Incorporating Dense Knowledge Alignment into Unified Multimodal Representation Models

Supplementary Material

6. Details of Model Training

In this section, we provide additional details on model training to demonstrate the reproducibility of DeKR.

6.1. Details of Hyperparameters

We present the training hyperparameter settings for DeKR-2B and 7B, respectively, as follows:

Table 9. Hyperparameters of DeKR Training

Hyperparameters	DeKR-2B	DeKR-7B
Initial Weight	Qwen2VL-2B	Qwen2VL-7B
Batch Size	16000	16000
Learning Rate	0.0001	0.0001
Weight Decay	0.01	0.01
Embeddding Dim.	4096	4096
Attention Type	Bidirectional	Bidirectional
Num Epochs	1 (320 steps)	1 (320 steps)
Scheduler	Cosine	Cosine
Warmup Ratio	0.01	0.03
Consumption	12 gpu*days	40 gpu*days
Precision	bf16	bf16
DeepSpeed	Zero2	Zero2
Image Size	448	448
Video Frames	8	8

Except for the aforementioned hyperparameters, we adhere to the original Qwen2VL settings for all other aspects.

6.2. Details of Image Resizing

The CLIP series models are constrained by Vision Transformers (ViT) as they forcibly scale image sizes to squares, thereby disrupting the original aspect ratio information. In contrast, DeKR scales images proportionally according to their aspect ratios while maintaining the same pixel counts. For example, given an input image of 640×480 pixels, to resize it to the dimension of 448, we first calculate the ratio of pixel counts: $640 \times 480/(448 \times 448)$. Then, we scale the image according to this ratio, resulting in a size of 517×388 pixels. Finally, to ensure divisibility by the patch size of 28, the image is interpolated to 504×392 pixels. This scaling method better preserves the image content without loss.

6.3. Details of Ranking

Inference process. This subsection further elucidates the inference process behind the DeKR results presented in Tables 7 and 8. To conserve inference time for the ranking task, we employ a coarse-to-fine ranking strategy. Initially, we utilize our DeKR representation model to retrieve results, selecting the top 30 ranked items for each sample. Subsequently, the DeKR-Rank model re-ranks these top 30 results. We have observed that the top 30 rankings produced by the DeKR representation model are already nearly perfect, ensuring that this two-step process does not compromise the model's prediction accuracy.

Training data. As described in Section 4.3, the training data for the ranking task consists of 500k image-text pairs randomly selected from DeKon5M as positive samples. Here, we further elaborate on the strategy for selecting negative samples. We compute the SigLip alignment scores for all samples and sort them based on these scores. After removing the positive samples from this sorted list, the remaining sequence serves as a difficulty-graded set of negative samples. For the Flickr task, we randomly select negative samples from four difficulty intervals: $[6, 24)$, $[24, 120)$, $[120, 600)$, and $[600, 3000)$.

Training hyperparameters. In the zero-shot training and fine-tuning of the DeKR-Rank model, several hyperparameters were slightly modified. The altered parameters are as follows:

Table 10. Altered Hyperparameters of DeKR-Rank Training

Hyperparameters	DeKR-Rank
Initial Weight	DeKR-7B
Learning Rate	0.00003
Num Epochs	2 (196 steps)
Warmup Ratio	0.01
Consumption	10 gpu*days

7. Details of Downstream Tasks

7.1. Image-Text Retrieval

COCO and Flickr are the most commonly used datasets for evaluating image-text retrieval tasks. The COCO test set is available in both 1K and 5K versions, we utilize the more challenging 5K version. This version comprises 5,000 images, each accompanied by five correct descriptions, totaling 25,000 description texts. Flickr only offers a 1K version, which includes 1,000 images, each with five correct descriptions, amounting to 5,000 description texts. For retrieval tasks, a correct recall is achieved as long as at least one accurate description is retrieved.

7.2. Video-Text Retrieval

For video retrieval, we evaluate two well-known datasets, MSR-VTT and DiDeMo, which are commonly used for comparing video retrieval models. The MSR-VTT test set comprises 1,000 videos and their corresponding descriptions, while the DiDeMo test set includes 1,004 videos and descriptive paragraphs. During testing, we uniformly select 8 frames as input.

7.3. Composed Image Retrieval

Composed image retrieval extends the traditional image retrieval task by requiring consideration not only of image content similarity but also of textual instructions during the retrieval process. We evaluated our approach on two datasets: FashionIQ and CIRR. FashionIQ is divided into three subsets—Dress, Shirt, and Toptee—comprising 1,085, 678, and 709 queries respectively. Each query consists of a query image and a textual instruction, with the corpus containing 15,415 images. In addition to testing these three smaller subsets, we followed the VISTA protocol to assess retrieval performance on the entire dataset. Full-set (*val-all*) retrieval utilizes the entire validation set as queries, totaling 6,016, and incorporates the training, validation, and test sets into the corpus, amounting to 74,381 images. This approach more effectively evaluates the model's composed retrieval capabilities. Furthermore, on CIRR, we employed full-set retrieval as the evaluation method, encompassing 4,181 queries and a corpus of 21,551 images.

7.4. Composed Text Retrieval

Composed text retrieval is an extension of text retrieval, incorporating additional image information during the retrieval process to enhance accuracy. WebQA is a typical composed text retrieval task, primarily retrieving multimodal content that composed images and text through text queries. The WebQA dataset comprises two subtasks: one is text retrieval task, containing 2,455 queries and a substantial corpus of 544,489 texts; the other is composed text retrieval task, featuring 2,511 queries and a corpus of 403,277 multimodal contents. Additionally, ReMuQ is a recently popular composed text retrieval benchmark that resembles the VQA task in its format. Each query in ReMuQ consists of a question and a related image, and the test set includes a total of 3,609 queries, with the retrieval objective being to identify the correct answer from 195,837 candidate answers.

8. Q & A

8.1. Why prefer MLLMs as representation models?

1) In the text domain, LLMs have extensive internal knowledge, making LLM-to-Vector representations perform better than traditional BERT-based retrieval models. Therefore, we believe that MLLMs also have the same potential in the multimodal domain. **2)** Current MLLMs significantly outperform CLIP series models in fine-grained understanding abilities. We believe these capabilities can be transferred to retrieval tasks.

8.2. Will the computational cost be higher?

The inference process is the same as CLIP model. Since MLLM-based models share the text encoder (LLM) for both visual and textual encoding, their computational cost is about 1.5 times that of similarly sized CLIP series models. This results in a slight increase in computation but brings significant performance improvements.

9. Further Case Presentations

In this chapter, we present additional cases, including some examples from DeKon5M and the results of DeKR on downstream tasks. Please refer to the following pages for detailed information. The figures [9,10,11,12,13] in the first five pages illustrate data samples for various tasks within DeKon5M, while the remaining figures [14,15,16,17] display the retrieval results of DeKR.

The image captures a moment of tranquility featuring a kestrel, a bird of prey, in its natural habitat. The kestrel, with its brown and gray plumage, is the focal point of the image. Its yellow beak and black eyes are clearly visible, adding a splash of color to its otherwise muted tones. The bird's head is slightly tilted to the left, as if it's attentively observing something in the distance. The background is a blurred green color, providing a stark contrast to the kestrel and drawing the viewer's attention to it. The image does not contain any discernible text. The relative position of the kestrel to the background suggests it is in the foreground, further emphasizing its presence in the image.

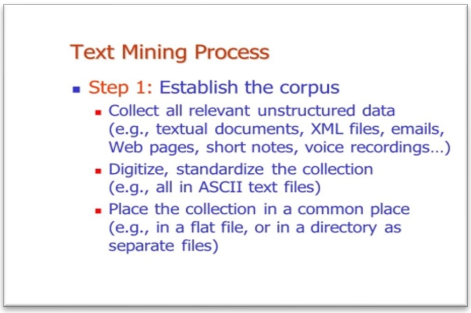
The image presents an aerial view of a wooden boardwalk winding its way through a lush green forest. The boardwalk, constructed from wooden planks, meanders through the dense foliage, disappearing into the verdant expanse of trees and shrubs.
On the left side of the boardwalk, a small wooden hut with a thatched roof nestles comfortably amidst the greenery. The hut, with its rustic charm, adds a touch of human presence to the otherwise untouched natural landscape.
The sky above is a clear blue, providing a striking contrast to the greenery below. In the distance, the faint outline of mountains can be seen, adding depth and dimension to the scene.

The image captures a serene scene of a clear blue sky dotted with fluffy white clouds. The clouds, scattered across the expanse of the sky, appear to be in motion, as if dancing in the wind. The sky is devoid of any text or human-made objects, offering a pure, unadulterated view of nature's beauty. The image is taken from a low angle, making the viewer feel as if they are looking up at the sky, further emphasizing the vastness and tranquility of the scene. The colors in the image are vibrant, with the blue of the sky contrasting beautifully with the white of the clouds. The image does not contain any discernible objects or actions, and there are no texts present. The relative positions of the clouds to each other and to the edges of the image create a sense of depth and perspective. The image is a testament to the simple yet profound beauty of the natural world.

Figure 9. Image-Text Retrieval Data with Dense Knowledge in DeKon5M.

An old photograph of a diner in the atlantic city boardwalk

A view of lake superior from the summit of acadia mountain



Text mining process step 1 establish the corpus collect all relevant unstructured data

Traffic light switches from green to red near the pedestrian crossing sign

View over tree into deep misty valley within daybreak. foggy and misty morning on the sandstone view point in national park

Figure 10. Image/Video-Text Retrieval Data with Sparse Knowledge in DeKon5M.

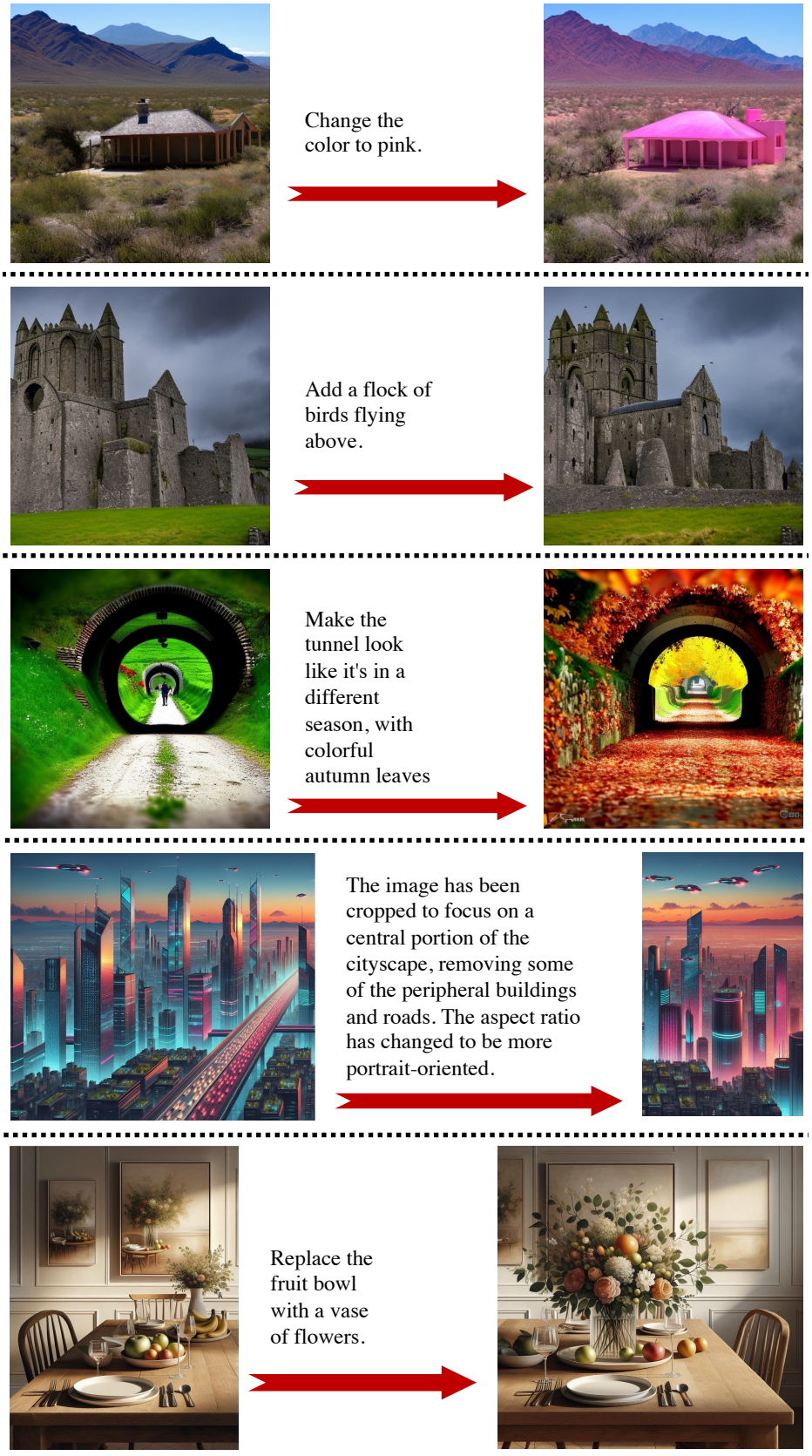


Figure 11. Image+Text to Image Retrieval Data in DeKon5M.

What activity do the two main men in the image likely enjoy, and how can you determine that?

The two main men in the image likely enjoy surfing, as evidenced by them walking on the beach and carrying surfboards. Both men are holding a surfboard alongside them, which is a clear indication that they have either just finished surfing or are heading out to surf. Their presence on a sandy beach further supports the notion that they are engaged in a water sports activity, specifically surfing.

Are the baseball players sitting on a bench inside the dugout?

Yes, the baseball players are sitting on a bench inside the dugout.

How has the presence of large windows affected the ambiance of the living room?

The presence of large windows in the living room has created a bright and well-lit ambiance, contributing to a welcoming atmosphere. The ample natural light coming in through the windows accentuates the stylish decor and living room furniture, making the space feel more pleasing and comfortable. Furthermore, the natural light enhances the colors and textures of potted plants, couches, tv, and other decorations, making the room feel more visually appealing and vibrant.

What is the bird doing in the image?

The bird is flying in the air, over the water and near the ocean shore as the wave crashes into the rock. It appears to be enjoying or simply carrying on with its activity amidst the dynamic ocean scene happening beneath it.

Figure 12. Image+Text to Text Retrieval Data in DeKon5M.

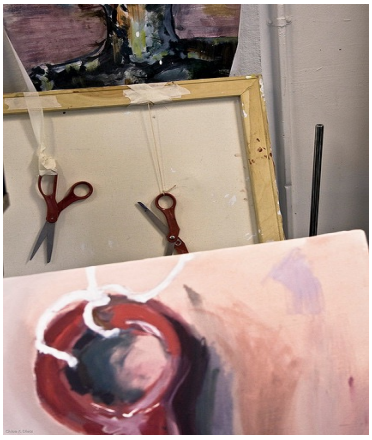
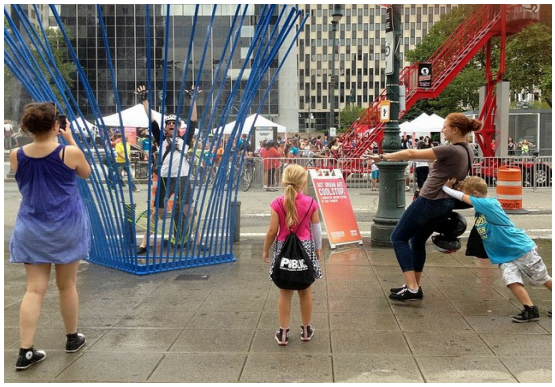
What are the tools needed for painting?	A work in progress. The unfinished painting on the easel captures the essence of creativity and the journey of artistic expression.	
How to celebrate achievement and success with a loved one?	The thrill of victory and the joy of companionshi p captured in one frame.	
What imaginative ice cream flavors are featured in a playful cartoon illustration of an ice cream shop?	As an ice cream connoisseur, I\'m intrigued by the whimsical flavors displayed at this cartoon ice cream parlor. From "Giggling Grape" to "Silly Strawberry," these imaginative treats not only tickle the taste buds but also ignite a sense of childlike wonder in every scoop.	

Figure 13. Text to Image+Text Retrieval Data in DeKon5M.



Top1

Woman taking a picture of someone standing behind a sculpture and a child pushing another woman towards the sculpture .

Top2

People stand near a large modern art sculpture .

Top3

People in a inner city courtyard watching a performance .

Top4

a group of people standing around a cage while holding frisbees

Top5

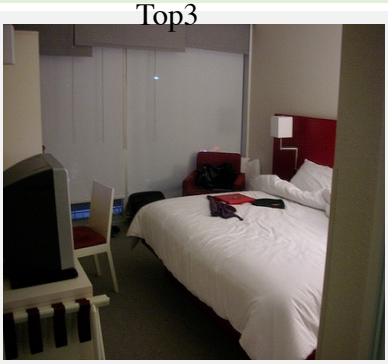
A person bending down to pick something up with people watching behind a fence .

A fluffy white chair that faces away from a television .

Top1

Top2

Top3



Top4

Figure 14. Results of DeKR on Image-Text Retrieval Tasks.

Top1	Top2	Top3	Top4	Top5
news reporters talk about a strange sight in part of san diego ✓	a news anchor is interviewing a person on screen	a man in a suit and a woman wearing brown giving the news	anchor talking about a shows	a news reporter talks about a shooting

Adding ingredients to a pizza

Top1

Top2 ✓

Top3

Top4

Figure 15. Results of DeKR on Video-Text Retrieval Tasks.

Figure 16. Results of DeKR on Composed Image Retrieval Tasks.

what colors are the cranes at the construction around the old at soho and king?

Top1

The cranes at the construction around the old Westinghouse building at Soho and King are red and white.

Top2

At the construction site of the ECB building, two cranes are visible.

Top3

The cranes are red and white.

Top4

A "media tower" – a scaffold for billboards, operated by Branded Cities – has been constructed on the northwest corner of Yonge and Dundas.

Top5

The cranes on the top level of construction at Shanghai Tower are red.

How many lit neon signs are in the windows of Thai Home restaurant in Sandy, Oregon?

Top1

Thai Home Restaurant - Sandy, Oregon

Top2

Regional Thai Taste restaurant Cheam London Borough of Sutton The Regional Thai Taste restaurant in Cheam Village in London Borough of Sutton. The restaurant opened in the early 1990s.

Top3

Thai Patio Restaurant

Figure 17. Results of DeKR on Composed Text Retrieval Tasks.