NoiseStat: A Lightweight Diagnostic Framework for Retrieval Robustness under Noise

Bodhisatta Maiti
Independent Researcher
Orlando, United States
bodhisatta.iitbhu@gmail.com

Debshree Chowdhury
Independent Researcher
Kolkata, India
debshreechowdhury@gmail.com

***Abstract*—The usage of Vision-Language Models (VLMs) is ubiquitous in modern AI systems. VLMs are often evaluated on clean data, whereas real-world data can be noisy and imperfect. This paper introduces a lightweight diagnostic framework to evaluate the retrieval robustness of VLMs under visual and textual noise. 10k image-text pairs were randomly sampled from the Fashion200k dataset, and 13 types of structured noise have been applied to the visual data and 5 types of structured noise have been applied to the textual data. In this study, we evaluate the performance of 5 VLMs in multiple retrieval scenarios: clean-text-to-noisy-image, noisy-image-to-clean-image and noisy-text-to-clean-image under various perturbations. This lightweight diagnostic framework can be utilized in a plug-and-play manner to evaluate the robustness of VLMs.**

***Index Terms*—vision–language models, multimodal retrieval, text-to-image retrieval, image-to-image retrieval, noisy queries, robustness evaluation, e-commerce information retrieval**

I. Introduction

Vision–Language models (VLMs) [1]–[5] tend to show good performance in clean benchmark data sets for tasks such as retrieval and zero-shot classification. However, real-world image and text data can be noisy or degraded. Most often, the evaluations of VLMs are done under ideal conditions making it difficult to gauge the effectiveness of the models under challenging scenarios. The robustness of models under perturbations has been studied extensively for tasks such as classification (e.g. ImageNet-C/ImageNet-P/ImageNet-R [6], [7]) but there are limited studies in the retrieval space. Qiu et al. [8] have proposed a broader evaluation framework incorporating multiple visual and textual perturbations to assess the robustness of image-text models. There remains a need for a lightweight, plug-and-play evaluation framework to evaluate the retrieval performance under structured noise. This paper introduces NoiseStat, a minimalistic diagnostic framework to evaluate the retrieval robustness of VLMs under 13 types of visual noise and 5 types of textual noise. In this study, 10000 image-text pairs were randomly sampled from the Fashion200k [9] dataset and used for the study. The retrieval performance was studied in various scenarios: clean-text-to-noisy-image, noisy-image-to-clean-image, and noisy-text-to-clean-image. The retrieval scenarios for this study have been chosen keeping in mind the e-commerce domain. The clean text to noisy image retrieval emulates those scenarios where users search for products in platform where the item or product photos are uploaded by users which might not be perfect, especially in marketplaces or used product type of platforms. The noisy image to clean image retrieval emulates scenario where the user uploads noisy image and tries to find similar clean catalog images from the e-commerce platform. The noisy text to clean image is the scenario where users accidentally make mistakes in typing their search query especially on mobile e-commerce applications. For all the retrieval scenarios, the recall@k metrics was calculated. The VLMs used of this study are CLIP [1], AltCLIP [2], FashionCLIP [10], SigLIP [4] and SigLIP-2 [5].

II. Methodology

The Fashion200k dataset was used for the study, 10000 image-text pairs were randomly sampled from the dataset. No additional preprocessing was done, the images and text descriptions were used as-is. The dataset was loaded from Hugging Face (Marqo/fashion200k).

A. Visual Perturbations

In this study, 13 structured noise types were applied to the images. There were 10 single noise perturbations (e.g., gaussian noise, motion blur, color jitter, etc.) and 3 tiered noise compositions (tier easy, tier medium, tier hard), which are combinations of 3-4 single noises. Real-world images can contain multiple noise types, to emulate that scenario and also to test the robustness of VLMs under challenging circumstances, a 3-tier system was created. In order to maintain reproducibility, all noises were applied using deterministic parameters (single severity level per noise). This design keeps the setup lightweight and fully reproducible, allowing other researchers to extend the same noise definitions or severity levels without requiring heavy compute resources. Table I lists the 13 noise types and the corresponding parameters. For the tiered noises, the same severity level values of atomic noises were used for consistency. Figure 1 displays a sample original image and the 13 noisy images based on the noise applied. In this paper, the original images are referred to as clean images.

B. Textual Perturbations

In this study, 5 structured noise types were applied to the text data. There were 4 single noise perturbations (typo, deletion, swap, insertion) and 1 tiered noise composition,

TABLE I
DESCRIPTION OF VISUAL NOISE TYPES APPLIED IN THIS STUDY

Noise Type	Method	Key Parameters
Gaussian Noise	Additive white noise	Mean = 0, Std = 25
Salt and Pepper Noise	Random black/white pixel flip	Amount = 0.03
Occlusion	Central black rectangle	Area = 15%
Gaussian Blur	PIL GaussianBlur	$\sigma = 2.0$
Color Jitter	torchvision ColorJitter	Brightness/Contrast/Saturation = 0.8, Hue = 0.1
Motion Blur	Directional kernel via OpenCV	Kernel size = 15, Angle = 0
Perspective Warp	Random inward corner shift	Distortion scale = 0.2
Grayscale	Convert to grayscale and back to RGB	–
Fog Overlay	White image blend simulating fog	Blend $\alpha = 0.4$
Rain Overlay	White pixel noise overlay + Gaussian blur + image blend	Drop prob = 0.3, Blur $\sigma = 1.5$, Blend $\alpha = 0.3$
Tier Easy	Fog + Gaussian Noise + Gaussian Blur + Perspective Warp	Composite
Tier Medium	Color Jitter + Rain + Grayscale	Composite
Tier Hard	Salt-and-Pepper + Motion Blur + Occlusion	Composite

TABLE II
DESCRIPTION OF TEXTUAL PERTURBATION TYPES APPLIED IN THIS STUDY.

Perturbation Type	Method	Key Parameters
Typo (Character-level)	Random character deletion, substitution, insertion, or repetition inside words	∼5 edits per text
Word Deletion	Randomly remove words	∼5 deletions per text
Word Swap	Randomly swap the positions of words	∼5 swaps per text
Extraneous Insertion	Insert irrelevant distractor tokens (“cheap”, “sale”, “free shipping”) into the text	∼5 insertions per text
Tiered Perturbation	Combination of typo, deletion, and swap perturbations	2–3 edits of each type per text

TABLE III
QUALITATIVE EXAMPLE OF TEXTUAL PERTURBATIONS APPLIED TO A PRODUCT DESCRIPTION.

Perturbation Type	Example Text
Original	white silk blouse with a high collar and long sleeves. the blouse has a button-down front and a loose fit. the material is silky and smooth to the touch.
Typo	white silk blouse with a high collar and long sleeves. the blouse has a button-down front aod a loose fit. the material is silky and smoqoth to te touch.
Word Deletion	white silk with a high collar and long sleeves. blouse has button-down front loose fit. the material is silky and smooth to the touch.
Word Swap	white silk blouse with a fit. collar and long the the blouse sleeves. a button-down front and a loose high has the is and silky smooth to material touch.
Extraneous Insertion	white silk blouse with a high collar and long sleeves. the blouse has a best button-down front free shipping and best a loose fit. the material is silky and smooth sale to the touch. sale
Tiered Perturbation	white and blouse wilth a high collar siik loose the has blouse button-down front and a long fit. the material is silky and smoooth to the touch.

which is a combination of all the single noises. The real world can contain multiple noises in a single text, to emulate that scenario and also to test the robustness of VLMs under challenging circumstances, a tiered text noise was created. Table II lists the 5 noise types and the corresponding methods and parameters. Table III displays a sample original text and the 5 noisy texts based on the noise applied. In the paper, the original text will be referred to as clean text.

C. Embedding and Retrieval Model

In this study, five VLMs were used to assess robustness under noise. The VLM's vision encoder was used to encode the clean and noisy images; whereas the clean and noisy text were encoded using the text encoder. All VLMs were used from the Hugging Face platform. The five VLMs and their Hugging Face variants are as follows:

1) CLIP: `openai/clip-vit-base-patch32`
2) AltCLIP: `BAAI/AltCLIP`
3) FashionCLIP: `patrickjohncyh/fashion-clip`
4) SigLIP: `google/siglip-base-patch16-224`
5) SigLIP-2: `google/siglip2-base-patch16-224`

D. Evaluation Setup

The framework evaluates the retrieval performance in multiple scenarios: clean text → noisy image, noisy image → clean image and noisy text → clean image.

Recall@K: for a given input query, the top-k retrieved results were compared with the ground truth data depending on the retrieval scenario. The k values chosen for the study are 1, 5 and 10. This is a basic and fundamental method to evaluate the retrieval results.

In addition to retrieval accuracy, the study also examines how stable the model representations remain under different noise conditions. This was measured using the mean

Fig. 1. Example original image and its 13 perturbed variants using atomic and composite visual noise types.

cosine similarity between clean and perturbed embeddings. The metric provides a simple way to observe whether a model preserves the overall direction of its feature space when exposed to visual or textual distortions. Although Recall@K reflects changes in ranking performance, cosine similarity offers a complementary view of embedding consistency at the representation level.

Together, Recall@K and cosine similarity provide complementary insights: one reflects retrieval accuracy, while the other captures representational stability. This pairing enables a more interpretable understanding of robustness, particularly for models whose embeddings remain locally consistent yet lose alignment under global retrieval ranking.

III. Results

A. Recall@K analysis for clean text → noisy image retrieval

The SigLIP family of models performed the best across all the various noise types. FashionCLIP's performance was better than AltCLIP and CLIP, however, it lagged behind the SigLIP family of models. AltCLIP performed better than CLIP; however, it lagged behind the other models. Across all the models, the atomic noises Salt-and-Pepper, Color Jitter, Grayscale and Motion Blur had the most negative impact on the retrievals. On expected lines, the performance of all the models dropped significantly in the tiered noises, especially the tier hard. Even though SigLIP-2 performed the best on the hard tiered noise, it's recall dropped by 68.17%, 62.16%, and 58.17% at recall levels 1, 5 and 10 respectively. The detailed recall values can be found in Table IV.

B. Recall@K analysis for noisy image → clean image retrieval

SigLIP-2's performance was best across all the noise types and recall levels followed by SigLIP. FashionCLIP's performance was better than AltCLIP and CLIP, however, it lagged behind the SigLIP family of models. AltCLIP performed better than CLIP however, it lagged behind remaining models in the study. The atomic noises Salt-and-Pepper and Motion Blur had the most degradations in the retrieval. The atomic noises Occlusion, Perspective Warp and Fog had the least degradations in the retrieval, with the SigLIP family of models showcasing recall@1 between 0.98-1 for these atomic noises. On expected lines, the recall performance dropped significantly in the tiered noises across all the models. Even though SigLIP-2 performed the best on the hard tiered noise, it's recall dropped by 89.70%, 81.10%, and 75.60% at recall levels 1, 5 and 10 respectively, when compared to the baseline recall of 1.00. The detailed recall values can be found in Table V.

TABLE IV
RETRIEVAL PERFORMANCE (RECALL@K) OF FIVE VLMS FOR THE **CLEAN TEXT → NOISY IMAGE** SCENARIO ACROSS 13 VISUAL NOISE TYPES.

Noise Type	**CLIP**			**AltCLIP**			**FashionCLIP**			**SigLIP**			**SigLIP-2**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Clean	0.087	0.214	0.297	0.163	0.361	0.473	0.242	0.484	0.597	**0.342**	**0.626**	**0.737**	0.333	0.621	0.734
Gaussian Noise	0.060	0.150	0.212	0.118	0.265	0.349	0.152	0.322	0.409	**0.288**	**0.547**	**0.656**	0.279	0.545	0.655
Salt-and-Pepper	0.038	0.102	0.147	0.079	0.192	0.262	0.111	0.240	0.315	**0.218**	**0.436**	**0.539**	0.206	0.424	0.529
Occlusion	0.061	0.166	0.231	0.120	0.280	0.382	0.177	0.368	0.468	**0.327**	**0.610**	**0.719**	0.319	0.599	0.716
Gaussian Blur	0.067	0.168	0.234	0.117	0.272	0.361	0.182	0.392	0.489	**0.282**	0.539	0.650	0.274	**0.546**	**0.662**
Color Jitter	0.044	0.117	0.171	0.094	0.223	0.299	0.120	0.271	0.352	**0.222**	0.446	0.549	0.220	**0.448**	**0.558**
Grayscale	0.036	0.101	0.146	0.077	0.185	0.262	0.113	0.258	0.348	**0.189**	0.408	0.521	0.187	**0.410**	**0.528**
Motion Blur	0.047	0.115	0.164	0.086	0.202	0.274	0.126	0.280	0.368	0.211	0.440	0.550	**0.212**	**0.442**	**0.556**
Perspective Warp	0.079	0.208	0.286	0.158	0.354	0.463	0.227	0.465	0.572	**0.334**	**0.620**	**0.730**	0.321	0.599	0.713
Rain (Weather)	0.054	0.141	0.197	0.117	0.268	0.353	0.161	0.335	0.426	**0.276**	**0.535**	**0.642**	0.254	0.505	0.615
Fog (Weather)	0.063	0.161	0.224	0.129	0.291	0.377	0.184	0.387	0.492	0.292	0.563	0.669	**0.293**	**0.567**	**0.677**
Tiered Noise (Easy)	0.033	0.089	0.129	0.065	0.151	0.208	0.086	0.193	0.254	0.166	0.344	0.437	**0.168**	**0.352**	**0.444**
Tiered Noise (Med.)	0.014	0.038	0.058	0.033	0.083	0.116	0.042	0.102	0.139	**0.087**	**0.203**	**0.270**	**0.087**	0.199	0.267
Tiered Noise (Hard)	0.011	0.036	0.056	0.031	0.077	0.110	0.026	0.061	0.084	0.095	0.214	0.283	**0.106**	**0.235**	**0.307**

TABLE V
RETRIEVAL PERFORMANCE (RECALL@K) OF FIVE VLMS FOR THE **NOISY IMAGE → CLEAN IMAGE** SCENARIO ACROSS 13 VISUAL NOISE TYPES.

Noise Type	**CLIP**			**AltCLIP**			**FashionCLIP**			**SigLIP**			**SigLIP-2**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Gaussian Noise	0.289	0.394	0.441	0.393	0.526	0.585	0.688	0.803	0.845	0.778	0.893	0.923	**0.840**	**0.939**	**0.960**
Salt-and-Pepper	0.176	0.274	0.318	0.251	0.376	0.434	0.387	0.522	0.578	0.602	0.753	0.802	**0.703**	**0.862**	**0.905**
Occlusion	0.790	0.890	0.913	0.857	0.917	0.933	0.919	0.952	0.962	0.995	**0.999**	**0.999**	**0.997**	**0.999**	**0.999**
Gaussian Blur	0.602	0.745	0.793	0.639	0.764	0.808	0.861	0.943	0.965	0.940	0.983	0.988	**0.975**	**0.996**	**0.998**
Color Jitter	0.480	0.588	0.633	0.560	0.660	0.697	0.747	0.820	0.843	0.786	0.867	0.889	**0.856**	**0.914**	**0.927**
Grayscale	0.677	0.825	0.870	0.695	0.828	0.871	0.873	0.948	0.965	0.969	0.992	0.996	**0.984**	**0.999**	**0.999**
Motion Blur	0.235	0.331	0.375	0.263	0.363	0.407	0.541	0.671	0.723	0.679	0.819	0.860	**0.735**	**0.866**	**0.903**
Perspective Warp	0.853	0.937	0.957	0.897	0.957	0.970	0.968	0.994	0.997	0.979	0.996	0.997	**0.991**	**0.999**	**0.999**
Rain (Weather)	0.379	0.490	0.536	0.498	0.628	0.672	0.729	0.835	0.873	0.809	0.907	0.931	**0.871**	**0.952**	**0.968**
Fog (Weather)	0.915	0.972	0.983	0.925	0.974	0.983	0.986	0.998	0.999	0.989	0.998	0.999	**0.998**	**1.000**	**1.000**
Tiered Noise (Easy)	0.036	0.070	0.092	0.065	0.110	0.137	0.146	0.248	0.295	0.246	0.394	0.462	**0.270**	**0.442**	**0.525**
Tiered Noise (Med.)	0.019	0.039	0.052	0.036	0.067	0.084	0.146	0.226	0.264	0.216	0.331	0.381	**0.255**	**0.373**	**0.424**
Tiered Noise (Hard)	0.006	0.014	0.021	0.012	0.027	0.038	0.024	0.042	0.055	0.056	0.115	0.150	**0.103**	**0.189**	**0.244**

C. Recall@K analysis for noisy text → clean image retrieval

SigLIP performed the best in majority of the noise types and recall levels closely followed by SigLIP-2. FashionCLIP's performance was better than AltCLIP and CLIP, however, the performance lagged behind the SigLIP family of models. AltCLIP performed better than CLIP, however, it's performance lagged behind the remaining models. The atomic noises didn't have any significant impact on the retrieval results when compared to the baseline. For example, in the atomic noise typo, SigLIP's recall@1 dropped by 11.40%, whereas CLIP's recall@1 dropped by 17.24%. In the tiered noise, we can see the negative impact on the retrieval. For example, SiGLIP's performance dropped by 18.13%, 14.22% and 12.08% are recall levels 1, 5 and 10 respectively. The detailed recall values can be found in Table VI.

D. Embedding Stability Analysis

To better understand how noise affects the underlying representations, mean cosine similarity was computed between clean and noisy embeddings for both image and text inputs (Tables VII and VIII). Higher values indicate that a model's embeddings remain locally consistent under perturbation, although such stability does not always translate to stronger retrieval performance. CLIP, for instance, preserves relatively high cosine similarity across most visual noise types yet shows the largest drop in Recall@K, suggesting that its representations remain directionally stable but lose their global alignment. In contrast, the SigLIP variants maintain comparable or slightly lower similarity values while exhibiting steadier retrieval accuracy, implying that they deform more coherently in the latent space. Text perturbations resulted in smaller changes overall, indicating that the language encoders were less sensitive to surface-level distortions such as typos, insertions, or word swaps. These findings underline that stability and robustness are not strictly correlated—high local similarity does not always ensure resilient retrieval rankings, highlighting the importance of analyzing both perspectives.

IV. FUTURE WORK

The current study was performed at a single noise severity level for both images and text, the future work will involve expanding the severity levels and evaluating the robustness of VLMs across multiple severity levels. Additionally the framework will be expanded to study the impact of image noises on the performance of generative VLMs (e.g., BLIP models [11], [12], LLaVA models [13], [14], Qwen-VL models [15], [16], etc.), especially on tasks such as captioning and question-answering. FashionCLIP consistently performed better than CLIP and AltCLIP. The next step would be to

TABLE VI
RETRIEVAL PERFORMANCE (RECALL@K) OF FIVE VLMS FOR THE **NOISY TEXT → CLEAN IMAGE** SCENARIO ACROSS 5 TEXTUAL PERTURBATION TYPES.

Text Noise Type	CLIP			AltCLIP			FashionCLIP			SigLIP			SigLIP-2		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Clean	0.087	0.214	0.297	0.163	0.361	0.473	0.242	0.484	0.597	**0.342**	**0.626**	**0.737**	0.333	0.621	0.734
Typo	0.072	0.186	0.264	0.145	0.326	0.432	0.206	0.426	0.536	**0.303**	**0.575**	**0.689**	0.282	0.557	0.673
Deletion	0.073	0.189	0.266	0.140	0.315	0.416	0.202	0.423	0.524	**0.292**	0.552	0.665	0.289	**0.553**	**0.667**
Swap	0.077	0.195	0.278	0.146	0.333	0.439	0.220	0.453	0.561	**0.313**	**0.581**	**0.697**	0.291	0.566	0.680
Insertion	0.081	0.204	0.285	0.141	0.324	0.430	0.225	0.449	0.559	**0.322**	**0.604**	**0.714**	0.287	0.567	0.681
Tiered	0.068	0.178	0.248	0.130	0.301	0.400	0.192	0.404	0.509	**0.280**	**0.537**	**0.648**	0.265	0.525	0.642

TABLE VII
MEAN COSINE SIMILARITY BETWEEN CLEAN AND NOISY **IMAGE** EMBEDDINGS FOR EACH VISUAL PERTURBATION.

Noise Type	CLIP	AltCLIP	FashionCLIP	SigLIP	SigLIP-2
Gaussian Noise	0.856	0.878	0.823	0.857	0.881
Salt & Pepper	0.815	0.823	0.742	0.810	0.836
Occlusion	0.913	0.897	0.874	0.943	0.952
Gaussian Blur	0.924	0.913	0.886	0.881	0.920
Color Jitter	0.887	0.893	0.842	0.860	0.894
Grayscale	0.934	0.925	0.886	0.913	0.933
Motion Blur	0.842	0.834	0.804	0.834	0.866
Perspective Warp	0.949	0.946	0.897	0.928	0.947
Rain	0.873	0.892	0.835	0.850	0.872
Fog	0.958	0.947	0.939	0.929	0.953
Tiered (Easy)	0.789	0.782	0.666	0.750	0.783
Tiered (Medium)	0.713	0.741	0.643	0.703	0.730
Tiered (Hard)	0.682	0.707	0.509	0.660	0.713

TABLE VIII
MEAN COSINE SIMILARITY BETWEEN CLEAN AND NOISY **TEXT** EMBEDDINGS FOR EACH TEXTUAL PERTURBATION.

Text Noise Type	CLIP	AltCLIP	FashionCLIP	SigLIP	SigLIP-2
Typo	0.946	0.963	0.957	0.925	0.938
Deletion	0.949	0.960	0.958	0.933	0.940
Swap	0.946	0.965	0.966	0.936	0.948
Insertion	0.934	0.968	0.940	0.918	0.929
Tiered	0.920	0.942	0.939	0.904	0.925

fine-tune the SigLIP family of models on fashion data and evaluate whether the fine tuned SigLIP models can outperform the current results. While the current study has incorporated embedding stability analysis using cosine similarity, future work will explore additional representation-level diagnostics to better characterize how models behave under varying noise intensities. In the long term, the objective is to release the framework as a lightweight Python package that can be used in a plug-and-play manner to assess the robustness of retrieval systems under noise. In essence, NoiseStat serves as an initial step toward a standardized, transparent benchmark for retrieval robustness—its lightweight nature encourages broader adoption and community validation.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[2] Z. Chen, G. Liu, B. Zhang, F. Ye, Q. Yang, and L. Wu, "AltCLIP: Altering the language encoder in CLIP for extended language capabilities," arXiv preprint arXiv:2211.06679, 2022.

[3] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, 2021, pp. 4904–4916.

[4] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 11975–11986.

[5] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, *et al.*, "SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," arXiv preprint arXiv:2502.14786, 2025.

[6] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," arXiv preprint arXiv:1903.12261, 2019.

[7] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8340–8349.

[8] J. Qiu, Y. Zhu, X. Shi, F. Wenzel, Z. Tang, D. Zhao, B. Li, and M. Li, "Benchmarking robustness of multimodal image-text models under distribution shift," *J. Data-centric Mach. Learn. Res.*, 2023.

[9] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, "Automatic spatially-aware fashion concept discovery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1463–1471.

[10] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue, "Contrastive language and vision learning of general fashion concepts," *Sci. Rep.*, vol. 12, no. 1, Nov. 2022, Art. no. 18952, doi: 10.1038/s41598-022-23052-9.

[11] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, Baltimore, MD, USA, 2022, pp. 12888–12900.

[12] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, Honolulu, HI, USA, 2023, pp. 19730–19742.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," arXiv preprint arXiv:2304.08485, 2023.

[14] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," arXiv preprint arXiv:2310.03744, 2023.

[15] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," arXiv preprint arXiv:2308.12966, 2023.

[16] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," arXiv preprint arXiv:2409.12191, 2024.