

Vision-Language-Vision Auto-Encoder: Scalable Knowledge Distillation from Diffusion Models

Supplementary Material

Contents

A Additional Details and Results	1
A.1 Data Pipeline	1
A.2 VQA Analysis	2
A.3 Impact of Stage-1 training	2
A.4 Captioner Arena	2
A.5 Visual Results By Our Captions	4
B Dataset License	7
B.1 Dataset used in training	7
B.2 Dataset used in testing	7
B.3 Open-sourced models	7

A Additional Details and Results

A.1 Data Pipeline

§4.1 details our data collection and filtering procedure. We additionally annotate a subset of the corpus with *Gemini 2.0 Flash*[2]. Figure1 shows the whole pipeline how we obtain our data for Stage-1 and Stage-2. Figure2 provide the token length distribution of our captions used for training Stage-2.

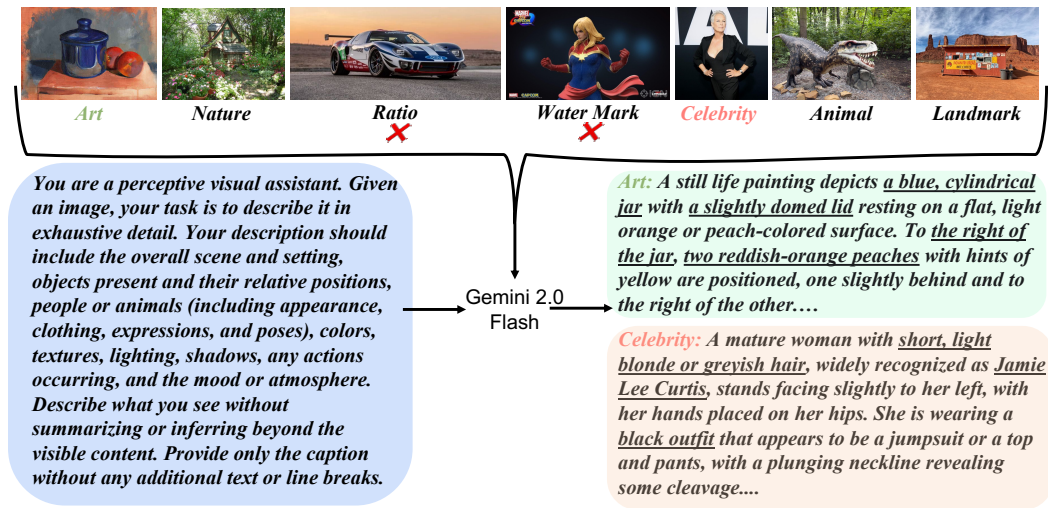


Figure 1: **An overview for filtering our 40 M dataset.** We curate a 40 M image–caption corpus by first filtering LAION-2B-en-aesthetic for high-quality images and then prompting Gemini 2.0 Flash with image-conditioned templates to generate rich, descriptive captions.

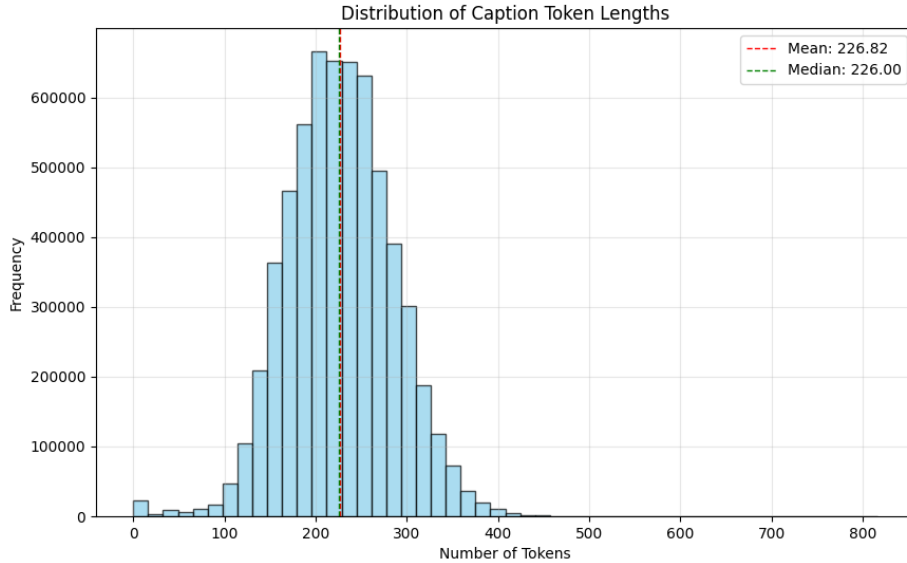


Figure 2: **Stage-2 caption length statistics.** Histogram of token counts for all captions ($N \approx 6M$). Most captions fall in the 170 – 280 token band, with mean $\mu = 226.82$ (red dashed) and median $\tilde{x} = 226$ (green dashed).

18 A.2 VQA Analysis

19 Following Wei *et al.* [3], we evaluate on OK-VQA with DeepSeek-V3 [1] under the strict *exact-match*
 20 metric. Our raw score is 45.31% (2 295 / 5 064), trailing the Gemini 2.0 Flash caption baseline of
 21 46.34% by 1.03% (52 questions). Among the 526 cases where Gemini is marked correct and our
 22 model wrong, we compute answer–answer cosine similarity in CLIP space and relabel pairs with
 23 similarity ≥ 0.8 , recovering 94 additional correct answers. The adjusted accuracy is therefore 47.17%
 24 This shows that the apparent deficit stems mainly from lexical mismatches rather than missing visual
 25 content. We show an example (one of the 94 cases) in Figure3.

26 A.3 Impact of Stage-1 training

27 Herein, we provide additional experiments to show impact of Stage-1 training. To be specific, we
 28 only train our Stage-2 with pretrained our VLV Encoder, and assess the generated captions with
 29 T2I tasks. Table1 reports the resulting FID scores (\downarrow) on MS-COCO 2014. Skipping Stage-1 (first
 30 three rows) yields very poor fidelity, even larger decoders cannot compensate, whereas with Stage-1
 31 training (grey row) drops FID to 12.2, confirming its critical role.

Decoder	1.0	2.0	3.0	4.0
Qwen-2.5 0.5B	57.29	64.02	66.92	68.18
Qwen-2.5 1.5B	49.15	56.66	59.15	60.14
Qwen-2.5 3B	45.63	51.03	53.28	54.40
Qwen-2.5 3B + Stage-1	12.22	7.55	9.58	10.84

Table 1: **Effect of Stage-1 training on FID (\downarrow).** Grey row shows the same 3B decoder after Stage-1 training; lower is better.

32 A.4 Captioner Arena

33 We assess caption quality by querying *Gemini 2.0 Flash* with a tailored rubric. Figure 4 displays a
 34 representative caption-arena outcome, together with Gemini 2.0 Flash’s rationale, confirming that our
 35 captions are on par with those from GPT-4o.



Caption by VLV: A close-up shot shows a bowl filled with three bright orange oranges, each glistening with water droplets. The bowl, which appears to be made of metal or a similar reflective material, has a dark, almost black interior. The oranges are positioned in the center of the bowl, with the one in the foreground taking up the majority of the frame. The other two oranges are partially visible in the background, slightly out of focus. The water droplets on the oranges create a sparkling effect, adding a sense of freshness and vibrancy to the image. The lighting is soft and diffused, casting gentle shadows and highlighting the textures of the oranges and the bowl. The overall mood is fresh and inviting, emphasizing the natural beauty of the fruit. The colors are vibrant and saturated, with the bright orange of the oranges contrasting against the dark background of the bowl. The composition is simple yet effective, drawing the viewer's attention to the oranges and their natural beauty. The focus is sharp on the oranges in the foreground, while the background is slightly blurred, creating a depth of field that draws the viewer's eye to the main subject. The image evokes a sense of simplicity and naturalness, highlighting the beauty of everyday objects. The overall impression is one of freshness, simplicity, and natural beauty.

Question: What states are these grown in?

Our Answer: Florida, California, Texas, Arizona

Gemini Answer: Florida, California

Ground-Truth:

["florida california", "florida california", "florida california", "florida california", "california", "california", "california", "california", "california", "california", "california", "florida", "florida"]

Caption by Gemini: A close-up captures several bright orange oranges with tiny water droplets clinging to their textured surfaces, situated in a silver or gray metallic bowl. The bowl's rim is visible along the bottom left quadrant of the image, showing slight discoloration and wetness from the condensation or water. The oranges are arranged in a clustered fashion, with one in the center and others partially visible around it, suggesting a group. The light is distributed unevenly, with highlights on the wet parts of the oranges and the bowl, and shadows cast on areas where they touch. The overall color palette is dominated by oranges, with the silver or gray bowl providing a neutral contrast. The texture appears rough due to the visible pores on the orange peels and the water droplets adding a reflective quality. The mood is fresh and inviting, potentially evoking a sense of healthy eating or refreshing citrus. On the left side of the photo is text, oriented vertically, in a thin, sans-serif font, that reads "Abi Booth Photography."

Figure 3: OK-VQA Example. Both our caption and Gemini caption do not mention the states information. But our caption not only capture the oranges but also the number of oranges. Our answers contain the right ones highlighting in green.



Qwen-2.5 caption: "The image shows a black and white cat lying on top of a laptop. The cat appears to be resting or lounging, with its body stretched out across the surface of the laptop. The laptop is open, displaying its keyboard, and the cat's head is positioned near the upper part of the laptop screen. The background seems to be indoors, possibly on a desk or table. The cat has a fluffy coat with a mix of black and white fur, and its eyes are partially closed, giving it a relaxed appearance."

Rating: 5.

Explanation: "The caption accurately describes the image, covering the main elements like the cat, laptop, and their relationship. It also mentions the cat's appearance and the indoor setting. There are no hallucinations."

GPT-4o caption: "A black and white cat with long fur and whiskers is lying stretched out inside an open laptop. The cat's body occupies the entire space where the screen would be, with its head resting on the left side of the laptop. Its ears are perked up, and it has an alert expression with its yellow eyes slightly narrowed. The keyboard of the laptop is visible in the foreground, featuring a silver color with white keys. The lighting is bright and even, casting soft shadows around the edges of the cat's fur. The texture of the cat's fur appears silky and smooth, while the laptop has a metallic and sleek appearance. The mood is lighthearted and quirky, suggesting a relaxed and playful use of the laptop as a resting place by the cat."

Rating: 6.

Explanation: "The caption accurately describes the key elements of the image, including the cat's appearance, position, and expression, as well as the laptop's features and overall setting. It also correctly identifies the spatial relationships between the objects."

VLV caption: "A close-up shot shows a black and white cat lying on top of a laptop, with the laptop's keyboard visible in the foreground. The cat's fur is a mix of black and white, with a prominent white patch around its face and chest. Its eyes are a striking yellow, and its whiskers are long and white. The cat's posture is relaxed, with its head turned slightly to the left. The laptop is silver and has a gray keyboard. The background is a plain white wall. The lighting is soft and diffused, creating subtle shadows on the cat's fur and the laptop's surface. The overall mood is calm and domestic. The focus is sharp on the cat and the laptop, while the background is slightly blurred. The texture of the cat's fur appears soft and fluffy, contrasting with the smooth surface of the laptop. The image is well-composed, with the cat positioned in the center of the frame, drawing the viewer's attention to its features and the laptop's keyboard."

Rating: 6.

Explanation: "The caption provides a comprehensive description of the image, covering the cat's appearance, the laptop, and the overall scene. It accurately describes the colors, textures, and spatial relationships between the objects. There are no hallucinations."

Figure 4: Captioner Arena Example. All captions show the correct objects without hallucinations. Both our caption and GPT-4o caption show the spatial relationship while Qwen-2.5 VL does not.

Your role is to serve as an impartial and objective evaluator of an **image caption** generated by a Large Multimodal Model (LMM). Based on the single image input, assess the caption on *three* main criteria:

1. **Coverage of image elements** – how well the caption mentions the salient objects, their attributes, actions, and contextual details.
2. **Absence of hallucinations** – the caption must not invent objects, attributes, counts, spatial relations, or other details not present or implied by the image.
3. **Object spatial layout consistency** – whether spatial relationships (*left/right, above/below, front/behind, center, background/foreground*) are described accurately.
 - Any incorrect or invented spatial relation is a hallucination.
 - Omitting an obvious spatial relation reduces coverage.
 - Stating a relation that is ambiguous or uncertain in the image is also a hallucination.

Evaluation protocol

Start with a brief explanation of your evaluation process. Then assign *one* rating using the scale below. **Output only the rating number—no extra text, symbols, or commentary.**

- 6 Comprehensive coverage, correct spatial layout, *no* hallucinations
- 5 Very informative, correct spatial layout, no hallucinations, minor omissions
- 4 Moderate coverage, correct spatial layout, no hallucinations, several omissions
- 3 Limited coverage, minimal spatial detail, no hallucinations
- 2 Informative but contains at least *one* hallucination (object or spatial)
- 1 Limited coverage *and* at least one hallucination (object or spatial)
- 0 Not informative and/or multiple hallucinations

36

37 A.5 Visual Results By Our Captions

38 We show some qualitative results of our captions of MS-COCO 2014 validation split in Figure5,
39 Figure6, Figure7, Figure8. In each figure, we show the raw images with the generated captions
40 by **VLV**, as well as images generated by SoTA T2I models, such as Midjourney, FLUX.1-dev and
41 ImageFX. Across all cases, our compact *caption embeddings* faithfully encode fine-grained semantics
42 and spatial layout, demonstrating that high descriptive fidelity can be maintained without sacrificing
43 brevity.



Caption: A medium shot captures a black bear standing in a grassy, rocky area. The bear is positioned in the center of the frame, facing slightly to the right, with its head turned towards the right side of the image. The bear's fur is a solid, glossy black, and its ears are small and rounded. It appears to be standing on its hind legs, with its front paws resting on the ground. The ground around the bear is covered with a mix of dry grasses and small, brown rocks. The rocks vary in size and color, ranging from light gray to dark brown. The grasses are a mix of green and brown, with some taller grasses reaching up towards the top of the frame. The lighting in the image appears to be soft and diffused, with no harsh shadows. The colors are muted and natural, with the black of the bear and the browns of the rocks and grasses dominating the scene. The overall mood of the image is peaceful and natural, capturing a moment of stillness in the bear's environment. The bear's posture and expression suggest a sense of calm and observation. The background is slightly blurred, drawing the viewer's attention to the bear in the foreground. The image is well-composed, with a clear focus on the bear and its surroundings. The texture of the bear's fur and the rough surface of the rocks are visible, adding depth and realism to the image.

Figure 5: Demonstrating the Ability to Capture Spatial Layout. The caption shows bear's layout (*in the center of the frame*) in this image as well as the bear's posture (*head turned towards the right side*), showing VLV's ability of capturing spatial layout.



Caption: A medium, eye-level shot depicts a train car with the number "1843" painted on it, positioned on a train track in front of a train station. The train car is predominantly gray with yellow accents on the front and a red stripe running along the side. The number "1843" is prominently displayed on both the front and side of the car. The train track is visible beneath the train car, and the train station is visible in the background. The train station appears to be made of concrete and steel. The sky is overcast, and the lighting is soft and diffused. The overall mood of the image is calm and serene. The train car is the main subject of the image, and it is positioned in the foreground. The train station is visible in the background, and the sky is overcast. The lighting is soft and diffused, and the overall mood of the image is calm and serene.

Figure 6: Demonstrating the Ability of OCR. VLV indeed have the ability of OCR, but in our training dataset, we filter the data by watermark probability less than 0.5. Thus VLV have potentials to improve OCR performance with enough data.



Caption: The image shows a luxurious outdoor bed setup on a sandy beach under a clear blue sky. The bed is elevated on a wooden platform and is covered with a white bedspread. A neatly folded stack of white towels sits on top of the bed. The bed is framed by white curtains that are tied back on either side, creating a sense of enclosure and privacy. The curtains are draped in a way that allows light to filter through, adding a soft glow to the scene. In the background, the turquoise ocean stretches out to the horizon, with several lounge chairs and beach umbrellas scattered along the shoreline. The sand is light-colored and appears to be well-maintained. The lighting is bright and sunny, casting soft shadows and highlighting the textures of the sand, wood, and fabric. The overall mood is serene and luxurious, evoking a sense of relaxation and vacation. The composition is well-balanced, with the bed as the focal point and the beach setting providing a natural and inviting backdrop. There are no people or animals visible in the image. The atmosphere is calm and peaceful, suggesting a perfect day at the beach. The colors are vibrant and natural, with the white of the bed and curtains contrasting against the blue of the sky and ocean. The textures are varied, from the smooth sand to the rough wood of the platform and the soft fabric of the curtains.

Figure 7: **Demonstrating the Ability of Capturing Complex Objects.** Caption enumerates almost every object and correctly describe their spatial relationships, highlighting VLV’s comprehensive scene understanding.



Caption: A medium shot captures a man skiing down a snow-covered slope. He is wearing a red beanie, dark sunglasses, and a light-colored ski jacket with black and red accents, paired with dark gray ski pants. A large black and gray backpack is slung over his shoulders. He is using ski poles to navigate the slope, and his expression is one of concentration and enjoyment. The background features a dense forest of bare trees, suggesting a winter setting. The ground is covered in snow, and the lighting is bright, casting shadows that add depth to the scene. The overall mood is one of outdoor adventure and winter sports. The texture of the snow is visible, adding to the realism of the image. The colors are muted, with the red of the beanie and jacket providing a pop of color against the predominantly white and gray tones of the scene. The man's ski tracks are visible in the snow, indicating his movement. The atmosphere is cold and crisp, typical of a winter day. The image is well-composed, with the man in the foreground and the forest in the background, creating a sense of depth. The lighting is bright and natural, enhancing the colors and textures of the scene. The overall impression is one of winter sports and outdoor adventure. The man's expression is one of enjoyment and concentration, adding to the overall mood of the image.

Figure 8: **Demonstrating the Ability of Capturing Human Posture.** Captions show details of human as well as his posture, demonstrating VLV’s fine-grained posture awareness.

44 **B Dataset License**

45 **B.1 Dataset used in training**

46 **LAION-5B**

47 License: Creative Common CC-BY 4.0 <https://laion.ai/blog/laion-5b/>

48 **B.2 Dataset used in testing**

49 **MS-COCO**

50 License: Creative Common CC-BY 4.0 <https://cocodataset.org/#termsofuse>

51 **VQAv2**

52 License: CC-BY 4.0 <https://visualqa.org/terms.html>

53 Dataset website: <https://visualqa.org/index.html>

54 **OK-VQA**

55 License:N/A.

56 Dataset website: <https://okvqa.allenai.org/>

57 **B.3 Open-sourced models**

58 **stable-diffusion-3.5-medium** (used for image generation).

59 [https://huggingface.co/stabilityai/stable-diffusion-3.5-medium/blob/main/](https://huggingface.co/stabilityai/stable-diffusion-3.5-medium/blob/main/LICENSE.md)
60 LICENSE.md

61 **Qwen-2.5** (used in stage-2 for LLM decoder).

62 <https://huggingface.co/Qwen/Qwen2.5-72B-Instruct/blob/main/LICENSE>

63 **Qwen-2.5-VL** (used in image captioning).

64 <https://github.com/QwenLM/Qwen2.5-VL/blob/main/LICENSE>

65 **Florence-2-Large** (used in image captioning).

66 <https://huggingface.co/microsoft/Florence-2-large/blob/main/LICENSE>

67 **LLaVA-v1.5** (used in image captioning).

68 <https://huggingface.co/liuhaotian/llava-v1.5-7b>

69 **References**

- 70 [1] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi
71 Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*,
72 2024.
- 73 [2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan
74 Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable
75 multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 76 [3] Chen Wei, Chenxi Liu, Siyuan Qiao, Zhishuai Zhang, Alan Yuille, and Jiahui Yu. De-diffusion makes text a
77 strong cross-modal interface. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
78 *Recognition*, pages 13492–13503, 2024.