Supplementary Materials:
FiVA: Fine-grained Visual Attribute Dataset for Text-to-Image Diffusion Models

In the supplementary material, we include links to the dataset, metadata, and documentation in Section A. We then introduce additional details on dataset construction in Section B. Further, we present more details on the experimental setup and additional experimental results in Section C. Finally, we discuss the limitations and future work of the project in Section D. Please also find the datasheet for the dataset in Section E.

A Dataset Information

A.1 Dataset Link and Documentation

Our dataset, metadata, and its license are currently maintained on huggingface [1] for users to download: `https://huggingface.co/datasets/FiVA/FiVA`. It contains the generated images and their metadata, the the original taxonomy of visual attributes and subjects to create the prompts, and the data filtering file. For each of the images, the main visual attribute type, keyword, subject, and prompt is stored in the metadata. A detailed documentation of dataset structure and usage as well as an example of the metadata can be found in the dataset card via the URL above. The Croissant link can be find here `https://huggingface.co/api/datasets/FiVA/FiVA/croissant`.

A.2 Author Statement and Data License

The authors bear all responsibility in case of violation of rights and confirm that this dataset is open-sourced under the Playground v2.5 Community License license.

B Additional Details on Dataset Construction

Details on attribute taxonomy and statistics. When constructing the attribute library, for `color`, `lighting`, `dynamics`, `artistic stroke`, and `focus and depth of field`, we create a list of short descriptions or keywords for each kind of subcategory together with a list of major subjects that can fit into the description. When constructing the prompt, we simply link the attribute and the subject with a comma. For two specific attribute types, namely `rhythm` and `design`, the visual results can hardly be presented simply via short descriptions or keywords. We use long descriptions with "[sks]" denoting the placeholder for subjects that might fit into the sentence. Prompts are created by replacing "[sks]" with each of the subject candidates. We show the visualization of a rough distribution of attributes and subjects in Figure S2, as well as an example of constructing a pair of images that share similar lighting conditions. We also show some more examples of images with different visual attributes in Figure S1.

[1] https://huggingface.co/

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

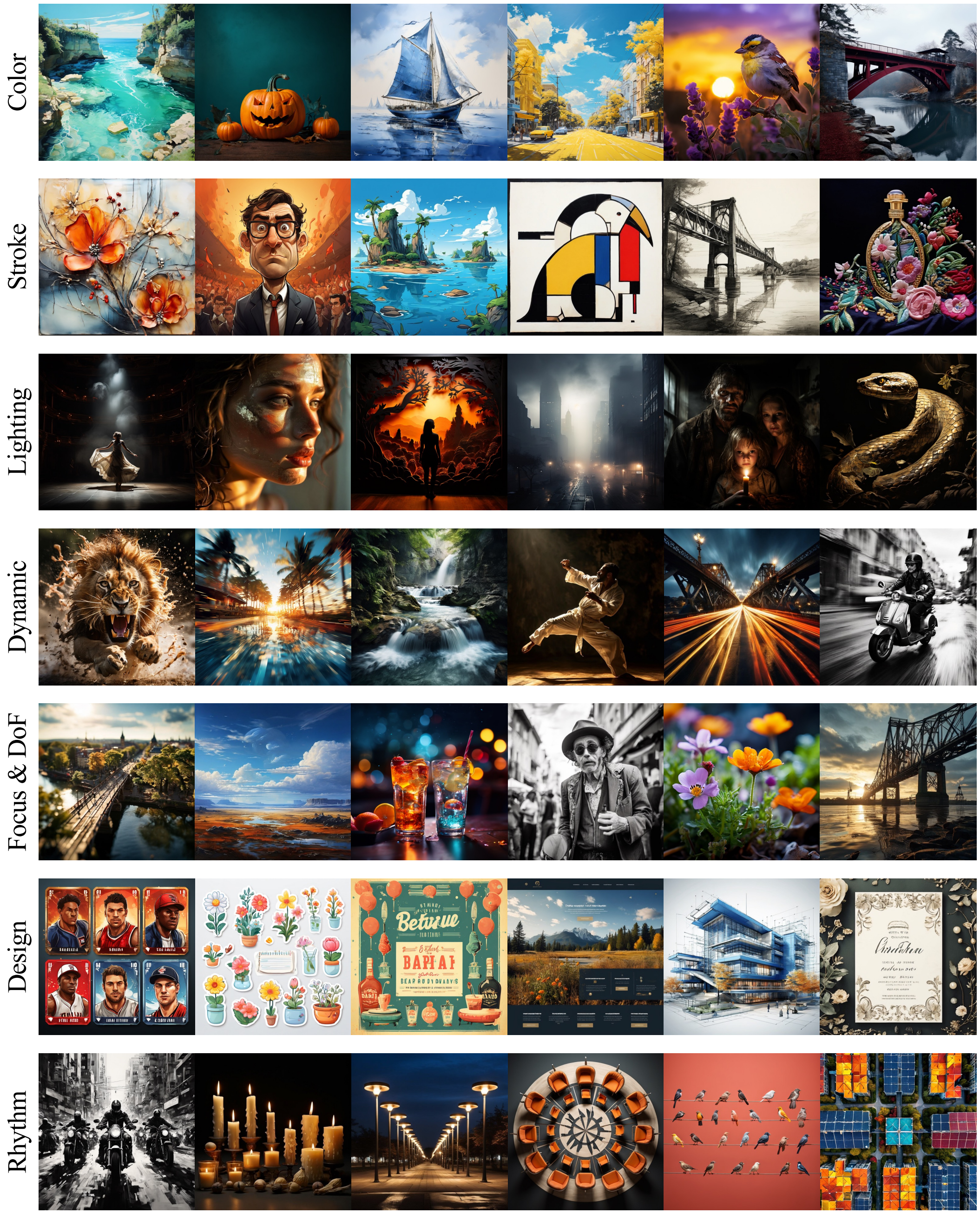


Figure S1: **More image examples with different visual attributes.**

Details on the Range-sensitive Data Filtering. To achieve attribute-consistent image pairs, we need to establish a set of ranges for each attribute where any two images maintain consistency. We organize images into a hierarchy of `Set/Major-subject/Sub-subject`, with the largest set being the aforementioned "group of suitable subjects." Figure S3a shows an example of the hierarchical structure of images related to the attribute 'lighting: moonlight' featuring 7 major-subjects and over 100 sub-subjects. Within this hierarchy, each sub-subject corresponds to a list of images, where each image belongs to that sub-subject and possesses the visual attribute of 'lighting: moonlight'.

We apply *Range-sensitive Data Filtering* to this hierarchy: We first validate the consistency within each specific `Major-subject`. Subsequently, we validate the `Set` encompassing all validated major-subjects. For any major-subject that failed to pass the validation, we then check their `Sub-subjects`.

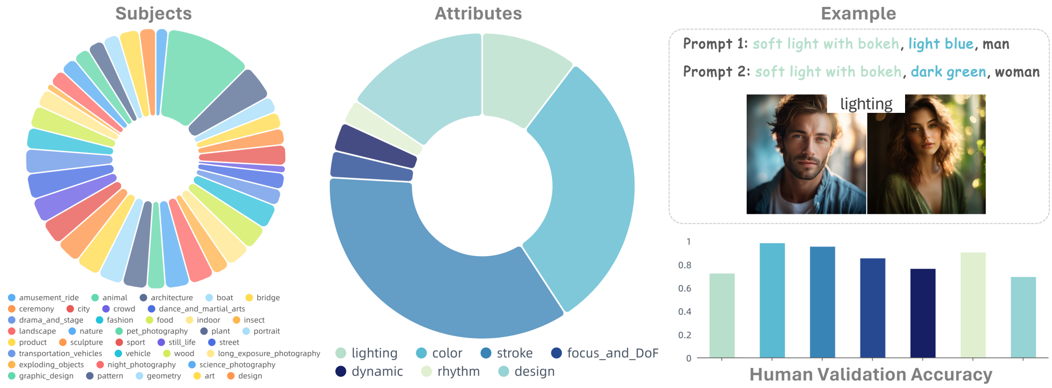
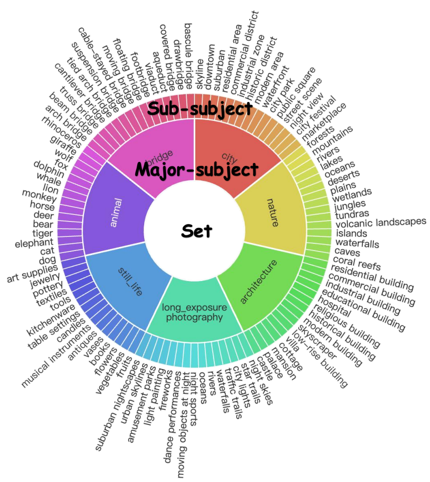
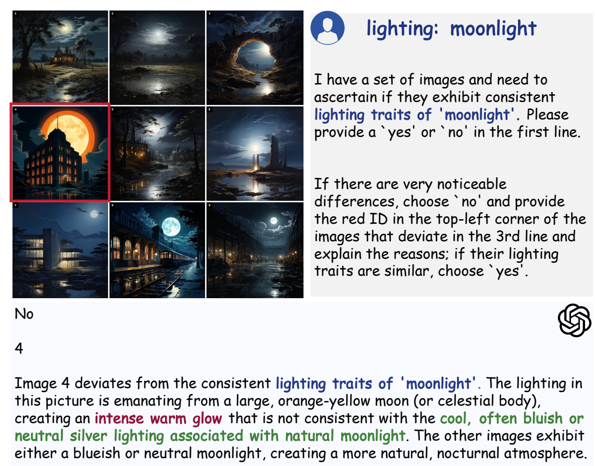


Figure S2: **Statistics and Analysis.** We visualize the rough distribution of visual attributes and subjects on the left. On the right, we show an example pair of images that shares similar lighting condition. We also visualize the attribute alignment accuracy via human validation here.



(a) **Hierarchy of Subject Tree.**



(b) **GPT4V based Range-sensitive Data Filtering.**

Figure S3: **Range-sensitive Data Filtering.** Taking the attribute *lighting: moonlight* as an example, **(a)** demonstrates the hierarchy of `Set/Major-subject/Sub-subject`. It lists the "group of suitable subjects" chosen when generating images related to *lighting: moonlight*, along with sub-subjects under each major-subject. Due to space limitations, only 15 sub-subjects are listed for each major-subject. **(b)** verifies whether the images under *major-subject: architecture* exhibit consistent *lighting* traits of *moonlight*. The result shows that Image 4 exhibits inconsistencies, with the reasons provided.

As shown in Figure S3b, from the range we want to verify, we sample 9 images and arrange them in a grid. Using GPT-4V, we assessed image consistency for a specific visual attribute. In our example, <major attribute> is *lighting*, and <specific attribute> is *moonlight*. For each range we want to verify, this sampling is repeated multiple times. If the mean proportion of inconsistent images remains below a predefined threshold of 0.1, we consider the images consistent for the selected visual attribute within the specified range.

Details on the Human Validation. For human validation, we clarify that there are 1,400 images in total, with 200 images for each attribute. These 200 images are randomly paired based on the same attribute description. Human experts are instructed to judge whether the paired images share similar visual attributes according to the specific attribute type. Notably, we do not require each image to be highly aligned with the text prompt that created it; we only seek visual similarity between the paired images. The accuracy for each attribute is visualized in Figure S2.

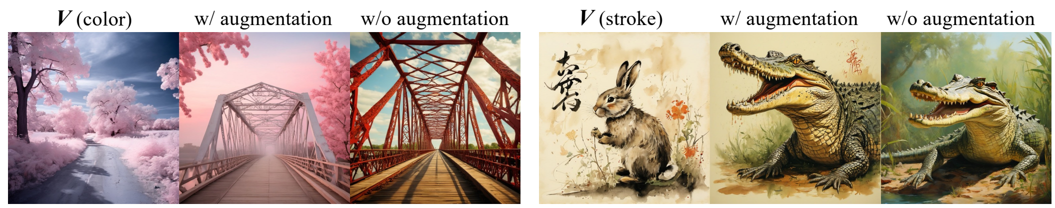


Figure S4: **Ablation on attribute input augmentation.** Models trained with tag augmentation handle slight deviations in input text during inference, while those without augmentation would fail in these cases.

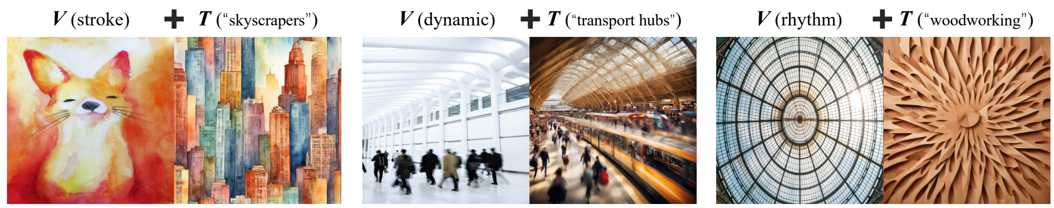


Figure S5: **Examples with real-world images.** We demonstrate that our adapter can be effectively extended to real-world images, which have a different distribution from generated images.

C Additional Experimental Details

C.1 Details on Experimental Setup

Implementation Details. For our methods, our framework's training and inference setting are similar to the IP-Adapter [5]. The learning rate is set to 2e-5, and weight decay is set to 1e-3 for stabilizing the training. The Q-former, channel projector, and multi-image cross-attention are trained, and other parameters are frozen. The training images are resized to 512×512, and the inference resolution is 1024×1024. The model is trained for three epochs with the randomly shuffled training dataset. For each target image, the attribute images are randomly sampled.

For the baseline methods, we adopt the official code base and hyper-parameters for IP-Adapter [5], DEADiff [3], and Style-Aligned [1], and we use the implementation in diffusers [2] for Dreambooth-Lora [4] with only the reference image as training source.

Details on Evaluation The validation set for the user study contains 100 reference images with different visual attribute types. The distribution of the validation set reflects the inherent diversity of each attribute. We involve three times more data for the GPT study under the same distribution, thanks to its ability to scale up.

For the CLIP-Score, we use `ViT-L-14` model, and report the cosine similarity between the text feature of the target subject and the image feature of the generated image. For the user study, we send questionnaires to 30 volunteers with randomly shuffled image options. We are using the `gpt-4-turbo-2024-04-09` model for GPT-4V API inference. Detailed instructions for GPT-4V can be find in Figure S6.

C.2 More Results

GPT Study Results Multi-modal Large Language Models (*e.g.*, GPT-4V(ision)) can offer a more scalable alternative to user studies, providing comprehensive analysis and judgment simultaneously. Specifically, we instruct the GPT-4V model to complete similar questionnaires as in the user study. An example of the instruction and GPT's output can be found in Figure S6. The GPT study results, shown in Table R1, demonstrate that our method outperforms the baselines in most attributes. However, the

[2]https://github.com/huggingface/diffusers

Table R1: **GPT study results on each attribute type.** The Attr&Sub-Acc here denotes the accuracy when both the attribute transferring and target subject are correct.

Methods	Attr&Sub-Acc							
	Color	Stroke	Lighting	Focus&DoF	Dynamic	Design	Rhythm	**Average**
DB-Lora	0.516	0.478	0.358	0.485	0.480	<u>0.600</u>	**0.607**	0.503
IP-Adapter	0.323	0.403	0.340	0.364	0.520	0.440	0.500	0.413
DEADiff	0.161	0.209	0.245	0.485	0.400	0.080	0.357	0.277
Style-Aligned	<u>0.581</u>	<u>0.552</u>	**0.396**	<u>0.606</u>	**0.600**	**0.660**	<u>0.571</u>	<u>0.567</u>
Ours	**0.780**	**0.647**	**0.396**	**0.727**	<u>0.560</u>	0.510	0.521	**0.592**

results for `Design` and `Rhythm` are not as strong, possibly due to the relatively small data scale for these two attributes.

Effect of the input attribute augmentation. During inference, users may present visual information in various ways. For example, "color" might be referred to as "hue" or "palette," and "dynamic" as "motion capture" or "action shot." Therefore, we add attribute name augmentation during Q-former training to accommodate diverse user inputs. As shown in Figure S4, when the input text slightly differs from the standard attribute names during inference, models trained with tag augmentation can still perform effectively, whereas those without augmentation fail to do so.

Results on real-world data. We show the generalization ability of the model to some real-world data collected from Unsplash [3] to verify the model's generation ability to some attributes beyond the training set. Results in Figure S5 shows that our adapter can be effectively extended to real-world images, which have a different distribution than generated images.

D Limitations and Future Works

The main limitation of the dataset is its heavy reliance on the capacity of the generative model, which might constrain the realism, range of available visual attributes, and attribute accuracy between paired data. For example, specific attributes like photographic composition techniques or creative photography can hardly be created in this way. This might also introduce some bias in appearance distribution introduced by the generative model. In the future, we will consider collecting some high-quality data from platforms with professional photographers and designers, and involve human annotation to create paired data, which can further enhance the dataset with a more realistic data distribution and more complex visual attributes.

E Datasheet for Datasets

The following section contains answers to questions listed in datasheets for the dataset.

E.1 Motivation

- For what purpose was the dataset created?
  The FiVA dataset is designed to facilitate research in fine-grained visual attributes, enabling user-friendly customization. It allows users to selectively apply desired attributes to create images that match their unique preferences and specific content requirements.

- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  The authors of this paper.

[3] https://unsplash.com/

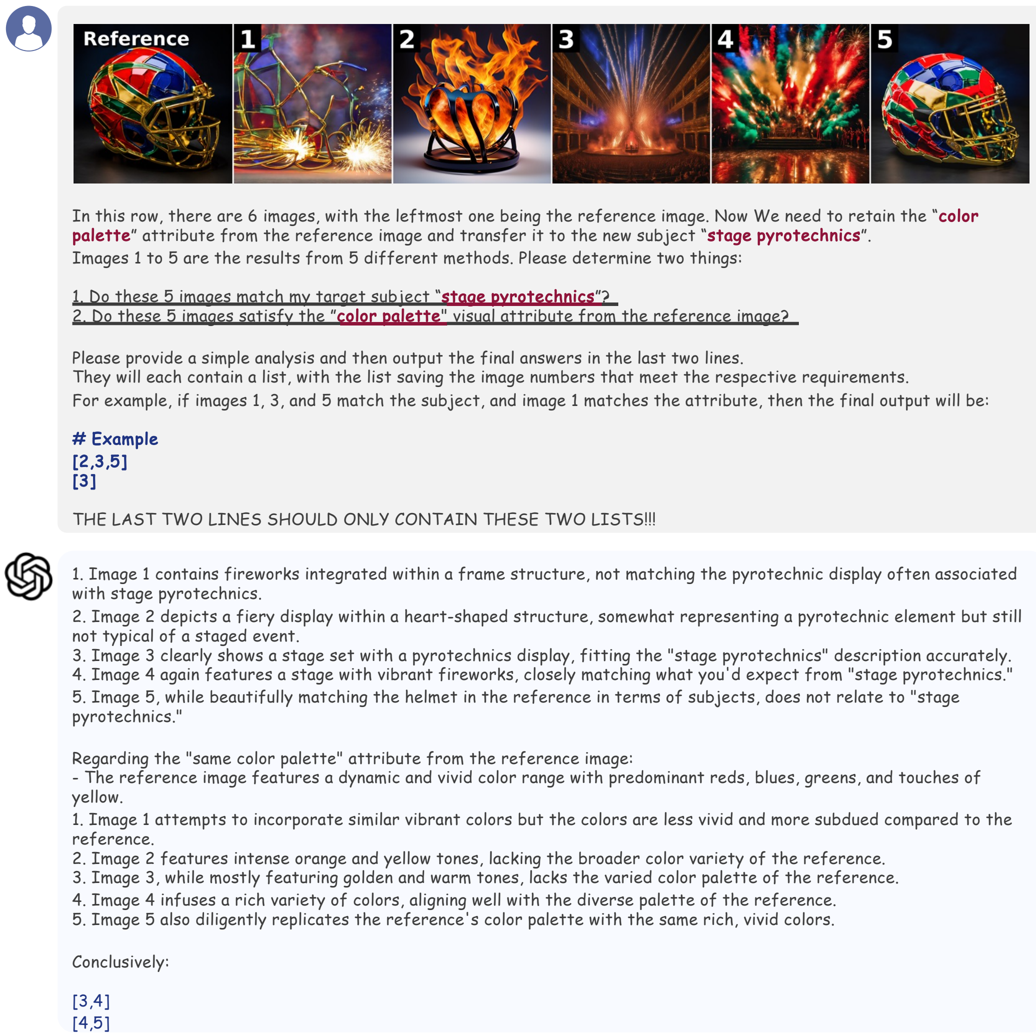


Figure S6: **An example of the input, instruction, and output of the GPT study.** GPT-4V shows sufficient ability in understanding the problem and providing comprehensive analysis and judgement to these questions that can hardly be evaluated by other pre-trained models.

- Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
  The Chinese University of Hong Kong and Shanghai AI Lab supported this work.

E.2 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  The FiVA dataset consists of a number of pairs of images that share similar visual attributes and corresponding meta data like attribute type and subject.
- How many instances are there in total (of each type, if appropriate)?
  The FiVA dataset contains 700K images generated by Playground-V2.5.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  The FiVA dataset is a new dataset generated using existing 2D generative models.

- What data does each instance consist of?
  Each instance contains an image with a prominent visual feature, such as color, stroke, lighting, and so on.
- Is there a label or target associated with each instance?
  Yes.
- Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  N/A.
- Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
  N/A.
- Are there recommended data splits (e.g., training, development/validation, testing)?
  Yes. We provide a small subset for validation.
- Are there any errors, sources of noise, or redundancies in the dataset?
  Yes.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
  The dataset is self-contained.
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?
  N/A.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  N/A.
- Does the dataset relate to people?
  Yes.
- Does the dataset identify any subpopulations (e.g., by age, gender)?
  N/A.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
  N/A.
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
  N/A.

E.3 Collection Process

- How was the data associated with each instance acquired?
  We used the open-source 2D generative model, Playground-V2.5 [2] to generate the dataset.
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?
  We develop an attribute library and subject tree to create the prompts, generate the images, and develop a range-sensitive filtering to enhance the pair-wise attribute alignment. We also perform human validation to verify the accuracy of the attribute alignment.
- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
  N/A.

- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
  The authors of the paper participated in the data collection and verification process.
- Over what timeframe was the data collected?
  The data was collected during April and May of 2024.
- Were any ethical review processes conducted (e.g., by an institutional review board)?
  N/A.
- Does the dataset relate to people?
  Yes.
- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
  We generated the image data.
- Were the individuals in question notified about the data collection?
  The data is not collected from individuals.
- Did the individuals in question consent to the collection and use of their data?
  The data is not collected from individuals.
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
  N/A.
- Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
  Yes.

E.4 Preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?
  Yes. We provide a data filter.
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  Yes.
- Is the software that was used to preprocess/clean/label the data available?
  Yes, we use Python to preprocess/clean/label the data.

E.5 Uses

- Has the dataset been used for any tasks already?
  Yes, for customized image generation.
- Is there a repository that links to any or all papers or systems that use the dataset?
  No.
- What (other) tasks could the dataset be used for?
  High-level perception tasks like aesthetic analysis.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  N/A.
- Are there tasks for which the dataset should not be used?
  N/A.

E.6 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  No.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  The dataset are released on Huggingface: https://huggingface.co/datasets/FiVA/FiVA/.
- When will the dataset be distributed?
  The dataset will be gradually released starting from June 2024. Due to its large scale, it will take some time for the dataset to be fully released, considering the uploading speed.
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  The dataset will be released under the Playground v2.5 Community License license.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  No.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
  No.

E.7 Maintenance

- Who will be supporting/hosting/maintaining the dataset?
  The authors of this paper.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  Please contact the first author of the paper.
- Is there an erratum?
  No.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  Yes.
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?
  N/A
- Will older versions of the dataset continue to be supported/hosted/maintained?
  Yes.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  Please contact the authors of the paper.

References

[1] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *ArXiv*, abs/2312.02133, 2023.

[2] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation, 2024.

[3] Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. 2024.

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[5] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.