

FineLIP: Extending CLIP’s Reach via Fine-Grained Alignment with Longer Text Inputs

Supplementary Material

6. Details of long-text-to-image generation

In Fig. 3, the captions used to generate images from top to down are listed.

- (1) *A rather low aerial view from an airplane at a football field with bleachers on each side. The football field is in the center of the frame and angled from bottom left to upper right. The image is blurry. The field is green with white football striping and markings. The goal posts are on each end. There is a red track and deck around the field. The bleachers on the left side have a control box on its top center. Behind the football to the right is a soccer field. To the left of the soccer field is a baseball field. The forefront is a city street void of traffic. The flat top of a commercial building is in the bottom right corner of the frame. Across the top of the frame in the background is a city neighborhood that is speckled with house rooftops and trees.*
- (2) *An outdoor close-up view looking up at a painting on a tall worn down gray colored wall that has cracks and dark markings spread throughout its surface. Towards the bottom of the painting is a large flower with dark pink petals and an orange stigma, below the flower are faded black painted block letters placed side by side that have a white faded blob on them. Above the flower and to its sides are small scattered out yellow painted stars that are circle shaped, and about a foot above the flower is a large painting of a beige colored dragonfly with a pink head and white wings. Above the gray colored wall is a silver metal caged fence, the clear light blue sky can be seen through the fence and above it.*
- (3) *A side view of a white Chevrolet Silverado pickup truck parked on the street outside of a two story home that’s painted a light blue color on the outside with white colors around the windows and corners across the home. The white Chevrolet truck has black rims and has slight damage and scratches along the visible body. The truck blocks the full view of the yard outside the home, but two large trees with numerous leaves are visible on the left and right of the view, with green grass and dry patches visible below. At the bottom of the view, healthy green grass can be seen in the bottom right, while green plants are visible to the left growing out of the dirt. The plants and grass are separated by a black plastic divider located in the bottom middle portion of the view. Numerous shadows are cast as the view is illuminated by sunlight, At the bottom of the view, the street is covered in shadows from nearby trees, branches, and*

leaves. The white Chevrolet pickup truck is partially bright from sunlight, but also a shadow is visible from the driver side mirror extending toward the bottom left of the view. In the background behind the truck, the tree to the left in the yard is bright from sunlight, while the tree to the right is slightly darker and cast in partial shadows.

- (4) *An indoor medium close-up front view of a doorway that has a white wooden frame and a white door that is currently swung open to the left side. The doorway leads into a lit up bathroom that has a light colored wooden floor made up of wooden panels placed side by side. On the right side of the bathroom is a white countertop that consists of one sink, and dark brown drawers and cabinet doors below it. There is a white bath mat placed on the floor, in front of the countertop. To the left side of the countertop is a white wall that has a rectangular shaped mirror mounted to it that is positioned vertically. Behind the white wall is a partial view of a white toilet that is pointed out towards the left, and to the left of the toilet are several white towels hanging from a towel rack that is mounted to the wall.*

Combining the Fig. 3 and the provided captions here, we can have the following observations:

- (1) For the first example, LAPS and our FineLIP can capture the keywords “red track and deck around the field”, while the others cannot.
- (2) In the second caption, the keyword “dragonfly” is only caught by Baseline and FineLIP, as shown at the bottom left of noteworthy flower.
- (3) Regarding the “white truck” in the third caption, CLIP(L/14) and SPARC fail, Baseline seems to just give a partial view, while LAPS and FineLIP presents the whole truck with higher quality.
- (4) Finally, in the last example, the generated image from FineLIP accurately reflects the keywords “cabinet” and “toilet” as specified in the caption, while other models fail to do so.

Although these visualized examples highlight the effectiveness of FineLIP compared to other state-of-the-art methods, many details (such as color, attributes, and locations) are still missing in the generated images. Furthermore, the gap between the generated images and the ground-truth images remains substantial. It would be both interesting and meaningful to explore how to effectively leverage long captions in the end-to-end training of text-to-image generation

		Flickr30k						COCO					
		Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
B/16	Pretrained-CLIP	0.441	0.682	0.771	0.248	0.451	0.546	0.517	0.768	0.843	0.327	0.578	0.682
	Baseline	0.470	0.719	0.805	0.334	0.564	0.659	0.553	0.800	0.869	0.409	0.669	0.766
	TULIP [19]	0.461	0.708	-	0.352	0.572	-	0.568	0.803	-	0.407	0.661	-
	FineLIP (Ours)	0.528	0.753	0.829	0.341	0.567	0.658	0.587	0.814	0.882	0.404	0.662	0.760
L/14	Pretrained-CLIP	0.485	0.727	0.809	0.280	0.493	0.587	0.560	0.795	0.869	0.353	0.600	0.701
	Baseline	0.544	0.788	0.862	0.419	0.653	0.739	0.608	0.836	0.899	0.472	0.721	0.809
	TULIP [19]	0.567	0.795	-	0.416	0.643	-	0.626	0.847	-	0.461	0.711	-
	FineLIP (Ours)	0.622	0.828	0.888	0.424	0.652	0.736	0.634	0.848	0.910	0.462	0.712	0.801

Table 4. Short caption cross-modal retrieval on *Flickr30k* and *COCO*. Best result is in bold.

models.

7. Long caption datasets details and additional results

L/14	Long-DCI [30]				IIW [8]			
	Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Baseline	0.568	0.763	0.585	0.754	0.968	1.000	0.978	1.000
SPARC [2]	0.428	0.638	0.465	0.669	0.903	0.988	0.933	0.998
LAPS [6]	0.604	0.778	0.592	0.758	0.985	1.000	0.983	1.000
FineLIP (Ours)	0.608	0.780	0.607	0.767	0.993	1.000	0.985	1.000

Table 5. Zero-shot cross-modal retrieval results on additional long caption datasets

Urban1k dataset contains 1,000 (image, long caption) pairs. The images selected from the Visual Genome dataset [14] depict busy urban scenes, with captions generated by GPT-4V [22] averaging 101 words. These detailed captions describe attributes such as object types, colors, and spatial relationships, making this dataset particularly suitable for evaluating zero-shot long caption cross-modal retrieval. **DOCCI dataset** offers long, human-annotated captions for 15,000 images, with an average description length of 136 words. *DOCCI*’s descriptions are highly detailed, covering complex visual challenges like spatial relationships between objects, counting, and text rendering. We utilize the official 5k test split for evaluation.

For long-caption image-text alignment, we added additional results on long-caption retrieval datasets, *Long-DCI* [31] and *IIW* [9] (see Tab. 5).

8. Evaluation on zero-shot short caption cross-modal retrieval and classification

Our approach, FineLIP, is specifically designed to handle long captions, so training and evaluation are conducted exclusively with long caption data. Interestingly, we find that FineLIP also gives a significant performance boost on tasks involving short captions, even though short captions are not explicitly used in training and are not the primary focus of

this paper.

To have the comprehensive evaluation, we adopt the widely used zero-shot cross-modal retrieval benchmarks for short captions: *Flickr30k* [34] and *COCO2017* [15]. In *COCO2017*, we use the 5k validation set, while for *Flickr30k*, we employ the entire 30k dataset following the standard practice. In Tab. 4, the Pretrained-CLIP²³ models refer to those released by OpenAI, trained on 400 million short-text image pairs from web sources, and demonstrate strong performance on short caption datasets. After fine-tuning on *ShareGPT4*, the long caption dataset, with traditional contrastive learning (Baseline), substantial improvements are observed across all metrics, benchmarks, and model sizes. FineLIP, which incorporates aggregation and alignment modules for more fine-grained cross-modal alignment, significantly outperforms the baseline in image-to-text retrieval and achieves comparable performance in text-to-image retrieval. TULIP, trained with long and short captions, delivers results that fall between these two approaches.

The results indicate that training with longer captions enhances the model’s generalization ability, improving performance on tasks involving short captions. This promising finding highlights the advantages of leveraging longer, more detailed captions and the importance of fine-grained alignment, even for short caption tasks. However, we also observe that FineLIP achieves text-to-image results comparable to the Baseline while providing significant gains in image-to-text retrieval. This pattern, consistent with findings from previous studies on cross-modal retrieval [18, 27, 29], needs further investigation.

For zero-shot classification, we evaluate on *DataComp-38* [7] (see Fig. 4), with Pretrained-CLIP, Baseline, and FineLIP achieving average accuracies of 0.643, 0.629, and 0.619, respectively.

²CLIP-ViT-B/16

³CLIP-ViT-L/14

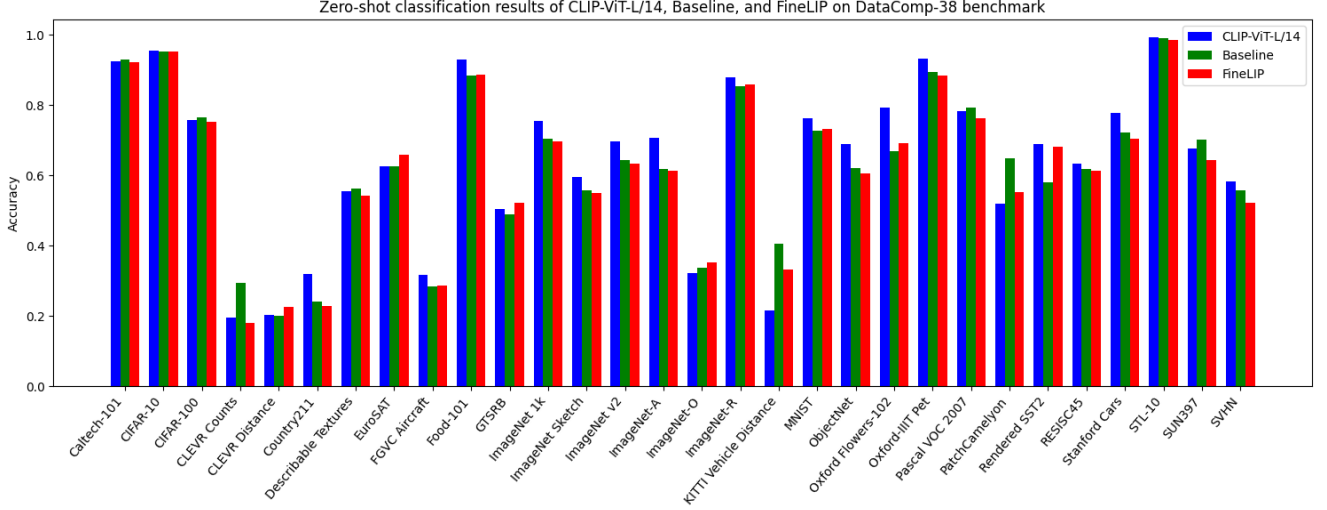


Figure 4. Zero-shot Classification on *DataComp-38*

	Urban1K				DOCCI			
	Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
ATRM + FILIP	0.315	0.525	0.436	0.657	0.354	0.620	0.426	0.694
ATRM + CLIM (Ours)	0.907	0.983	0.893	0.975	0.771	0.954	0.795	0.958

Table 6. Effect of CLIM in cross-modality alignment

9. Importance of CLIM

We conducted an ablation study on the CLIM block. Keeping ATRM fixed, we tested two cross-modality alignment methods. When using baseline contrastive loss, which is generally designed to perform on global features (the feature of [CLS] token for image and the feature of [EOS] token for text), no improvement is observed as ATRM only refines local tokens. ATRM + FILIP (Tab. 6) underperformed our solution FineLIP (ATRM + CLIM), likely due to FILIP’s training instability, as noted in its original paper [33] and SPARC [2].


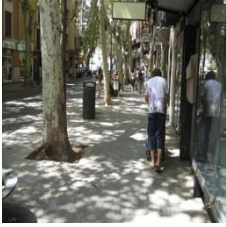






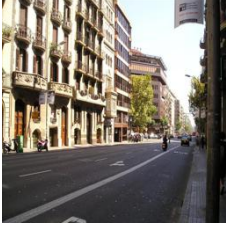
10. Visualization of long caption image-text retrieval

Fig. 5 presents the Text-to-Image (T2I) retrieval results for various models on *Urban1k* dataset. The grid shows the top-5 retrieved images for a given text query, ranked by the similarity score between the text and image features. The higher the score, the more relevant the image is to the query. The ground truth image (GT) does not appear within the top 5 retrieval results for CLIP (L/14). For the Baseline, SPARC, and LAPS, the top 1 retrieval results are not the GT but rather images that are highly similar to it. In contrast, FineLIP retrieves the GT in the top position, demonstrating its superior ability to capture fine-grained details and to distinguish among highly similar images.

Fig. 6 illustrates the Image-to-Text (I2T) retrieval re-

sults on the *Urban1k* dataset. The figure shows the top-5 retrieved captions for a given image, ranked by similarity scores. FineLIP manages to retrieve the ground truth caption in the top position, while others models fail. In particular, for models like SPARC and LAPS, the correct caption is not even within the top-5 results. This example shows that FineLIP is capable of extracting discriminative fine-grained textual features even when the text input is long.

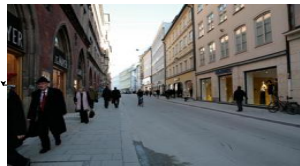
These visual results emphasize FineLIP’s advantage in enabling more precise and nuanced retrieval, especially for tasks involving complex text-to-image associations that demand detailed comprehension. FineLIP’s ability to leverage long captions and fine-grained alignment methods allows it to retrieve highly relevant content that other models may overlook. By extending the token limit and using aggregation before token-level alignment strategy, FineLIP achieves more accurate and contextually relevant retrieval than existing models.

					CLIP(L/14)
0.2813	0.2716	0.2711	0.2705	0.2684	
					Baseline
0.2423	0.2393	0.2379	0.2331	0.2281	
					SPARC
0.2424	0.2383	0.2363	0.2290	0.2284	
					LAPS
0.3997	0.3958	0.3870	0.3846	0.3492	
					FinELIP
0.4371	0.4200	0.4124	0.3913	0.3583	

Query Text: *This image captures a bustling urban street scene at twilight with the warm glow of the sun lighting the treetops. The street is lined with neatly aligned trees and traditional-style street lamps. On the right, there is a row of parked motorcycles, behind which sits a terrace of a café bustling with patrons. A pedestrian walkway runs parallel to the road, where people are strolling leisurely. On the left, the walkway is bordered by an **ornate building with stonework details and protruding balconies**. Two individuals are walking towards the camera, one of whom is holding a blue bag. The vehicles on the road hint at moderate traffic flow.*

Figure 5. Top-5 text-to-image retrieval results on *Urban1k* dataset [36] for L/14 variants of CLIP, Baseline, SPARC [2], LAPS [6] and FineLIP (Ours), with retrieval scores. The correct retrieved images are marked with green boxes. CLIP ignores the caption in bold due to the 77 token limit.

<p>This image features a bustling urban street scene, possibly in a downtown area with a mix of pedestrians and vehicles. A variety of people are seen walking on the sidewalk, some carrying shopping bags. The architecture is a combination of tall buildings, with a prominent skyscraper visible in the backdrop. American flags hang from some of the buildings, indicating a sense of patriotism or a special occasion. A NYPD (New York Police Department) van is parked on the roadside, suggesting the location could be New York City. There are trees and greenery in the background, possibly a park or planted area for aesthetic enhancement. The weather appears to be clear and sunny, casting shadows on the pedestrians and highlighting the colors of the surroundings.</p> <p>0.2805</p>	<p>The image captures an urban street scene at an intersection with pedestrians crossing the road. The traffic light for the pedestrians is red, yet they are walking, which suggests they might be crossing against the signal. There's a group of approximately eight individuals in casual attire, some wearing shorts and t-shirts, suggesting warm weather. A white van is visible and parked by the roadside, and a silver sedan waits at the intersection. The architecture is varied, with a red-brick building on the corner featuring large arched windows, indicative of classic urban design. Tall buildings line the street in the background, hinting at a dense city environment. The sky is clear, hinting at a sunny day.</p> <p>0.2776</p>	<p>The image depicts a bustling urban street scene under a mostly clear sky with some clouds. The architecture suggests a European city with a mix of classical and modern building facades featuring storefronts with green awnings. The street is busy with pedestrians who appear to be going about their daily activities. Vehicles, including a prominent white car, are visible, suggesting a shared road space with pedestrians. Pedestrian crossings and road signs are noticeable, guiding traffic and people. An overpass with a signboard is seen further down the street. The lighting indicates daytime with shadows cast on the pavement, creating a contrast of light and shade.</p> <p>0.2726</p>	<p>This image captures a bustling city street scene during the day with pedestrians and cyclists. The architecture is European in style, with classic facades featuring arched windows on the building to the left and simpler, clean lines on the right. There's a mix of warm-toned and neutral-colored buildings under a clear blue sky. Several shops, including a notable fashion retailer, line the street, displaying brightly lit window fronts. The street itself is wide, likely pedestrianized, with a smooth surface and no visible traffic marks, suggesting it may be a shopping or walking area. Some people are in business attire, indicating the area's mix among casual visitors and working professionals.</p> <p>0.2703</p>	<p>This image captures a lively street scene on a sunny day. The architecture suggests a European setting with traditional buildings featuring red brickwork and signage indicating businesses such as a coffee shop. A clear blue sky serves as the backdrop. Pedestrians are walking about, some partially obscured, wearing casual attire suited for warm weather; several men are dressed in T-shirts and shorts. A multi-directional signpost is visible, featuring white text on a black background, contributing to the urban ambience. A bicycle is parked on the side, and a mix of trees and built structures can be seen in the background.</p> <p>0.2660</p>
<p>The image shows an urban street scene, likely in a European city, with a focus on a wide road lined with classical architecture. The buildings have uniform facades with ground-floor retail shops, including a Hackett store. Pedestrians are visible, with two individuals walking past the storefronts. The street has multiple lanes marked for traffic; on the nearest lane, there's a cyclist in a blue top, blurred in motion. Cars and iconic red double-decker buses are visible in the distance, suggesting a busy thoroughfare. Traffic lights are on red, halting some vehicles. Two British flags are mounted on a building, indicating this may be in the United Kingdom.</p> <p>0.2449</p>	<p>This image portrays an urban street scene with a row of multi-story buildings. On the left, a narrow, four-story building is painted blue, sandwiched between taller brick buildings. One of the buildings on the right features signage for "Restoration Hardware" on its cream-colored facade, indicating a retail store on the ground floor. Several pedestrians are crossing the street in the foreground, and a few cars are parked at the curb. There's a traffic light at the corner, and a small barren tree is visible in front of the blue building. The atmosphere seems calm, with clear skies and daylight.</p> <p>0.2424</p>	<p>An elderly couple walks their small dog on a city sidewalk in front of a closed commercial establishment with metal shutters down. The building facade is made of stone, with ornate architectural details and balconies featuring intricate iron railings. Three flags are displayed above the entrance: a Catalan flag, an EU flag, and an American flag. The shop fronts have signage for "TOUS." The street appears calm with a car and a motorcycle parked in the background. The color palette includes various shades of grey from the building and pavement, muted tones from the walking couple's attire, and vibrant hues from the flags.</p> <p>0.2394</p>	<p>This image captures a bustling city street scene during the day with pedestrians and cyclists. The architecture is European in style, with classic facades featuring arched windows on the building to the left and simpler, clean lines on the right. There's a mix of warm-toned and neutral-colored buildings under a clear blue sky. Several shops, including a notable fashion retailer, line the street, displaying brightly lit window fronts. The street itself is wide, likely pedestrianized, with a smooth surface and no visible traffic marks, suggesting it may be a shopping or walking area. Some people are in business attire, indicating the area's mix among casual visitors and working professionals.</p> <p>0.2379</p>	<p>This image captures a busy urban street scene on a sunny day with clear blue skies. A crowd of people is walking along the sidewalk, engaged in various activities like shopping and conversing with each other. In the foreground, a woman in a blue top and pink skirt walks towards the viewer. There's a storefront with a large, illuminated sign reading "MARKS & SPENCER," suggesting a retail environment. The architecture is a mix of modern and traditional, with a glass facade on the left contrasting with the brick buildings on the right. The pedestrians wear casual summer clothing, with some carrying shopping bags, indicating a leisurely atmosphere.</p> <p>0.2366</p>
<p>The image depicts a quaint urban street scene with two adjoining storefronts, each featuring large display windows and signage. On the left, the store showcases electrical devices, while on the right, artistic lamp designs are visible in the window display. Between the stores is an entrance with ornate white door frames. A mature individual, dressed in a long black coat and carrying a bag, walks past the lamp store. In front of the image, a bicycle is parked on the sidewalk, indicating a bike-friendly area as suggested by the bike lane symbol painted on the street. The architecture and shops exude a European feel.</p> <p>0.2505</p>	<p>In this image, we see a busy urban environment with a clear blue sky, suggesting a sunny day. Various individuals appear engaged in different activities: walking, standing, and cycling. A classic red telephone box stands on the left, while scaffolding is visible in the background, indicating construction or renovation at a building. Several people are dressed in typical summer attire, such as shorts and T-shirts, in colors like white, blue, green, and black, and at least one bicycle is parked. Some of the individuals carry bags, and others seem to be waiting or looking at something outside the view of the image, possibly a storefront to the right.</p> <p>0.2495</p>	<p>The image captures a busy urban street scene on an overcast day. On the left, a pedestrian walks by a bank named "CaixaNova," with blue signage and an ATM adjacent. The corner building hosts a business with a dark facade and signage that reads "FORNE ASSOCIADOS / Marketing Imobiliário." Central to the image is a black station wagon driving away from the viewer with a visible license plate. The right side features a pharmacy with a blue cross emblem and a clothing store with mannequins in the display. Pedestrians are gathered near a zebra crossing, waiting to cross, and there is a pale yellow building at the end of the perspective.</p> <p>0.2491</p>	<p>The image shows an urban street scene, likely in a European city, with a focus on a wide road lined with classical architecture. The buildings have uniform facades with ground-floor retail shops, including a Hackett store. Pedestrians are visible, with two individuals walking past the storefronts. The street has multiple lanes marked for traffic; on the nearest lane, there's a cyclist in a blue top, blurred in motion. Cars and iconic red double-decker buses are visible in the distance, suggesting a busy thoroughfare. Traffic lights are on red, halting some vehicles. Two British flags are mounted on a building, indicating this may be in the United Kingdom.</p> <p>0.2461</p>	<p>This image captures a lively street scene on a sunny day. The architecture suggests a European setting with traditional buildings featuring red brickwork and signage indicating businesses such as a coffee shop. A clear blue sky serves as the backdrop. Pedestrians are walking about, some partially obscured, wearing casual attire suited for warm weather; several men are dressed in T-shirts and shorts. A multi-directional signpost is visible, featuring white text on a black background, contributing to the urban ambience. A bicycle is parked on the side, and a mix of trees and built structures can be seen in the background.</p> <p>0.2441</p>
<p>The image shows an urban street scene, likely in a European city, with a focus on a wide road lined with classical architecture. The buildings have uniform facades with ground-floor retail shops, including a Hackett store. Pedestrians are visible, with two individuals walking past the storefronts. The street has multiple lanes marked for traffic; on the nearest lane, there's a cyclist in a blue top, blurred in motion. Cars and iconic red double-decker buses are visible in the distance, suggesting a busy thoroughfare. Traffic lights are on red, halting some vehicles. Two British flags are mounted on a building, indicating this may be in the United Kingdom.</p> <p>0.4110</p>	<p>The image shows an urban street scene on a sunny day with clear blue skies. The street is lined with a mix of three-story brick buildings featuring ground-floor businesses, including a store with a green storefront named "Staple 2." Pedestrians wait to cross at a zebra crossing where the pedestrian signal shows a red hand, indicating a "do not walk" command. A blue bicycle is parked at a bike rack, and various banners are mounted on light poles. The scene has a combination of bare and budding trees, suggesting early spring, and shadows cast on the road hint at midday sunlight.</p> <p>0.3854</p>	<p>This image portrays an urban street scene with a row of multi-story buildings. On the left, a narrow, four-story building is painted blue, sandwiched between taller brick buildings. One of the buildings on the right features signage for "Restoration Hardware" on its cream-colored facade, indicating a retail store on the ground floor. Several pedestrians are crossing the street in the foreground, and a few cars are parked at the curb. There's a traffic light at the corner, and a small barren tree is visible in front of the blue building. The atmosphere seems calm, with clear skies and daylight.</p> <p>0.3795</p>	<p>In this image, we see a busy urban environment with a clear blue sky, suggesting a sunny day. Various individuals appear engaged in different activities: walking, standing, and cycling. A classic red telephone box stands on the left, while scaffolding is visible in the background, indicating construction or renovation at a building. Several people are dressed in typical summer attire, such as shorts and T-shirts, in colors like white, blue, green, and black, and at least one bicycle is parked. Some of the individuals carry bags, and others seem to be waiting or looking at something outside the view of the image, possibly a storefront to the right.</p> <p>0.3756</p>	<p>The image depicts a quaint urban street scene with two adjoining storefronts, each featuring large display windows and signage. On the left, the store showcases electrical devices, while on the right, artistic lamp designs are visible in the window display. Between the stores is an entrance with ornate white door frames. A mature individual, dressed in a long black coat and carrying a bag, walks past the lamp store. In front of the image, a bicycle is parked on the sidewalk, indicating a bike-friendly area as suggested by the bike lane symbol painted on the street. The architecture and shops exude a European feel.</p> <p>0.3627</p>
<p>This image captures a bustling city street scene during the day with pedestrians and cyclists. The architecture is European in style, with classic facades featuring arched windows on the building to the left and simpler, clean lines on the right. There's a mix of warm-toned and neutral-colored buildings under a clear blue sky. Several shops, including a notable fashion retailer, line the street, displaying brightly lit window fronts. The street itself is wide, likely pedestrianized, with a smooth surface and no visible traffic marks, suggesting it may be a shopping or walking area. Some people are in business attire, indicating the area's mix among casual visitors and working professionals.</p> <p>0.4051</p>	<p>The image shows an urban street scene, likely in a European city, with a focus on a wide road lined with classical architecture. The buildings have uniform facades with ground-floor retail shops, including a Hackett store. Pedestrians are visible, with two individuals walking past the storefronts. The street has multiple lanes marked for traffic; on the nearest lane, there's a cyclist in a blue top, blurred in motion. Cars and iconic red double-decker buses are visible in the distance, suggesting a busy thoroughfare. Traffic lights are on red, halting some vehicles. Two British flags are mounted on a building, indicating this may be in the United Kingdom.</p> <p>0.3990</p>	<p>This image portrays an urban street scene with a row of multi-story buildings. On the left, a narrow, four-story building is painted blue, sandwiched between taller brick buildings. One of the buildings on the right features signage for "Restoration Hardware" on its cream-colored facade, indicating a retail store on the ground floor. Several pedestrians are crossing the street in the foreground, and a few cars are parked at the curb. There's a traffic light at the corner, and a small barren tree is visible in front of the blue building. The atmosphere seems calm, with clear skies and daylight.</p> <p>0.3869</p>	<p>In this image, we see a busy urban environment with a clear blue sky, suggesting a sunny day. Various individuals appear engaged in different activities: walking, standing, and cycling. A classic red telephone box stands on the left, while scaffolding is visible in the background, indicating construction or renovation at a building. Several people are dressed in typical summer attire, such as shorts and T-shirts, in colors like white, blue, green, and black, and at least one bicycle is parked. Some of the individuals carry bags, and others seem to be waiting or looking at something outside the view of the image, possibly a storefront to the right.</p> <p>0.3834</p>	<p>The image shows an urban street scene on a sunny day with clear blue skies. The street is lined with a mix of three-story brick buildings featuring ground-floor businesses, including a store with a green storefront named "Staple 2." Pedestrians wait to cross at a zebra crossing where the pedestrian signal shows a red hand, indicating a "do not walk" command. A blue bicycle is parked at a bike rack, and various banners are mounted on light poles. The scene has a combination of bare and budding trees, suggesting early spring, and shadows cast on the road hint at midday sunlight.</p> <p>0.3816</p>



Ground Truth: This image captures a bustling city street scene during the day with pedestrians and cyclists. The architecture is European in style, with classic facades featuring arched windows on the building to the left and simpler, clean lines on the right. There's a mix of warm-toned and neutral-colored buildings under a clear blue sky. Several shops, including a notable fashion retailer, line the street, displaying brightly lit window fronts. The street itself is wide, likely pedestrianized, with a smooth surface and no visible traffic marks, suggesting it may be a shopping or walking area. Some people are in business attire, indicating the area's mix among casual visitors and working professionals.

Figure 6. Top-5 image-to-text retrieval results on *Urban1k* dataset [36] for L/14 variants of CLIP, Baseline, SPARC [2], LAPS [6] and FineLIP (Ours), with retrieval scores. The correct retrieved texts are marked with green boxes. CLIP ignores the text in bold due to the 77-token limit. Zoom in for better visualization.