napkin and a metal pen. The scene suggests a reminder that hard times teach valuable lessons, as captured in handwritten text.

Original: happy kids sitting on the grass.

Ours: A group of happy blond girls sitting on the grass, wearing red, yellow, and blue shirts, with one girl having a small nose and another having a closed eye.

Original: A man playing tennis in a tennis court

Ours: A man in a white shirt and red shorts plays tennis on a court with a blue racket and white socks. He wears dark sunglasses and stands in front of a chain link fence and a green tree.

Original: A sign in my bed room.

Ours: A black letter hangs on a white wall with the text "Shack" in my bedroom.

Figure 4. Examples of the proposed FUSECAP enriched-captions dataset.

GIT: a man and a woman posing for a picture.

OFA: a man and a woman sitting on a motorcycle

Prismer: A man and a woman sitting on a motorcycle.

BLIP2: a man and woman sitting on a motorcycle outside

Ours: a smiling man and woman pose for a picture in front of a red motorcycle, with the man wearing a blue shirt and black glasses, and the woman wearing a white shirt

GIT: a cow laying in a field with a tree in the background.

OFA: a white cow sitting in a grassy field

Prismer: A white cow laying on a lush green field.

BLIP2: a cow is laying in the grass on a sunny day

Ours: a white cow rests in a lush green field under a clear blue sky, with a large green tree in the background and a dark shadow cast on the ground

GIT: a man riding a small motorcycle in a parking lot.

OFA: a man riding a motorcycle in a parking lot

Prismer: A man riding a motorcycle in a parking lot.

BLIP2: a man riding a motorcycle in a parking lot with tents

Ours: a man wearing a white shirt and blue jeans rides a motorbike in a parking lot surrounded by white and yellow tents, with a white line marking the edge of the parking

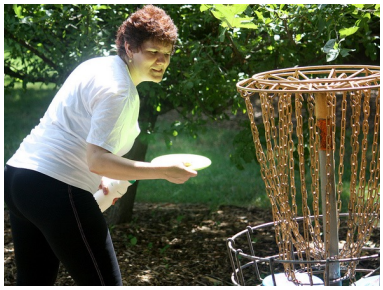
GIT: a man riding a horse with a dog in the background.

OFA: a man riding on the back of a white horse

Prismer: A man riding a horse next to a small dog.

BLIP2: a man riding a horse with a dog in the field

Ours: a man wearing a red hat and blue jeans rides a white horse with a long tail, while a small white dog follows closely behind



GIT: a woman is holding a frisbee in her hand.

OFA: a woman throwing a frisbee into a metal basket

Prismer: woman throwing a frisbee into a metal cage.

BLIP2: a woman is holding a frisbee in front of a metal cage

Ours: a woman with brown hair wearing a white shirt and black pants throws a white frisbee into a metal basket on a green grassy field

GIT: an elephant is riding in the back of a truck.

OFA: an elephant is riding in the back of a truck

Prismer: An elephant is riding in the back of a truck.

BLIP2: a large elephant is riding in the back of a truck

Ours: a brown elephant stands on the back of a truck on a gray road surrounded by green trees and grass under a cloudy sky, with a white line marking the edge of the road

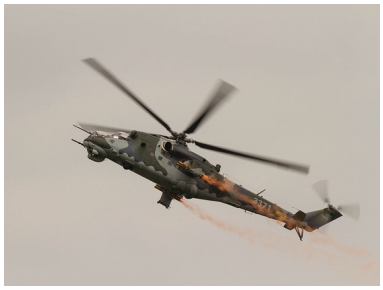
GIT: a bird flying over a city with mountains in the background.

OFA: a bird flies over the allegheny mountains

Prismer: A bird flying over a lush green hillside.

BLIP2: a bird flying over a green valley with mountains in the background

Ours: a black bird soars over a lush green valley, with a brown building in the foreground and a white and blue sky overhead



GIT: a helicopter is flying through the air with smoke coming out of it.

OFA: a helicopter flying in the sky

Prismer: A helicopter flying through a gray sky with propellers.

BLIP2: a helicopter is flying through the air with smoke coming out of the tail

Ours: a military helicopter with black blades and propellers flies through a gray sky, emitting orange and red smoke

GIT: a bunch of carrots and other vegetables on a counter.

OFA: a pile of vegetables sitting next to a white blender

Prismer: A close up of a blender and vegetables on a table.

BLIP2: a pile of vegetables sitting on a counter next to a juicer

Ours: a colorful assortment of vegetables, including carrots, celery, and lettuce, are arranged on a counter next to a white blender with a green handle

GIT: a woman holding an umbrella outside a shop.

OFA: a woman holding an umbrella in front of a shop

Prismer: A woman holding a white umbrella over her head.

BLIP2: a woman holding an umbrella outside a store on a cobblestone street

Ours: a woman stands in front of a store, holding a white umbrella she wears blue jeans, a white scarf, and blue boots in the background, there is a wicker basket

Figure 5. Comparative illustration of captions generated by our BLIP$_{\text{FUSECAP}}$ and other top-performing models.