Supplementary Material for Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering

Feng Gao[1] Qing Ping[2] Govind Thattai[2] Aishwarya Reganti [2] Ying Nian Wu [1] Prem Natarajan[2]
[1]University of California, Los Angeles, [2]Amazon
f.gao@ucla.edu, ywu@stat.ucla.edu, {pingqing,thattg,areganti,premknat}@amazon.com

A. Additional Details for Methodology

A.1. Generative Multi-Passages QA Details

Hyper-parameters To better illustrate the implementation of the generative multi-passages QA model, we introduce some key hyper-parameters in Table 1.

Hyper-Parameter	Value
Max Input Length	300
Max Decoding Length	20
Early Stopping	True
Pad to Max Length	True
Max Number of Beams	3
Learning Rate	0.0001
LR Scheduler	Linear
Total Optimization Step	20000

Table 1. Hyper-parameters of the generative multi-passages QA model, not including hyper-parameters for T5 backbone.

Input Format Different from the default input format to the pre-train a T5 model, we use a alternative formatting for the input sequences. We concatenate the question, the visual context and one retrieved Wikipedia knowledge passage as the input sequence, without any special token such as "*[SEP]*" between them. The question has a prefix "*question:* " before it. The visual context is the concatenation of image caption, dense labels and OCR text. The knowledge passage consists of a Wikipedia title and a Wikipedia paragraph. The two are concatenated by putting a prefix "*title:* " and a prefix "*context:* " before them respectively.

Vocabulary We also want to highlight the effect of the different sizes of QA model vocabulary. As in Table 2, we notice a trend that models with larger vocabulary sizes achieve higher performance. In particular, models using the default vocabulary (PICa and TRiG) perform better on OK-VQA dataset.

Method	Size	VQA Score
KRISP w/o VQA2 pre-train	2,250	32.31
Weakly Supervised VRR (C)	11,060	36.78
RVLESK	14,456	39.04
PICa (5 Ensembles)	50,257	48.00
Ours (6 Ensembles)	32,128	**50.50**

Table 2. The vocabulary size and performances of different SOTA methods on OKVQA. (C) represents classification. Some numbers may not be public accessible and we only report the numbers directly from the authors.

B. Additional Details for Ablation Study

B.1. Answer Accuracy in Beam-Search

In the main paper, we argue that the ground-truth answers of an OK-VQA question might be a semantically-similar cluster, such as (*swimsuit*, *bath suit*, *bikini*). This may also hold true for the question answering models, in terms of both classification models (top-k class prediction) and generative models (top-k beam prediction).

	Exact Match	VQA
Top-1	53.59%	49.35
Top-2	65.99%	61.61
Top-3	**71.78%**	**67.48**

Table 3. Ablation on Different k in Beam-Search Decoding.

We report the performance of our generative question answering model using top-1/2/3 beam-search decoding. As shown in Table 3, we can find that the both the Exact Match (EM) and VQA score increase as the k of beam-search increase. This suggests that while the top-one answers only achieve 49.35 VQA score, their semantically-similar candidates could reach as high as 67.48 VQA score. Therefore, we call out for new metrics that compare two sets of answers instead of top-one answer versus many ground-truth answers.

B.2. Backbone Model Size

To further illustrate the effectiveness and the efficiency of our model, we also compare the performance with various backbone model size. In Table 4, we show the VQA scores of different model backbones and highlights the approximate size of them.

Method and Backbone	Size/# Params	VQA Score
MAVEx (VilBert)	1.02GB	39.04
VRR (RoBerta-Large)	1.33GB	39.20
PICa (GPT-3)	175*B* params	48.00
Ours (w/ T5-Base)	0.85GB	46.50
Ours (w/ T5-Large)	2.75GB	**50.50**

Table 4. Performances and size of the backbone models in different methods. Since GPT-3 is not fully accessible, we only indicate the number of parameters of it which is 175 billion.

C. Additional Details for Error Analysis

To further understand the behavior of our TRiG framework, we conducted several error analysis.

Question Keywords / Types First, we investigate whether the model is likely to predict correctly over some question keywords than others. As in Figure 1-top, we can observe that majority of the questions contain the keyword "*what*", where our model is more likely to make correct predictions. On the other hand, for questions containing keywords such as "*how*" and "*why*", our model is more likely to make mistakes. We hypothesize that the "*how*" and "*why*" questions usually entail longer answers, which is harder for the generative model to predict. For example, for the question *why is this sign here*? (a sign for animal protection), the ground-truth answers are (*protect animal*, *safety*, *don't feed animal*, *direct*).

Figure 1. Distribution of correct/incorrect predictions. Top: Distributions of predictions over different question keywords. Bottom: Distribution of predictions over different question types.

Second, we report prediction distribution over 10 question types that are available from the OK-VQA dataset. As in Figure 1-bottom, we can see that our model is more likely to predict correctly on category of sports and recreation. On the contrary, our model makes more mistakes in Vehicles and Transportation and Plants and Animals.

The Impact of Visual Context and Knowledge Passages First, we would like to further investigate the effectiveness of the image-to-text transformation module, since it is the first stage in our TRiG framework. Shown in Figure 2-A, we find that if the visual contexts contain the ground-truth answers, the generative question answering model is more likely to generate a correct answer. In contrast, the model makes more mistakes if the visual contexts do not contain the ground-truth answers.

Second, we also investigate how the retrieved passages impact the generative question answering model. As is illustrated in Figure 2-B, we find that if the top-5 passages that contain the ground-truth answers, our generative question answering model is much more likely to predict correct answers. On the opposite side, if top-5 passages do not contain the ground-truth answers, it is more likely for the QA model to make a wrong prediction.

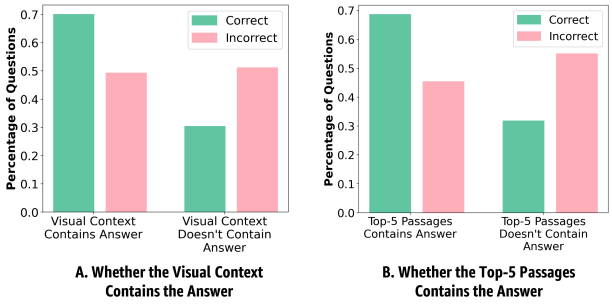


Figure 2. Error Category Break-Down.

Manual Error Review We also conducted manual eyeballing on 50 random examples where the model has made wrong predictions. We look into each example with all the available information (question, caption, dense labels, OCR text, knowledge passages, ground-truth answers) and attribute each example to one or more error categories. A brief statistics is shown in Table 5. Please note that the percentages of error types are not mutually exclusive because some wrong cases may fall in multiple categories.

We can observe that the first contributing factor to the errors is in image-to-text transformation (50%). The second category is the answer annotation ambiguity (30%), where the predicted answers are reasonable according to human judgement, but do not match any ground-truth an-

Category	Percentage
Image-to-Text	50%
Annotation	30%
Dense Passage Retrieval	24%
Generative QA	22%

Table 5. Ablation on Different k in Beam-Search Decoding.

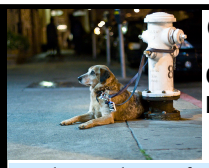
swers. There are also failures related to dense passage retrieval (24%) and generative QA model (22%). For more details of each error category, please see the examples in page 5-6.

Correct Examples

Q: What location do these vehicles stop?

GT: station, train station, train stops
Pred: station

C: The old fashioned train is setting on the track.
DL: old black train | square closed rectangular small glass window | orange yellow window | black engine | yellow train | black pant railroad steel | empty train | gray metal track | train black yellow car
K: " ...Train station A train station, railway station, railroad station, or depot (see below) is a railway facility or area where trains regularly stop to load or unload passengers or freight. It generally consists of at least one track-side platform and a station building (depot) providing such ancillary services as ticket sales and waiting rooms... "



Q: What is holding this dog here?

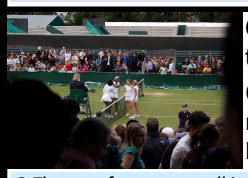
GT: leash, fire hydrant
Pred: leash

C: A dog is tied up to a fire hydrant.
DL: silver white fire hydrant | sitting brown dog | blue leash | brown black floppy ear | black number | black blue parked car ...
K: " ...is a rope or similar material used to control an animal by attaching to it or to a separate object on it; some leashes clip or tie to a collar, harness, or halter, while others go directly around the animal's neck or head. Leashes take many forms; for example: There are also bicycle dog leashes, especially designed for people who enjoy taking their pet in a ride with the bike... "

Q: Which object in the picture is described as sharp?

GT: scissors
Pred: scissor

C: A pair of black handled scissors lying on a roll of tape.
DL: plastic black handle | open round silver hole | round open plastic| black gray hole | black metal scissors | silver metal screw
K: " ...Scissors are hand-operated shearing tools. A pair of scissors consists of a pair of metal blades pivoted so that the sharpened edges slide against each other when the handles (bows) opposite to the pivot are closed. Scissors are used for cutting various thin materials, such as paper, cardboard, metal foil, cloth, rope, and wire. blade angles ideal for cutting hair... "



Q: Who won the most trophies of this sport?

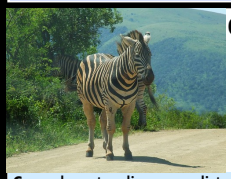
GT: serena Williams, ken rosewall, roger federer
Pred: serena williams

C: There are four woman talking at a tennis game.
DL: green chair | walking sitting standing watching person | playing tennis | watching playing woman | gray blond hair |watching stadium
K: " ...Williams sisters The Williams sisters are two professional American tennis players: Venus Williams (b. 1980), a seven-time Grand Slam title winner (singles), and Serena Williams (b. 1981), twenty-three-time Grand Slam title winner (singles), both of whom were coached from an early age by their parents Richard Williams and Oracene Price... "

Q: What is this machine used for?

GT: tell time, time tell, time, to tell time
Pred: tell time

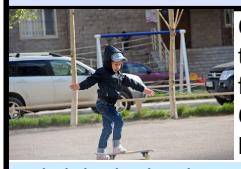
C: Signs are attached to a light pole, featuring a large clock.
DL: round white black clock | rectangular black white blue sign | clear blue sky | leafy large tall green tree | white brown sign green sign
O: Tojletlcs Funiculaire [|Halle $' PIERRE Thouire Arenes
K: " ...A clock is an instrument used to measure, keep, and indicate time. The clock is one of the oldest human inventions, meeting the need to measure intervals of time shorter than the natural units: the day, the lunar month, and the year. Devices operating on several physical processes have been used over the millennia... "



Q: What food does this animal eat?

GT: plant, grass, vegetation
Pred: vegetation

C: a zebra standing on a dirt road with trees.
DL: striped standing black white zebra| bushy long black tail | short black white mane | white black striped head
K: " ...zebras are highly water-dependent and are usually found within 25km of a water source. In one study, the zebra's diet was estimated to be 92% grass, 5% herbs, and 3% shrubs. Unlike many of the large ungulates of Africa, the plains zebra does not require (but still prefers) short grass to graze. It eats a wide range of different grasses, preferring young, fresh growth where available... "



Q: What video game was made with the name of arguably the most famous athlete in this sport?
GT: tony hawk, shaun white
Pred: tony hawk

C: The kid is skateboarding on the street while wearing a jacket
DL: growing green grass blue jean | gray blue skateboard | white blue hat | open white glass | closed window | stone brick building | wood pole black jacket | pink white shoe
K: "... (with Tony Hawk, Michael Phelps, and Alex Rodriguez) in 2008 and (alongside Jimmy Kimmel) in 2010. In a 2008 video promoting Nike's Hyperdunk shoes, Bryant appears to jump over a speeding Aston Martin. The stunt was considered to be fake, and the "Los Angeles Times" said a real stunt would probably be a..."

Q: Which part of the body might be particularly benefited by the use of this beverage?
GT: eye, brain
Pred: eye

C: Two glasses of juice are on a cutting board near diced vegetables.
DL: small orange | sliced carrot | plastic white metal | silver sharp blade | cut red sliced orange carrot | black sharp metal silver knife | filled half full glass | full clear glass | cut orange sliced carrot
K: "... juices which, unlike Western juices, usually depend on carrots and fruits instead of large amounts of tomato juice for their flavor. In general, vegetable juices are recommended as supplements to whole vegetables...which found that juices provide similar health benefits... "

Correct Examples

Q: What sport might this animal be used for?

GT: horse race, race, polo
Pred: polo

C: Woman outside her car approaching to pet a horse in fence.
DL: driving black silver parked car | smiling standing woman | leafy tall large green tree | standing white gray horse | wood fence | short brown red hair | clear dark black glasses
K: " ...to bring race horses to the track, to accompany them as they warm up for exercise, and then pick them back up after they run. Pony riders are required to wear helmets and safety vests when on the track with their charges. control of the ponied horse. The pony horse must have a calm and steady disposition... "

Q: What flavour of cake is this?

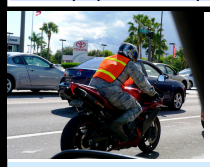
GT: vanilla, lemon, lemon vanilla
Pred: vanilla

C: A tall white cake with red flowers on top and some orange pots.
DL: large white cake | yellow sign | frosted chocolate | white cupcake | white fence | frosted white chocolate cupcake
O: Vanzlla LEMON
K: " ...Additional ingredients can be used, such as orange juice, orange muscat, milk, white dessert wine, or Riesling wine, orange oil or tangerine oil (or both), almond extract and vanilla extract. Some variations exist, such as being prepared without the use of flour. It can also be prepared as an upside-down cake... "

Q: Why are they carving pumpkins?

GT: halloween
Pred: halloween

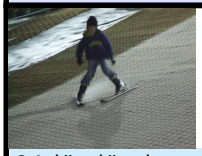
C: Two boys carving pumpkins while a lady watches.
DL: empty wine clear glass | big round large orange pumpkin | red bowl | standing woman | playing smiling young standing boy | blue jean | kitchen dark black glasses | brown wood cabinet
K: " ...Pumpkins are commonly carved into decorative lanterns called jack-o'-lanterns for the Halloween season in North America... The practice of carving pumpkins for Halloween originated from an Irish myth about a man named. The turnip has traditionally been used in Ireland and Scotland at Halloween... "



Q: Why he is having an orange vest?

GT: safety, to be visible to other, for protection, visibility in traffic
Pred: safety

C: A man is riding a motorcycle on a street in traffic.
DL: parked black blue car | yellow orange vest | parked gray silver car | white line | chrome round blue silver mirror...
O: TOYOTA
K: " ...Orange is the colour most easily seen in dim light or against the water, making it, particularly the shade known as safety orange, the colour of choice for life rafts, life jackets or buoys. Highway temporary signs about construction or detours in the United States are orange, because of its visibility and its association with danger... "



Q: When was this sport invented?

GT: 1850, 1700s', 1930
Pred: 1850

C: A skiier skiing down a slope with their skiis.
DL: long gray silver black ski | black glove | white snow black boot | standing young skiing boy | black blue jacket | black glove | black hat gray black boot | gray blue leg
K: " ...History of skiing Skiing, or traveling over snow on skis, has a history of at least eight millennia. Originally purely utilitarian, starting in the mid-1800s skiing became a popular recreational activity and sport, becoming practiced in snow-covered regions worldwide, and providing a market for the development of ski resorts and their related communities... "

Q: "What nationality is this food?

GT: american, germany
Pred: american

C: A hotdog on a plate with two green things.
DL: cooked brown long hot dog | white paper | white table | white black shadow | round white plate | cast black dark shadow
K: " ...Japanese Fusion Dogs are not actually from Japan but are a Pacific Northwest invention that pairs hot dogs with Japanese and Asian condiments like wasabi, kimchi and teriyaki. In October 2016 the Malaysian Islamic Development Department ruled that hot dog vendors must rename their product or risk not getting halal certification... "

Q: What food do these animals eat?

GT: hay, grass
Pred: hay

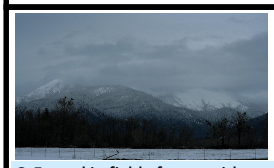
C: A man walks a horse, while people take photographs.
DL: black brown horse | black pant long brown tail | pink purple flower | standing walking man | large black camera | blue saddle | gray black hoof|white bag | black shoe | cement stone | brown wood...
K:"... Horses are grazing animals, and their major source of nutrients is good-quality forage from hay or pasture. They can consume approximately 2% to 2.5% of their body weight in dry feed each day. Therefore, a adult horse could eat up to of food. Sometimes, concentrated feed such as grain is fed in addition to pasture or hay..."

Q: What is the purpose of this vehicle?

GT: transportation, travel transport good, carry freight
Pred: transport good

C: A train makes its way down a train track.
DL: long red yellow train | circular small round window | cloudy blue sky | large clear glass windshield | yellow black engine | tall gray metal pole | black yellow front black yellow stripe | gray metal pole train
K: "... goods. Overland trains are used to carry cargo over rough terrain. Much of the world's freight is transported by train, and the rail system in the United States is used mostly for transporting freight rather than passengers and also more energy efficient than transporting freight by road. Rail freight is most economic when goods... "

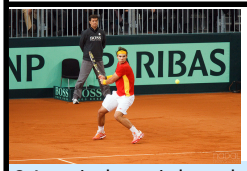
Failures Related to **Image-to-Text Transformation**



Q: What kind of trees are shown?

GT: evergreen, pine, fir
Pred: maple

C: Fenced in field of snow with mountains and overcast sky.
DL: white wire metal fence | leafless bare tree | covered snowy snow | covered mountain | white blue cloudy gray sky
K: "...The forests in the national park, which exhibit the characteristics of European-Siberian vegetation...Other notable trees include broadleaves such as oak (5%), alder, aspen, maple, dogwood... Coniferous trees predominate in the hemiboreal zone, but a significant number of deciduous species, such as aspens, oaks, maples... "



Q: In what language is the background advertisement being shown in?

GT: spanish, italian
Pred: english

C: A tennis player winds up a backhand.
DL: white short | empty black chair | playing watching standing man | large black letter | yellow orange red shirt | yellow tennis black racket | gray pant | green wall | yellow headband |metal fence | brown clay orange court
O: BOSS RIBAST VP BOss OSS napol
K: "...Tennis shot was pioneered in the 1970s by Guillermo Vilas and Yannick Noah,... Forward-facing between-the-legs shots are also occasionally employed; they are sometimes called front tweeners. The Bucharest Backfire is an over-the-shoulder backward shot, generally used to recover lobs. ..."

Q: What type of beer is that?

GT: craft, stella artois, stout, beer
Pred: budweiser

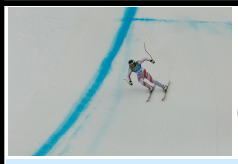
C: A glass of beer sitting next to a laptop.
DL: wine full clear tall glass | white gray silver keyboard | open blue on screen laptop | apple gray mouse pad | brown silver wood white table
K: "Beer writer Michael Jackson proposed a five-level scale for serving temperatures: well chilled for light beers (pale lagers)...Pale ale is a beer which uses a top-fermenting yeast and predominantly pale malt. It is one of the world's major beer styles...Budweiser Budweiser is an American-style pale lager produced by Anheuser-Busch ...it has grown to become one of the largest selling beers in the United States..."

Q: What is the meat called on the sandwich?

GT: pulled pork, brisket, pork, meat
Pred: beef

C: A plate of food that has some french fries and a burger.
DL: silver white napkin | slice cut sliced pickle | wine clear glass | metal silver fork | white bun | round white plate | golden french fries | brown white label
O: JQNes
K:"...The corned beef sandwich is a sandwich prepared with corned beef. The salt beef style corned beef sandwiches are traditionally served with mustard and a pickle..."

Failures Related to **Dense Knowledge Retrieval**



Q: Name the material used to make this skating board shown in this picture?

GT: fiberglass, plastic
Pred: wood

C: A person is skiing down a mountain next to a blue line in the snow.
DL: red white ski | blue line | snowy white snow | black blue white green vest | red blue white pant | white black helmet
K: "...Skateboard A skateboard is a type of sports equipment used primarily for the sport of skateboarding. It usually consists of a specially designed maplewood board combined with a polyurethane coating used for making smoother slides and stronger durability... Snowboards are generally constructed of a hardwood core which is sandwiched between multiple layers of fibreglass..."

Q: How would you cook this side dish?

GT: steam, steamed, pan fry, boil
Pred: grill

C: A white plate with some broccoli and meat.
DL: cooked green broccoli | dark shadow burnt sliced brown cooked grilled fish | white plate
K: "...fries known as steak fries. Chili, rice, pasta, or beans are also common sides. A side salad or a small serving of cooked vegetables often accompanies the meat and side, with corn on the cob, green beans, creamed spinach, asparagus, tomatoes, mushrooms, peas, and onion rings being popular... New side orders introduced within the past decade, such as rice and couscous, have grown to be quite popular throughout Europe..."

Q: Where is this bus headed to?

GT: acton, london, high street
Pred: taipei

C: A double deckered bus on a city street.
DL: double decker red bus | brick paved concrete gray sidewalk | parked red car | old brick white building | metal black pole | electronic digital yellow number | metal black bus stop | red mirror open black bus stop
O: 427 Acton Orjalan VN37365 First TEFDL5Z
K:"... double-decker buses on longer-distance routes, most notably commuter buses crossing the Bosphorus Bridge linking the European and the Asian sides of the city...."

Q: Who staged this room?

GT: staged 4 more, design, stage4more
Pred: home depot

C: The dog is resting on the floor in the living room.
DL: an brown dog | beige white wall | stacked book | tan white gray pillow | illuminated lit on lamp | colorful multi colored red rug | gray green couch | framed large mirror | open white window | brown wood coffee table
O: Before Staging After Staging stagedl more HOME STAGING REDESIGNS
K: "...Prior to filming, director Guillem Morales worked hard on a story board. For Shearsmith, the small space added to the need to meticulously plan the production process...Gleen Forbes, the set designer, thought that this made the show look cheap..."

Failures Related to **Generative Question Answering**

Q: Is this red wine or grape juice?

GT: red wine, wine
Pred: grape juice

C: A woman holding two wine glasses, one in each of her hands.
DL: empty clear wine glass | checkered plaid red scarf | silver gold ring | happy eating young smiling woman | big smiling white teeth | open dark brown eye | big large nose | short blond brown hair
O: bohemiantnaveler com
K: "...Some common types of wine glasses are described below. Glasses for red wine are characterized by their rounder, wider bowl...A wine glass is a type of glass that is used to drink and taste wine. ..."

Q: What type of car is this?

GT: old, vintage, wood car, station wagon
Pred: t

C: A classic car with a lady inside sitting in a parking lot.
DL: white surfboard | closed open green door | white brown brick building | green black tire | hanging white black sign | tinted clear open glass windshield | parked old green car | looking smiling sitting woman
O: 49 Juelytly| SealouGe
K: "...A classic car is an older automobile; the exact definition varies around the world. The common theme is of an older car with enough historical interest to be collectable and worth preserving... Division by separate eras include: antique cars (brass era cars such as the Ford Model T)..."

Q: What is the name of this type of small oven?

GT: toaster, toaster oven, microwave oven, ge
Pred: convection

C: A tray of muffins sits in an open oven while two more sit on plates.
DL: metal silver oven | marble tile white tiled countertop | brown small muffin | square white small plate | brick stone gray wall | metal silver microwave | fried cooking sliced cooked brown muffin | black silver metal tray
K: "... Toaster ovens function the same as a small-scale conventional oven. Toaster ovens typically have settings to toast bread and a temperature control... A convection microwave oven is a combination of a standard microwave and a convection oven..."

Q: What is the object of this game?

GT: score, hit ball run base, computation, run base
Pred: run

C: A baseball player is running to a base.
DL: black belt | black helmet | red gray white pant | white line | standing man | sitting baseball watching player | red white jersey | green baseball dugout | white black shoe | baseball green grass
O: PAC _ IFIC
K: "...The objectives of the offensive team are to hit the ball into the field of play, and to run the bases, having its runners advance counter-clockwise around four bases to score what are called 'runs'..."

Failures Related to **Ambiguous Answer Annotation**

Q: Name the model of train shown in this picture?

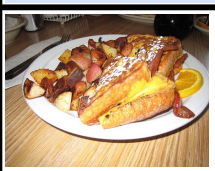
GT: subway, lionel, passenger, cummute
Pred: commuter

C: A red train traveling past a white train.
DL: red train | blue white train | red door open glass red window | yellow line | steel railroad | empty train | white green sign | gray yellow platform | white green sign | empty train station
O: DLR Station
K: "...Passenger operations include Amtrak, Metra, the Chicago Transit Authority's 'L' and Chicago's South Shore Line trains. The museum had an earlier model railroad layout..."

Q: What is the man doing with his phone?

GT: watch video, picture, video tape, take photo
Pred: take picture

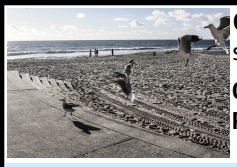
C: A person is holding up their cell phone to take a picture.
DL: up raised open holding white hand | raised up extended bent long thumb | red black phone | thin light hairy wrist
K: "...A selfie is a self-portrait photograph, typically taken with a smartphone which may be held in the hand or supported by a selfie stick...Smartphones can use their front camera (of lesser performance as compared to rear camera) facing the user for purposes like self-portraiture (selfie) and videoconferencing..."



Q: How might this be prepared?

GT: fried, pan, frypan
Pred: grilled

C: A plate of french toast and breakfast potatoes.
DL: grilled sliced fried cooked potato | brown wood table | sliced toasted grilled fried fish | silver white napkin | silver knife | glass dark black bottle | cooked red bacon
K: "...A baked potato, or jacket potato, is a potato that has been baked for eating. When well cooked, a baked potato has a fluffy interior and a crisp skin...French toast French toast is a dish made of bread soaked in eggs and milk, then fried..."



Q: How would you dress for this setting?

GT: short, bath suit, bikini, summer
Pred: swimsuit

C: A lot of seagulls flying around at the beach.
DL: sitting standing small walking black gray bird | cast black dark shadow cloudy white blue sky | standing walking person | small white cloud | gray calm large blue water | black sandy wet sand | white flying gray seagull
K: "...Beach balls are also a popular prop used in swimsuit photography and to promote or represent beach-themed events or locations...The video featured one dancer and Kumi sporting several fashions, including a crop top with black shorts and gold chains and a bustier with tight-fitting leggings..."