MID-Space: Aligning Diverse Communities' Needs to Inclusive Public Spaces

Shravan Nayak[1,2] **Rashid Mushkani**[1, 2] **Hugo Berard**[2] **Allison Cohen**[1]
Shin Koseki[1,2] **Hadrien Bertrand**[1]

[1]Mila – Quebec AI Institute, [2]Université de Montréal

Abstract

The ability to create visualizations of urban public spaces is a unique skillset that confers disproportionate power and influence over the city's architectural outcomes. Our goal is to democratize that power; putting easy-to-use visualization tools in the hands of marginalized community members so that they can expand their influence over the spaces they occupy. Furthermore, we aim to finetune these visualization tools using images that align with localized notions of equitable, diverse and inclusive public space. To achieve this, we built the MID-Space dataset. It contains preferences for urban public spaces based on criteria such as inclusivity, diversity and comfort. In this paper, we discuss our dataset development process, analyze the annotations obtained and demonstrate the potential for aligned models.

1 Introduction

Public spaces are built to serve the economic, social and political needs of local communities [7]. Urban designers and landscape architects play a key role in shaping these environments, employing visual tools such as drawings, images and 3D models as a primary medium to communicate their ideas [2]. These visual representations guide the development and experience of public spaces, influencing how these spaces are realized and how different community needs are addressed [5]. However, as sociologists have long argued [9], space is not a neutral backdrop but is socially produced through relationships of power and control. Consequently, the visual representations used by designers are not merely aesthetic choices but inherently shape who benefits from or is marginalized by these choices.

As cities continue to evolve and diversify, we should consider opportunities to democratize the visualization process beyond the designers to include a wider group of community members. Newfound AI capabilities, particularly text-to-image (T2I) models, reduce the barriers to entry, enabling people of all backgrounds to take part in the visualization process. The remaining challenge is how to ensure that the generated image aligns with what the user had in their own mind, creating an authentic illustration of what they would like to see.

To this end, we present the **M**ontreal **I**nclusive and **D**iverse Spaces (MID-Space) dataset, to align AI-generated visualizations with pluralistic human needs and preferences. Developed in coordination with a dozen community organizations through a series of workshops, the dataset includes textual prompts, corresponding AI-generated images, and annotations about pair-wise preferences with respect to one (or multiple) criteria. This dataset can serve as a resource among AI alignment researchers seeking to produce imagery that aligns with the preferences of the individual. Furthermore, it can serve as a tool among urban designers, used to obtain crowd-sourced input on the design of public spaces. In summary, the main contributions of our work are twofold:

(i) We introduce MID-Space, a preference alignment dataset designed to bridge the gap between AI-generated visualizations and diverse communities' preferences in public space design.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(ii) We demonstrate the dataset's utility by fine-tuning Stable Diffusion XL with preference alignment techniques, enhancing AI output alignment with pluralistic human values.

2 Related Work

Preference Alignment Datasets Preference alignment datasets have significantly improved the quality and alignment of T2I models [1, 4, 6, 17, 21]. Notable examples include Simulacra Aesthetic Captions [14], which contains 238,000 synthetic images rated for aesthetics, and Pic-a-pick [8], which features over 500,000 preference data points. Similarly, Image Reward [20] assesses images on alignment, fidelity, and harmlessness, besides ranking multiple images for the same prompt. Human Preference Score (HPS) [19] and its follow-up, HPS v2 [18], offer large-scale binary preference pairs, and train reward models that accurately capture human preferences. Relative to the literature, the MID-Space dataset is unique in terms of its: i) domain focus in urban design; ii) six alignment criteria that emphasize markers of equity and inclusion; and iii) annotators representing the needs and priorities of minority populations in Montreal.

Generative-AI in Urban Planning AI is beginning to play a transformative role in the domain of urban planning. For instance, Stable Diffusion was used to integrate AI-generated graffiti with building façades to simulate interactions between evolving city structures [16]. Similarly, researchers have investigated how generative algorithms reshape collective memory by studying visitor engagement with real and AI-generated images, providing insights into spatial perception [15]. AI models have also been developed to generate day-to-night street views, enhancing perceptions of safety and auditing urban environments [10]. Furthermore, AI-generated visuals of car-free cities have been shown to significantly increase public support for sustainable transport policies [3]. To the best of our knowledge, this is the first work that explores alignment research in the context of urban planning.

3 MID-Space: Dataset Creation

The MID-Space dataset was developed in multiple phases, demonstrating a commitment to and collaboration with vulnerable community groups in Montreal. We outline the steps involved in the creation of the dataset below. The project followed best practices for citizen collaboration and community engagement and was approved by the appropriate Research Ethics Committee.[1]

Criteria Selection To create a dataset of equitably designed public spaces, we first needed to determine how equity would be defined and evaluated. The process began in 2023 with three workshops, each involving 18 to 23 participants from diverse community groups. Participants reviewed images of Montreal's public spaces and engaged in discussions about the most important attributes that made the spaces inclusive. From these workshops, six criteria emerged: accessibility, safety, diversity, inclusivity, invitingness, and comfort. More details can be found in Appendix A

Annotator Selection We worked with a group of 16 annotators who were identified with the support of twelve community organizations. Our criteria for annotator selection included those living in Montreal and having at least one (and in some cases, multiple) minority identity markers. The demographic markers of our annotators included women, LGBTQ+ individuals, Indigenous peoples, immigrants, people with disabilities, racial and ethnic minorities, and the elderly.

Prompt Collection Participants were then convened for an additional session wherein they were divided into five groups of 4-5 participants. Each group included a combination of 2-3 citizens, one urbanism professional, and one AI expert. These groups were tasked with generating as many prompts as possible based on distinct scenarios that encompassed common public space typologies, amenities, and ambiances in Montreal. From this workshop, 440 prompts in total were collected. To expand the dataset further, these prompts were input into a large language model (GPT-4o [11]), which generated an additional 2910 prompts. We employed three different prompting strategies, detailed in Appendix B, to ensure the synthetic prompts mirrored human input and captured diverse public space typologies and features.

[1] Details omitted for anonymity.

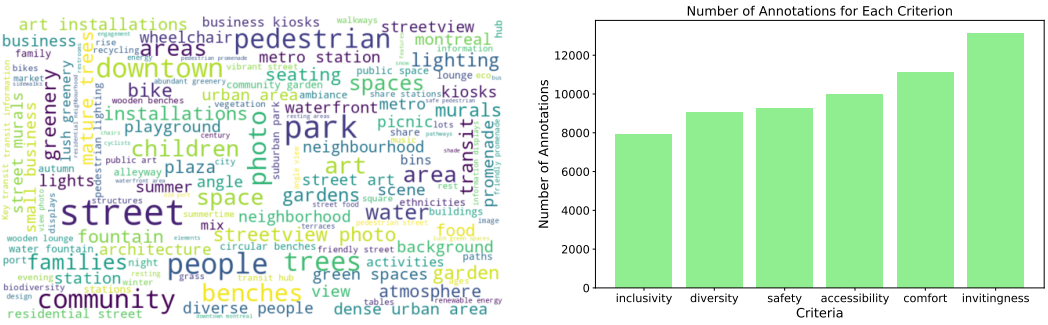


Figure 1: *(Left)* Word cloud containing various entities related to public spaces present in our prompts. *(Right)* Number of preference annotations collected from participants for six criteria.

Image Generation We used Stable Diffusion XL [13] to generate images for the collected prompts. To ensure consistency in output style and avoid confounding variables, we opted to use a single text-to-image model. To make the images differentiable, we created 20 images per prompt by varying key parameters such as guidance scale, steps, and seed. From these images, we selected the four that were most distinct using a greedy strategy based on CLIP similarity scores (Appendix C).

Human Annotation We created an accessible interface to ensure annotators of all backgrounds could annotate effectively. The participants were asked to compare two images which were shown side by side. Annotators evaluated each pair of images along three of the six criteria. Participants used a slider to indicate their preference on a continuous scale, ranging from -1 (indicating preference for the left image) to +1 (indicating preference for the right image). If the annotator had no clear preference between the images for a given criterion, the slider could be left in the middle. Before starting, participants attended workshops to familiarize themselves with the platform and the evaluation criteria. More details about the platform are provided in Appendix D.

4 Data Analysis

This section offers an overview of the composition and characteristics of our dataset, with a specific focus on the prompts and human annotations found within the MID-Space dataset.

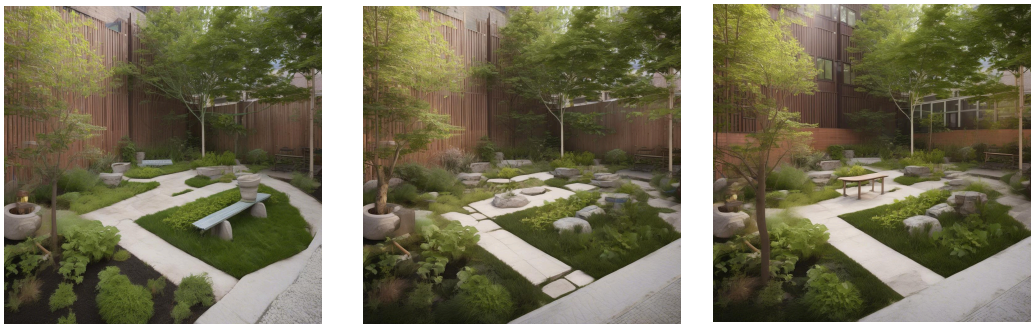
Prompts Our dataset consists of 3,350 prompts that cover a wide range of public space scenarios. The prompts vary in length, with an average of 37 words. They encompass over 3,000 distinct public space concepts, converting various amenities, typologies, and ambiances. Figure 1 (*left*) illustrates the diversity of concepts represented in our prompts.

Human Annotation A total of 16 participants, representing diverse communities, contributed to the annotation process, resulting in 42,131 annotations. Each annotation reflected preferences for up to three criteria. However, in 13,266 instances, no criteria were selected, potentially due to the complexity or ambiguity of the options. Some participants exhibited this indecisiveness more frequently than others. For the remaining annotations, participants contributed between 1,000 and 7,300 annotations each, with an average of two criteria selected per image pair. This offers a valuable opportunity to develop models that capture individual preferences. Figure 1 (*right*) displays the distribution of annotations across the six criteria. In total, we obtained 60,425 image pairs annotated for one of the six criteria. Notably, the "inclusive" criterion received the fewest annotations, possibly due to its ambiguity or the difficulty in evaluating this attribute. In contrast, "inviting" and "comfortable" garnered the most annotations, likely because these criteria are more closely tied to aesthetics and are easier to assess.

5 Experiments

To evaluate the capacity of the MID-Space dataset to adapt model outputs, we finetuned Stable Diffusion XL using preference alignment techniques, specifically Direct Preference Optimization

Prompt: A quiet, inclusive meditation garden in a busy urban area.



SDXL original | SDXL-DPO for inclusion criterion | SDXL-DPO for all six criteria

Figure 2: Images generated for the prompt given above by various models with the same seed, number of generation steps (50) and guidance scale of 5.

[17]. For each criterion we trained a new model, aligning each model with community preferences. We also developed a comprehensive model that integrates all six criteria to explore how different criteria interact and influence one another. More details can be found in Appendix E.

Qualitative Results The generated images were evaluated by experts in urban planning and AI on two types of spaces: indoor and outdoor. While the MID-Space dataset includes only outdoor public spaces, we tested the model's performance on indoor environments to assess its generalization. General trends such as the presence of people, improved seating arrangements, and better handling of surface transitions and spatial enclosure were observed. However, the finetuned model struggled to visualize accessibility features such as ramps, tactile surfaces, and signage. Examples for the two cases are outlined below and more can be found in Appendix F

Outdoor Public Space Analysis: Figure 2 provides an illustrative example for this case. The image generated by SDXL (left) lacked cohesion with undefined paths and poor accessibility. The inclusivity-finetuned version (middle) had better enclosure but still had uneven paths and limited accessibility. The model finetuned on all six criteria (right) showed smoother transitions, defined paths, and improved seating, enhancing accessibility and spatial integration.

Indoor Public Space Analysis: Figure 7 shows an example for this case. The initial design (left) featured poorly integrated artwork in a rigid space. The diversity-finetuned version (middle) did not improve the artwork integration but displayed more vibrant and well-lit pieces with added connecting passages in the corridor. Similarly, the version finetuned on all criteria (right) introduced abstract artwork but still struggled with artwork integration. However, it added continuous lighting, improved ventilation, and opened the side walls of the corridor, making the space more connected. The presence of people contributed to a more dynamic environment.

6 Conclusion and Limitations

In this paper, we introduced the MID-Space dataset, designed to align AI-generated visualizations with diverse human preferences in urban public space design. Through community engagement workshops, we identified six key criteria for creating equitable public spaces. We demonstrated the utility of the MID-Space dataset by fine-tuning Stable Diffusion XL using DPO resulting in AI models capable of generating visualizations that reflect localized values of equity and inclusion.

However, a subset of annotations included contradictory preferences across criterias. To properly account for those, better training methods are needed, that account for the criteria as signals, and more generally, by creating approaches that consider the pluralistic nature of this dataset.

MID-Space has limitations including: a small sample size, challenges in generalizing annotations to the level of the demographic group and de-noising techniques that reduced nuance. In the future, we look to enhance the number of annotators and develop more robust guidelines to improve annotation consistency. Additionally, we want to explore better training techniques to capture more nuance and further enhance model performance and alignment with human preferences.

Acknowledgments

This research was funded by the *Soutien aux initiatives avec les collectivités et les entreprises – Collaboration avec les organismes communautaires – Projet de partenariat communautaire de l'Université de Montréal*. Additional support was provided by Mitacs and the *Fonds de recherche du Québec - Société et culture* Doctoral Research Scholarships. We extend our gratitude to the following community organizations in Montreal, whose collaboration and support in connecting us with citizen collaborators have been invaluable: the Congolese Community Center of Montreal, Altergo, La Maisonnée, Cummings Centre, Projet Changement - Community Center for Seniors, Women's Center of Plateau Mont-Royal, LGBTQ+ Community Center of Montreal, Marguerite-Bourgeoys Hub, RÉZO, Afrique au Féminin, L'Agence On est là!, and Montreal Women's Groups Table. This research was enabled in part by compute resources provided by Mila.

References

[1] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2024.

[2] James Corner. *The Agency of Mapping: Speculation, Critique, and Invention*. Reaktion Books, London, 1999.

[3] Rachit Dubey, Mathew D. Hardy, Thomas L. Griffiths, and Rahul Bhui. Ai-generated visuals of car-free us cities help improve support for sustainable policies. *Nature Sustainability*, 7:399–403, 2024.

[4] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models, 2023.

[5] David Harvey. The right to the city. *New Left Review*, 53:23–40, 2008.

[6] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference, 2024.

[7] Jane Jacobs. *The Death and Life of Great American Cities*. Random House, New York, 1961.

[8] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023.

[9] Henri Lefebvre. *The Production of Space*. Blackwell Publishing, Oxford, 1974.

[10] Zhiyi Liu, Tingting Li, Tianyi Ren, Da Chen, Wenjing Li, and Waishan Qiu. Day-to-night street view image generation for 24-hour urban scene auditing using generative ai. *Urban Studies*, 61(1):78–94, 2024.

[11] OpenAI. Gpt-4 technical report, 2024.

[12] Barbara Plank and Gertjan van Noord. Effective measures of domain similarity for parsing. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[14] John David Pressman, Katherine Crowson, and Simulacra Captions Contributors. Simulacra aesthetic captions. Technical Report Version 1.0, Stability AI, 2022. url https://github.com/JD-P/simulacra-aesthetic-captions .

[15] Eduardo Rico, Sheng-Yang Huang, Julian Besems, and Wanqi Gao. (in)visible cities: What generative algorithms tell us about our collective memory schema. *Journal of Architectural Computing*, 18(2):215–229, 2023.

[16] Naai-Jung Shih. Ai-generated graffiti simulation for building façade and city fabric. *Journal of Urban Architecture and Design*, 12(3):45–60, 2023.

[17] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023.

Definitions

Public space - Public space is an area where everyone can go, like parks and streets, designed for people to meet, play, and relax together in cities and towns.

Inclusion (Inclusive) - Spaces where everyone is welcome and feels respected. These are places that do not discriminate against anyone.

Safe / Secure - Spaces where public safety is ensured through various measures. Spaces where one feels calm and safe, free from dangers related to physical elements, pollution, or any other concerns that could diminish a sense of security.

Comfortable - Well-equipped spaces with quality facilities that provide material comfort; places where one feels at ease and protected from the elements.

Inviting - Spaces that attract and engage people through appealing elements and activities; places that encourage community participation and interaction.

Diverse - Spaces that cater to the diversity of social groups and to the variety of services, activities, and functions. These are places offering a range of uses and meeting the needs of different cultures, ages, and abilities.

Accessibility - Urban spaces that are easily accessible and navigable for everyone, regardless of physical ability. These include features such as ramps, wide walkways, clear signage, and tactile indicators for safe and convenient access throughout the area.

Figure 3: Definitions of the six criteria shown to annotators for dataset annotations

[18] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

[19] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference, 2023.

[20] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

[21] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model, 2024.

Appendix

A Criteria Selection for Evaluation

MID-Space includes annotations based on six criteria: accessibility, safety, diversity, inclusivity, invitingness, and comfort. Figure 3 shows the definitions provided to the annotators for each criterion for the annotation task. Prior to the annotation process, we conducted a detailed workshop to explain the criteria and provide guidance on their use, ensuring annotators had a clear understanding while allowing for individual interpretation.

B Prompt Augmentation

We collected 440 prompts from annotators, each representing distinct scenarios related to public space typologies, amenities, and ambiances in Montreal. To expand the dataset, we supplemented these with synthetic captions generated by GPT-4o [11]. To ensure diversity, we incorporated a wide range of concepts drawn from public space literature, such as typologies like parks and wide walkways, amenities such as seating areas and streetlights, natural elements like forests and vegetation, locations such as suburban Montreal and old industrial ports, people such as First Nations and children, architectural elements like duplexes and houses, transportation modes including bike lanes and trams, artistic features such as street murals and art sculptures, different times like

winter, nighttime, and Christmas, and animals such as dogs and raccoons. These concepts were used to increase the diversity of the concepts in the generated prompts. We employed three distinct prompting strategies to ensure that the synthetic prompts retained similarity to those generated by humans while covering a wide range of topics relevant to public space design. The specific strategies and prompts used for the LLM are outlined below.

Method 1 We randomly sampled 8-16 prompts from the human-generated set and used them as in-context examples for the LLM to generate new prompts.

System prompt used Method 1

Your task is to craft detailed and imaginative prompts suitable for diffusion models like Stable Diffusion. These prompts should generate images illustrating the variety of Montreal's public spaces, capturing the community's diverse aspirations and values.
Each prompt must be rooted in a specific scenario related to Montreal's public spaces. You will be provided with the scenario, keywords, and examples of prompts related to these scenarios. Using this information, your task is to create a series of diverse, con¯textually rich, and relevant prompts following a style similar to the ones given as examples. These should aim to generate images showcasing Montreal's public spaces from varied perspectives.

Method 2 We provided the LLM with a detailed scenario which was also provided to the annotators during the initial prompt collection phase. Along with this we also provided several keywords related to the public space concepts mentioned earlier. Additionally, we included 8 randomly selected in-context samples relevant to the scenario guiding the model to generate new prompts based on these concepts.

System prompt used Method 2

Your task is to craft detailed and imaginative prompts suitable for diffusion models like Stable Diffusion. These prompts should generate images illustrating the variety of Montreal's public spaces, capturing the community's diverse aspirations and values. To achieve this, you will construct prompts using specific keywords provided for the following categories:
Typology: The type of spaces you want to depict
Elements: Distinct elements to include in your scene
Context: The scenarios in which your elements are placed
Style: The artistic style or technique that the image should emulate, defining its visual appearance
Mood: The overall mood or atmosphere of the image

You will also be given a few examples that have been generated using these keywords. Using all this information create a complete, coherent prompt similar in style to the examples. Aim for creativity and diversity in your prompts, ensuring they cover several aspects of the keywords given. These should aim to generate images showcasing Montreal's public spaces from varied perspectives.

Note:
1. Ensure your prompts integrate some of the provided keywords to encapsulate the community's desired visions of Montreal's public spaces but ensure that style and length is same as the examples.
2. Do not mention the style and mood explicitly. Use keywords that bring out these attributes naturally.
3. The style and the length of the prompts should be similar to the examples given. The prompt should be less than 77 tokens.

Method 3 This method used a template-based approach where specific keywords related to public space concepts in the original prompts were masked. We instructed the LLM to replace these masked keywords with concepts from a wide variety of public space themes. This ensured the prompts closely followed the structure of the human-generated ones while incorporating a diverse range of concepts.

System prompt and incontext sample used Method 3

Your task is to craft detailed and imaginative prompts suitable for diffusion models like Stable Diffusion. These prompts should generate images illustrating the variety of Montreal's public spaces, capturing the community's diverse aspirations and values.

For this task, you will be provided with a templated sentence containing several placeholders. Each placeholder represents a specific category (e.g., [Typology], [Location], [Activity], [Amenity]). Alongside the templated sentence, you will receive a list of words or phrases corresponding to each category. Your objective is to select the most appropriate word or phrase from each list to fill in the placeholders, creating a meaningful and grammatically correct sentence.

The structure of the templated sentence might require minimal modifications to ensure grammatical correctness and cohesiveness once the placeholders are filled.

Example 1:

Template: a [Typology] for [People] in [Location]

Keywords:
Typology: artistic eco friendly park, pedestrian street, all identities, two-story residential street, park, neighbourhood public space, urban square, wide walkway
People: elderly person, adults, first nations, children, teenagers, adults and elderly people, black and white families, a mother and her child, people, various ethnicities
Location: plateau, wellington neighbourhood, old port, montreal 's chinatown, old montreal, Montreal, downtown montreal, mont royal street

Output: A neighbourhood public space for children, teenagers, adults and elderly people in Montreal

<more examples>

We calculated Jensen-Shannon Divergence (JSD) [12] scores to measure the difference between the generated prompts and the human-provided data. Higher JSD values indicate greater divergence from human prompts, while lower values suggest closer alignment. We found that Method 1 and Method 2 produced prompts that deviated more from the human prompts, with JSD scores of 0.53 and 0.58, respectively, compared to Method 3, which had a JSD of 0.4. These results demonstrate that all three methods contribute to generating a diverse range of prompts, with Method 3 being more aligned with the human data.

C Image Generation

We used Stable Diffusion XL to generate images for each prompt. During an initial annotation workshop, participants indicated that it was difficult to meaningfully differentiate between the images. To address this, we generated 20 images per prompt by varying several hyperparameters, including seed, denoising steps, and guidance scale. From these 20 images, we applied a greedy selection method to identify the 4 most diverse images based on CLIP similarity scores. The method, detailed in Algorithm 1, selects images that minimize similarity to previously chosen ones, ensuring diversity among the selected images.

D Human Study Details

Figure 4 shows the web interface we developed to collect the data. Users could express their preferences by moving a slider to the left or right, depending on which image they preferred and how much they preferred it. Additionally, users could click on the purple dot next to each criterion to view its definition.

E Experiment Details

For fine-tuning Stable Diffusion XL using Direct Preference Optimization (DPO), we closely followed the hyperparameters from the original paper. We used an effective batch size of 64 and a learning rate of $1e-8$ with a 20% linear warmup. The β parameter was set to 5000. Individual models were fine-tuned for 500 steps, while the comprehensive model, which incorporates preference pairs from the entire dataset, was trained for 1500 steps. We found that increasing the training steps led to improved results, and we believe further training could enhance performance even more. All experiments were conducted on a single NVIDIA A100 80GB GPU.

We normalized the continuous preference values into binary values to better align with the requirements of DPO. However, future work could explore the use of continuous values to capture more nuanced preferences,

Algorithm 1 Selecting the 4 Most Diverse Images Using CLIP Similarity Scores

Input: similarity matrix S of size $n \times n$, number of diverse images $k = 4$
Output: Indices of the selected diverse images $selected\_indices$
$n \leftarrow \text{len}(S)$
$selected\_indices \leftarrow []$
$first\_index \leftarrow \arg\min(\text{mean}(S, \text{axis} = 1))$
Append $first\_index$ to $selected\_indices$
for $j = 1$ to $k - 1$ **do**
    $min\_similarity \leftarrow \infty$
    $next\_index \leftarrow -1$
    for $i = 0$ to $n - 1$ **do**
        if $i \notin selected\_indices$ **then**
            $current\_similarity \leftarrow \max(S[selected\_indices, i])$
            if $current\_similarity < min\_similarity$ **then**
                $min\_similarity \leftarrow current\_similarity$
                $next\_index \leftarrow i$
            end if
        end if
    end for
    Append $next\_index$ to $selected\_indices$
end for
return $selected\_indices$

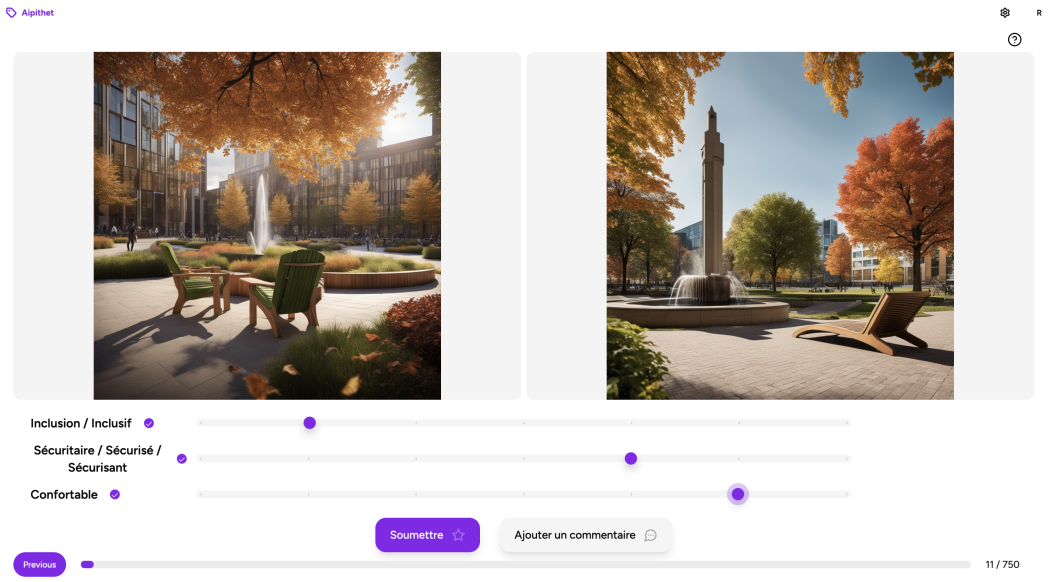


Figure 4: Web interface used to collect the data for MID-Space

potentially improving model alignment with finer-grained human judgments. When training the comprehensive model on the entire dataset, we encountered cases with conflicting annotations. Specifically, for some image pairs, one image was preferred for certain criteria while the other was favored for different criteria, as annotators could select up to three criteria per image. To resolve these conflicts, we applied a majority voting system, selecting the image with the most criteria in its favor as the preferred one. In cases of a tie, the winner was chosen randomly. We found 5,050 such instances where annotations conflicted. We believe that this issue could be better addressed by developing more sophisticated algorithms that account for the criteria as signals, and more generally, by creating approaches that consider the pluralistic nature of this dataset.

F Qualitative Results

Prompt: A bike path separated from vehicular traffic by barriers.

SDXL original | SDXL-DPO for safe criterion | SDXL-DPO for all six criteria

Figure 5: Images generated for the prompt given above with a particular focus on safety, by various models with the same seed, number of generation steps (50), and guidance scale of 5.

Prompt: A shopping mall with wide aisles, ramps, and accessible restrooms.

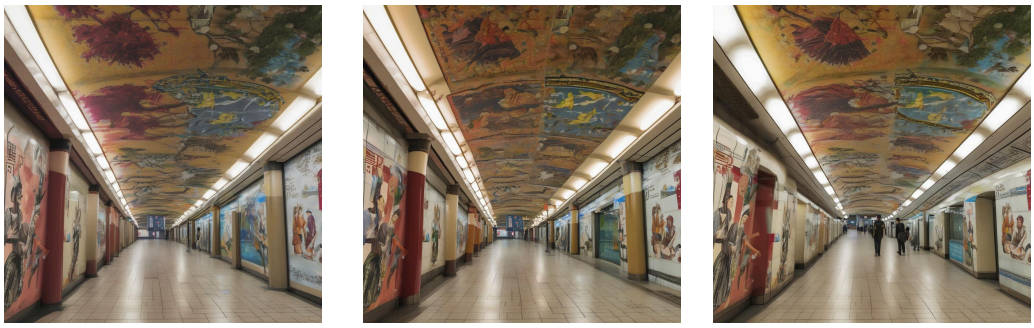
SDXL original | SDXL-DPO for accessibility criterion | SDXL-DPO for all six criteria

Figure 6: Images generated for the prompt given above with a particular focus on accessibility by various models with the same seed, number of generation steps (50), and guidance scale of 5.

Figure 5 shows the original SDXL design (left) of the bike path, which presents a basic separation between vehicles and cyclists using metal barriers. However, the narrow space and lack of path differentiation create a utilitarian feel, with bollards potentially obstructing smooth transitions. In the version finetuned for safety (middle), the barriers are more prominent, enhancing separation but maintaining a rigid, functional layout without features like bike parking or rest areas. The version finetuned for all criteria (right) offers wider lanes and smoother transitions, improving accessibility and comfort, though the image quality becomes more distorted.

Figure 6 shows the original design of the shopping mall (left), which features wide aisles but lacks clear differentiation for ramps. In the version finetuned for accessibility (middle), ramps remain problematic, and there is an increase in the number of stairs. In the version finetuned for all criteria (right), transitions between spaces and surfaces are smoother, though ramps are still difficult to navigate. Aside from an increase in the number of people, greenery, and stairs, there is little improvement, and the model has yet to fully grasp ramps or accessible restrooms. There is some indication of added seating areas, though they are not clearly visible. This result suggests that SDXL struggles to generate features such as ramps, which are crucial for accessibility. Since SDXL cannot effectively produce these features, they might be absent in the dataset, preventing the model from learning to improve on them.

Prompt: A metro station decorated with artwork representing different heritages.



SDXL original | SDXL-DPO for diversity criterion | SDXL-DPO for all six criteria

Figure 7: Images generated for the prompt given above with a particular focus on diversity, by various models with the same seed, number of generation steps (50), and guidance scale of 5.