How to Train your Text-to-Image Model: Evaluating Design Choices for Synthetic Training Captions

Supplementary Material

A. Further Experimental Details

We here provide further experimental details not eluded to in the main body of the paper.

A.1. Synthetic Caption Sampling

For the Llava-Next models, we used the sglang [44] as the inference framework, whereas we relied on lmdeploy [6]for InternVL2-LLama3-76B. Unless expressly stated otherwise, we used the default sampling hyperparameters at a batch size of 128 and image input resolution of 256p. Through careful prompt engineering, we decided to use the following captioning prompt.

```
Give a description of the objects
and scene in the image as if the
description can be used to prompt
a text-to-image generator model to
generate images.  Do not start the
description with "The image".  Do not
exceed <X> words.
```

To generate long, dense captions, for example, we would set `<X>` to 60.

A.2. Training Framework

We used an adapted version of the `accelerate` training script for Stable Diffusion provided by HuggingFace. The script is readily available online as part of the `diffusers` library[5].

B. Image Resolution

To reduce compute requirements when generating textual descriptions, Practitioners will often input downsampled images to the VLM. In Fig. 10, we compare the downstream performance of models whose training captions only differ in the image resolution used at caption generation. Generating the captions on full-resolution images yields slightly better T2I models. However, the differences are small and within the calculated confidence intervals. Consequently, using lower-resolution images during captioning is a valid strategy to reduce computational requirements with only minimal tradeoffs. For all subsequent experiments, we, therefore, use a slightly reduced resolution of 256p when generating captions.

Caption Setup	VQA-Score
Original LAION-2B	$0.6022^{\pm 0.212}$
InternVL-2 75B (60)	$\mathbf{0.6581}^{\pm 0.213}$
InternVL-2 75B (Fixed Random)	$0.6502^{\pm 0.211}$
InternVL-2 75B (Diverse Random)	$0.6551^{\pm 0.209}$
InternVL-2 75B (Fixed Random Persona)	$0.6311^{\pm 0.212}$

Table 1. Evaluation on Gen-AI Bench supports the general findings from the main paper. Superscript numbers indicate the standard deviation. Colors corresponding to the ones in the main body of the paper are included for convenience.

C. GenAI-Bench Results

In addition to the evaluation of our own benchmark using pick-score, we corroborate our main findings using GenAI-Bench with the VQA-Score metric [22]. Furthermore, we here report the *absolute* scores instead of the difference to the base model. Thus showcasing that our continual learning setup does not confound the analysis conducted. We show the results of our evaluation in Tab. 1.

From these results, we can draw the same conclusions as those in the main body of the paper.

1. Training on long, dense captions results in better prompt following of the downstream text-to-image model
2. Randomizing training caption length does not adversely affect text alignment while providing the benefits outlined in Sec. 4.2
3. We observe no measurable benefit in additionally diversifying captions across epochs
4. Attempting to make caption diversity explicit through personas results in a performance drop

D. Inter Epoch Diversity

In addition to the experiment discussed in Sec. 4.2, we conducted a more rigorous evaluation of inter-epoch diversity for data-constrained scenarios.

Experimental Setup For this set of experiments, we assume a fixed training budget of 1M samples. Consequently, training one epoch on our established dataset serves as the baseline for downstream text-to-image performance.

Next, we artificially decrease the number of available images for training by subsampling from the original training

[5] https://github.com/huggingface/diffusers/blob/main/examples/text_to_image/train_text_to_image.py

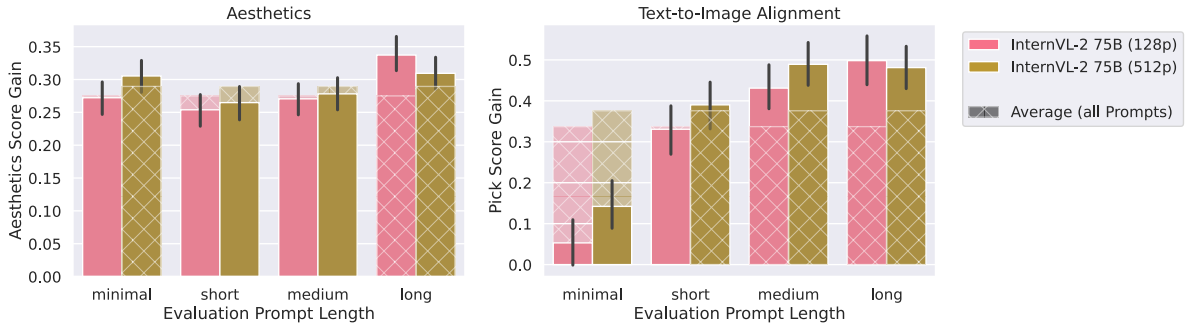


Figure 10. Comparison of training on captions sampled for low or high-resolution images. Decreasing image resolution for the VLM captioning can be a worthwhile trade-off. Scores are shown as the improvement over the SDv1.1 base model.

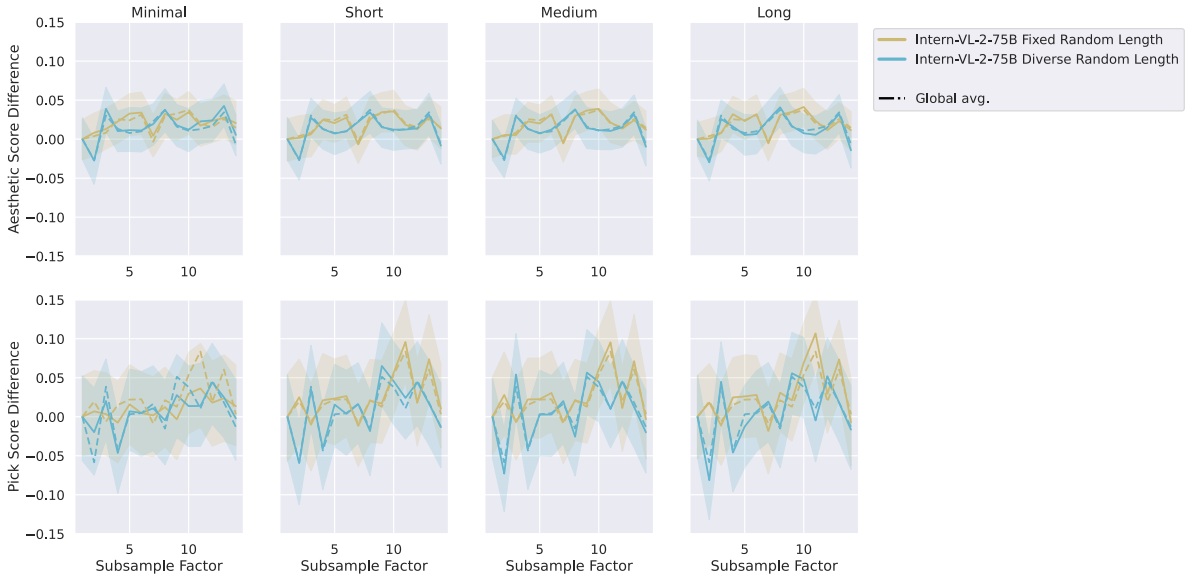


Figure 11. Comparison of training on diversifying captions over epochs in data-constrained environments. Subsampling factor refers to the subset size of training images and respective epochs. Scores are depicted as difference to the model trained on all 1M images.

set. For example, at a sub-sampling factor of 2x, we sample 500k images on which we would train for two epochs. Consequently, the total number of training steps remains the same.

To ensure that each subsample is representative of the entire training set, we chose not to draw samples randomly. Instead, we build roughly 1.4k image clusters using SigLIP embeddings [42] of all 1M images. We then sample from each cluster based on its entropy.

We compare two approaches to multi-epoch training for subsampling factors from 2x to 15x. The baseline approach always uses the same random-length caption for each image, whereas the diverse setting shows a different caption at each epoch.

Results We depict the results in Fig. 11. First, we observe that text-to-image training is surprisingly stable in this setting. Even when training on only 66k text-image pairs for 15 epochs, the downstream performance remains similar to training on the full set of 1M images. Consequently, we found no benefit in using a different caption at each epoch.

In fact, we only start to see deterioration in text-to-image alignment when reducing the number of samples below 5000 pairs (cf. Fig. 12). For image aesthetics, the model still produces comparable outputs when trained on only 2000 samples. We argue that the robustness in training can largely be attributed to the fact that we sampled from semantic image clusters. These results further strengthen the recommendation from the main paper regarding training data diversity.

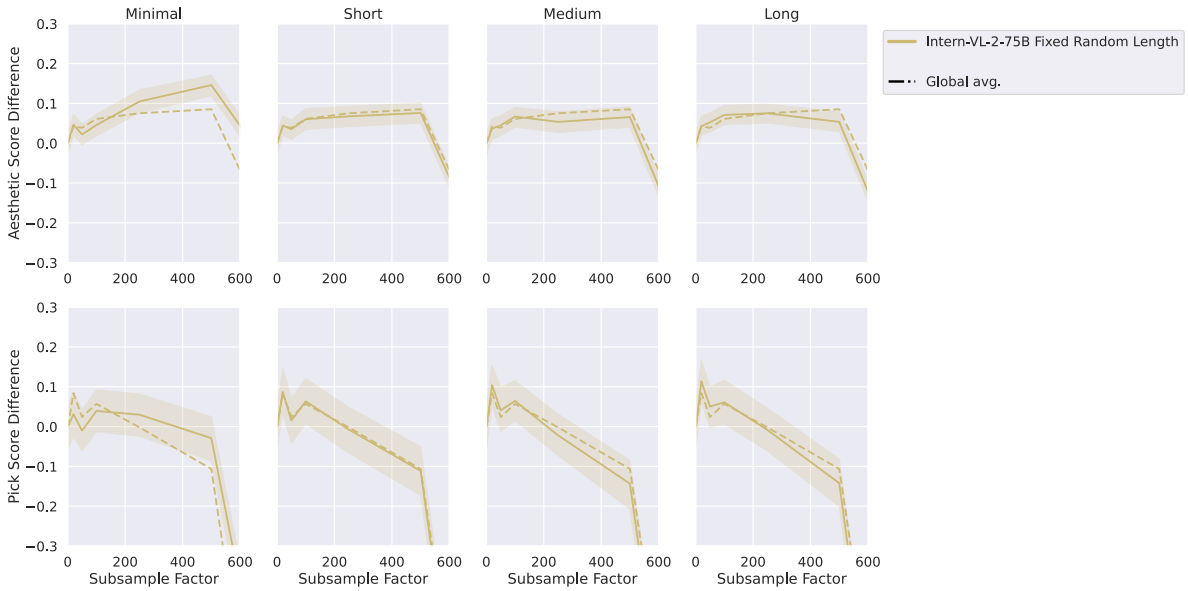


Figure 12. Diffusion models behave suprisingly robust in data-constrained training. Subsampling factor refers to the subset size of training images and respective epochs. Scores are depicted as difference to the model trained on all 1M images.

E. Personas

In the following, we provide further details and examples of the explicit persona sampling discussed in Sec 4.2. For any given image, we first prompt the VLM to asses which personalities are suitable for that image. For example, a description from a biologist focusing on flora and fauna would not be particularly helpful if no natural elements were present in the image. To that, end we use the following prompt, which we prepend to the 20 persona descriptions outlined in Tabs. 2 and 3.

```
Ehe task is to describe the objects and scene
in the image as if the description
can be used to prompt a text-to-image
generator model to generate images.
For the given roles provide a True or
False rating if the role is suitable to
give a caption for the image.
```

Subsequently, we prompt the model to generate a response JSON with one description per selected role. Here, we use an instruction prompt similar to the one described in Sec. A.1. We sampled at least three captions of different lengths for each role, resulting in at least 20 captions per image. We depict some qualitative examples in Fig. 13. The different persona captions provide a lot of diversity on subjective terms like *cinematic*, *vivid*, or *impressive*. In practice, such formulations are often used in prompting text-to-image models and can also be found in GenAI-Bench.

F. Bias

This section provides additional details on our experiments relating to model bias. In the main body, we showed that bias behavior varies across checkpoints, even if all were trained on the same images. Using different training captions led to significant shifts in the gender ratio of the generated images.

Fig. 14 presents more nuanced results, illustrating four random occupations alongside the average distribution. We included these four occupations for clear visualization purposes, but the findings here generalize to the entire set of 150 occupations.

On the left (Fig. 14a), we examine the gender ratio within the training dataset. The blue bars represent the gender ratio in the images, while the orange bars correspond to the ratio in the original LAION captions. The green and red bars show the gender ratios in dense and random-length captions generated with InternVL2, respectively. Consistent with the main results, the gender distributions for dense and random-length captions are closely aligned for individual examples and the average distribution. These distributions are also similar to the image gender ratio (blue), indicating that the captioning model effectively annotated the images and accurately captured respective gender expression. In contrast, the original LAION captions exhibit considerable noise and deviate significantly from the image distribution and synthetic caption distributions. Quantitative results further support these findings: for 91% of the occupations, the synthetic caption distributions differ no more than 5% from each other. Conversely, only around 10% of occupations

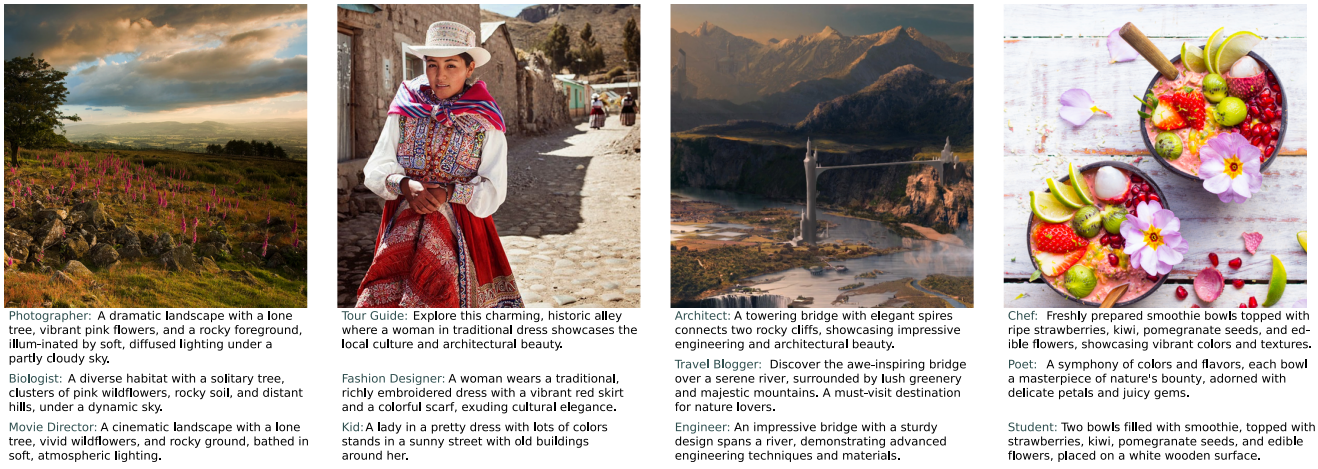
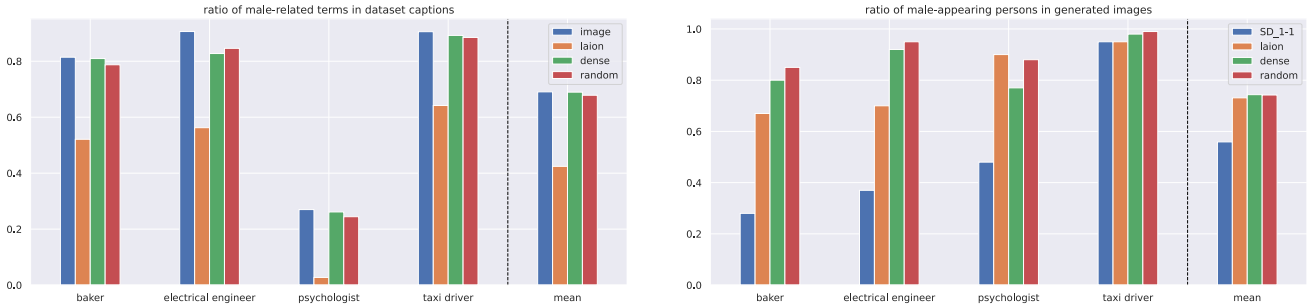


Figure 13. Qualitative examples of diverse persona captions.



(a) Gender ratio in captions

(b) Gender ratio in generated images

Figure 14. Bias investigation across four occupations. (a) depicts the training image gender distribution (blue), and the other bars depict the caption distribution. Generally, both synthetic captions (green and red) follow the image distribution well, whereas original Laion captions are much less well-describing and deviate substantially from the image distribution. (b) depicts the gender distribution in generated images with baseline SDv1.1 and the tuned versions. After continual pre-training (right side, blue vs. other bars), we observe an increase in male-dominated bias. The generated images follow the training image and caption distribution very well except for "psychologist", where there is a complete mismatch between training and outcome distributions.

show a difference of 5% or less between LAION and synthetic caption distributions.

On the right (Fig. 14b), we explore the gender ratio of images generated by the models. The blue bars depict results for the baseline SDv1.1 model, while the other bars represent results for SDv1.1 fine-tuned on original LAION captions (orange), dense captions (green), and random-length (red) captions, respectively. Once again, we observe that the models trained on synthetic captions produce similar gender ratios in their outputs, whereas the model trained on LAION captions differs significantly. These results indicate that gender distribution in the caption strongly influences the output bias of a model trained on them.

Interestingly, the caption distribution (left figure) does not clearly predict occupations' output distribution (right figure). Looking at the occupational examples of "baker", "electrical engineer", and "taxi driver" we find that the gender ratio in the generated images roughly follows the image and synthetic caption distribution in the dataset. However, the generated distribution from the LAION model is substantially different from its caption distribution. Furthermore, for "psychologist", we find that while the gender ratio in dataset captions and images is below 30%, in generated images, the ratio is around 80%. We found that this occupation is underrepresented in the dataset, leading us to assume that the training distribution is less reflective of the outcome distribution.

Upon further analysis, we observe that all three of our checkpoints tend to generate more male-appearing images than baseline SDv1.1. Though unwanted, this seems expected for models trained on synthetic captions (green/red), as those images and captions are already overall more male-related. However, despite LAION captions containing a significantly lower proportion of male-related content, the

model trained on them still shows a strong tendency to generate male-appearing individuals. We hypothesize that the trained model's gender bias is mainly influenced by the respective image distribution, although the correlation is not particularly strong. These findings suggest that the resulting output distribution cannot be easily predicted or controlled by the image or caption distribution alone. Instead, there seems to be a more complex interplay between the images, captions, and the various pre-trained components (such as the text encoder, e.g., CLIP or T5) of the T2I model.

Persona	Description
Photographer	Focuses on technical aspects like lighting, composition, focal points, exposure, and depth of field. The description highlights how these elements work together to create a visually striking or harmonious image.
Artist	Emphasizes emotions, aesthetics, color theory, and the overall atmosphere. This role leans on subjective interpretation, creating a narrative that may evoke feelings or inspire creativity.
Student	Provides a straightforward, observational description of what is depicted, often focusing on what is learned or noticed. The language is usually simple and direct, highlighting key visual elements.
Kid	The description is often simple, imaginative, and filled with curiosity. It tends to focus on vibrant and exciting details, sometimes personifying objects or describing them with enthusiasm.
Scientist	Describes the scene with precision, often breaking down biological, physical, or chemical details. The focus is on facts, measurements, and processes that may not be immediately obvious but are scientifically relevant.
Historian	Provides context around the historical significance of the objects, architecture, or figures in the scene. It often includes dates, origins, cultural relevance, and how the past informs the present.
Poet	Uses metaphorical and symbolic language to evoke emotion and mood. The description is often lyrical and less concerned with technical accuracy, instead aiming to capture an essence or deeper meaning.
Architect	Focuses on the structure, materials, geometry, and spatial design. Descriptions include functional and aesthetic details, often emphasizing how the space interacts with its environment or inhabitants.
Fashion Designer	Describes garments, fabrics, textures, and the fit of clothing on the body. Attention is given to color schemes, trends, and how the overall look conveys style or message.
Chef	Describes food in terms of appearance, texture, and taste, often highlighting the freshness or quality of ingredients. Presentation, plating, and culinary techniques are emphasized.
Movie Director	Provides a cinematic description of the scene, focusing on atmosphere, lighting, framing, and potential narrative. This role often includes emotional undertones or the suggestion of a storyline.
Tour Guide	Offers a guided description, highlighting landmarks, cultural relevance, or the scenic beauty of a location. The tone is informative and inviting, meant to engage and educate the audience about the place.
Psychologist	Analyzes the psychological states of people or the emotional tone of the scene. The description is introspective, focusing on body language, facial expressions, and possible underlying emotions or tensions.
Travel Blogger	Describes the experience of being in a place, often using enthusiastic, sensory-rich language. The focus is on personal experience, beauty, and what makes the location worth visiting.
Mechanic	Focuses on the technical and mechanical aspects of machines or vehicles, describing their functions, condition, and efficiency. The language is often practical and concerned with performance and maintenance.
Biologist	Describes natural elements such as flora, fauna, ecosystems, or weather, with attention to scientific classifications and behaviors. The focus is on living organisms and their interaction with the environment.
Detective	Analyzes the scene for clues, providing detailed observations that suggest a narrative or mystery. The language is often investigative, with an emphasis on small, telling details that could reveal a larger story.

Table 2. Overview of personas and respective prompt description for explicit diversification. Table continues in Tab. 3.

Persona	Description
Meteorologist	Focuses on weather patterns, climate conditions, and atmospheric details. The description often includes information about wind, temperature, and the impact of weather on the environment.
Interior Designer	Describes the layout, furniture, color schemes, and design elements of a room or space. The focus is on how the design choices create a functional and aesthetically pleasing environment.
Engineer	Focuses on the design, functionality, and technical specifications of structures, systems, or machinery. Descriptions emphasize problem-solving, durability, and how elements work together to achieve a purpose.

Table 3. Continuation of Tab. 2