PictoSD: A Low-Resource Adaptation Strategy for Diffusion Models in Pictogram Generation

Anonymous ACL submission

Abstract

Pictograms are widely used to support communication through visual representations of concepts and actions. However, adapting recent text-to-image diffusion models to the pictogram domain remains challenging under limited data, particularly for culturally specific concepts. In this paper, we introduce PictoSD, a low-resource adaptation strategy for pictogram generation based on pre-trained diffusion models. Our approach specializes Stable Diffusion to the pictogram domain exploiting SaRA, a parameter-efficient fine-tuning method that preserves semantic priors while learning the characteristic visual style of pictograms. To support adaptation under limited data, we construct and curate a text-paired subset from ARASAAC by combining keywords with refined captions generated through a vision-language model. We further improve visual consistency and structural coherence through Perturbed-Attention Guidance (PAG). PictoSD enables pictogram generation from both textual and visual inputs by combining concise keywords or prompts with image-guided spatial control through ControlNet. Experimental results show that PictoSD produces visually consistent and semantically coherent pictograms while outperforming conventional full fine-tuning in terms of the Visual-Linguistic Harmony Index (VLHI). To illustrate the potential of our approach for culturally grounded communication, we generate a small collection of pictograms representative of Mexican culture.

1 Introduction

Pictograms are graphical representations of concepts such as actions, objects, people, animals, or places, and are widely used to support communication through visual symbols (Pereira et al., 2022). Due to their ability to transmit information in a clear and simple manner (Barros et al., 2014), pictograms can help people with complex communication needs express ideas that would be difficult to communicate using only words (Vicente et al., 2024). However, pictogram design is typically manual and time-consuming. Furthermore, since the interpretation of a pictogram highly depends on cultural background (Vaschalde et al., 2018), creating pictograms for culturally specific concepts becomes particularly challenging.

Recent text-to-image (T2I) diffusion models open new opportunities for automatic pictogram generation. Nevertheless, adapting these models to the pictogram domain remains difficult under low-resource conditions. In particular, pictogram generation requires preserving the semantic knowledge of a pre-trained model while learning the characteristic visual style of pictograms from relatively few examples. Moreover, text prompts alone may not be sufficient to describe the desired visual structure of a pictogram, especially for actions, abstract concepts, or culturally grounded scenes.

In this work, we introduce PictoSD, a low-resource adaptation strategy for pictogram generation based on Stable Diffusion (SD) (Rombach et al., 2022). PictoSD supports both textual and visual inputs, allowing users to combine keywords or prompts with reference images to guide the generated pictograms. To adapt the model to the pictogram domain, we leverage SaRA (Hu et al., 2025), an efficient fine-tuning method that preserves semantic priors while learning the pictogram style. In addition, we construct and curate a text-paired subset from ARASAAC (Rodrigo and Corral, 2013; Paolieri and Marful, 2018) by enriching keywords with refined captions generated and manually corrected through a vision-language model. This allows PictoSD to learn from both concise concept-level inputs and more detailed textual descriptions.

To further improve visual consistency and controllability, PictoSD incorporates guidance and spatial conditioning mechanisms during generation. As a result, PictoSD can generate pictograms that

combine textual intent with visual structure while preserving semantic consistency and the characteristic pictogram style.

Particularly in Mexico, there is a significant gap in pictogram resources, as no pictogram database currently reflects the cultural and linguistic diversity of the country. To illustrate the potential of PictoSD in this direction, we generate a small collection of pictograms representative of Mexican culture. Remarkably, PictoSD is able to capture the pictogram style from relatively few examples, making it easily adaptable to other regions and cultural contexts. The contributions of our work are summarized as follows:

- We present PictoSD, a low-resource adaptation strategy for pictogram generation based on a pre-trained T2I diffusion model.
- We construct and curate a text-paired pictogram subset from ARASAAC by combining keywords with refined captions generated through a vision-language model.
- We adapt Stable Diffusion to the pictogram domain while preserving semantic consistency under limited training data.
- We enable pictogram generation from both textual and visual inputs through guided and spatially controlled generation.
- We demonstrate the potential of PictoSD for culturally grounded generation by creating a small collection of pictograms representative of Mexican culture.

2 Related Work

2.1 Pictogram and Icon Generation

Recent advances in generative models have enabled the automatic creation of icons and pictograms from textual descriptions. Early approaches employed GAN-based architectures (Yang et al., 2021; Chen et al., 2022b; Mino and Spanakis, 2018), while more recent methods leverage autoregressive Transformers (Wu et al., 2023) and text-to-image (T2I) diffusion models (Lin et al., 2024a) to generate icons from textual prompts and style references.

Despite these advances, most existing approaches focus on stylistic icon generation and provide limited support for representing complex communicative concepts. In addition, these methods generally rely only on textual descriptions or style references, without incorporating explicit spatial guidance from input images. PictoAI (Drews et al., 2025) is the only approach we found that explicitly attempts to generate visually consistent pictograms automatically. Nevertheless, it is limited to monochromatic cartographic pictograms and presents scalability constraints. In contrast, our approach focuses on low-resource adaptation of diffusion models for pictogram generation while enabling multimodal control through both textual and visual inputs.

2.2 Low-Resource Adaptation of Diffusion Models

Adapting pre-trained diffusion models to new visual domains has recently received considerable attention. Existing approaches include additive adapters (Chen et al., 2022a; Mou et al., 2024), selective parameter optimization (Guo et al., 2021), and low-rank adaptation strategies such as LoRA (Hu et al., 2022; Han et al., 2024). Although these approaches enable efficient adaptation, preserving the semantic priors of large T2I models under limited data remains challenging.

To address this issue, Hu et al. (2025) propose SaRA, a parameter-efficient fine-tuning method inspired by model pruning. SaRA optimizes parameters with relatively small absolute values while preserving most of the pre-trained weights frozen, achieving effective adaptation in a memory-efficient manner. In this work, we leverage SaRA to adapt Stable Diffusion to the pictogram domain using relatively few examples while maintaining semantic consistency across generated concepts.

2.3 Controllable and Guided Image Generation

Although modern T2I diffusion models can generate high-quality images from text, textual prompts alone often provide limited control over the spatial structure and composition of the generated samples. To overcome this limitation, previous work has explored controllable generation through external conditioning mechanisms such as Composer (Huang et al., 2023), T2I-Adapter (Mou et al., 2024), and ControlNet (Zhang et al., 2023). These methods incorporate additional spatial signals, including sketches, segmentation maps, depth maps, and keyposes, to guide image generation.

Recent work has also explored guidance strategies to improve generation quality during sampling. In particular, Perturbed-Attention Guidance (PAG)

(Ahn et al., 2024) enhances structural consistency without requiring additional training or external classifiers.

Different from previous approaches, our work combines low-resource adaptation, semantic preservation, multimodal controllability, and guided sampling into a unified strategy for pictogram generation. In particular, PictoSD allows users to generate pictograms from both concise keywords and detailed prompts while also incorporating input images to guide the intended visual structure of the generated pictograms.

3 Method

In this section, we describe PictoSD, a strategy for automatically generating visually consistent pictograms with multimodal user inputs: text and images. PictoSD consists of three main components: domain adaptation, structure enhancement, and control transfer (see Figure 1). Together, these components allow the model to preserve the semantic knowledge of a pre-trained T2I diffusion model, learn the pictogram style from limited data, and incorporate input images to guide the spatial structure of the generated pictograms.

3.1 Low-Resource Domain Adaptation

Stable Diffusion (SD) (Rombach et al., 2022) is a T2I diffusion model capable of generating high-quality images from text prompts. In our work, we use SD as the base model and specialize it to the pictogram domain while preserving its ability to represent general concepts.

Designing a pictogram requires strong semantic knowledge to represent a concept accurately, and directly modifying most of the parameters of a pre-trained model might lead to forgetting its semantic priors. To address this issue, we leverage SaRA (Hu et al., 2025), an efficient fine-tuning method for pre-trained diffusion models.

SaRA identifies parameters with relatively small absolute values in the pre-trained weights P_0 through a sparse binary mask M. The selected parameters are optimized through

$$P_M = P_0 \odot M, \tag{1}$$

while the remaining parameters are frozen to preserve prior knowledge. This is especially valuable for pictogram generation, as it enables the model to learn the characteristic visual style of pictograms (intuitively into those relatively small parameter values) without forgetting object and conceptual semantics.

To avoid overfitting, SaRA constrains the rank of the trainable matrices using the nuclear norm-based loss:

$$L_{rank} = ||P_M||_* = \sum_i \sigma_i(P_M), \tag{2}$$

where σ_i are the singular values of P_M.

In order to specialize Stable Diffusion to the pictogram domain with SaRA, we follow the procedure shown in Figure 2. We first collect 1,000 text-paired pictograms from the ARASAAC database, ensuring that they cover a wide variety of concepts, including actions, animals, descriptions, objects, people, places, and idioms. To guarantee visual consistency, all pictograms are taken with a white background and resized to 512×512 pixels.

While each pictogram in ARASAAC is paired with some keywords that describe the concept it represents, using only these keywords as text prompts could degrade model performance in the domain adaptation process, as they may not constitute high-quality textual descriptions (Betker et al., 2023; Chen et al., 2024; Sinha et al., 2024). Hence, adopting the methodology of prior work from the T2I field (Chen et al., 2024; Lin et al., 2024a), we employ the vision-language model LLaVA (Liu et al., 2023) to generate refined textual descriptions for the collected pictograms. Specifically, we prompt LLaVA with the instruction: "*The concept represented by this image is 'keyword'. Describe the image in one short sentence with as much detail as possible.*" Here, *keyword* is a keyword associated with the corresponding pictogram. This produces concise yet semantically rich descriptions. We manually correct fragments of the descriptions that distort the visual content of the pictograms to improve their quality. Finally, all pictograms are paired with both keywords and moderately detailed curated descriptions. Following the practice of Lin et al. (2024a), the text prompt for each pictogram is uniformly sampled from either its keyword or description during the fine-tuning process. This allows the model to learn from both concise concept-level inputs and more detailed textual prompts.

3.2 Structure Enhancement

Diffusion models often struggle to produce images with well-defined structures (Hong et al., 2023), which is problematic for pictogram generation, where images must remain simple and in-

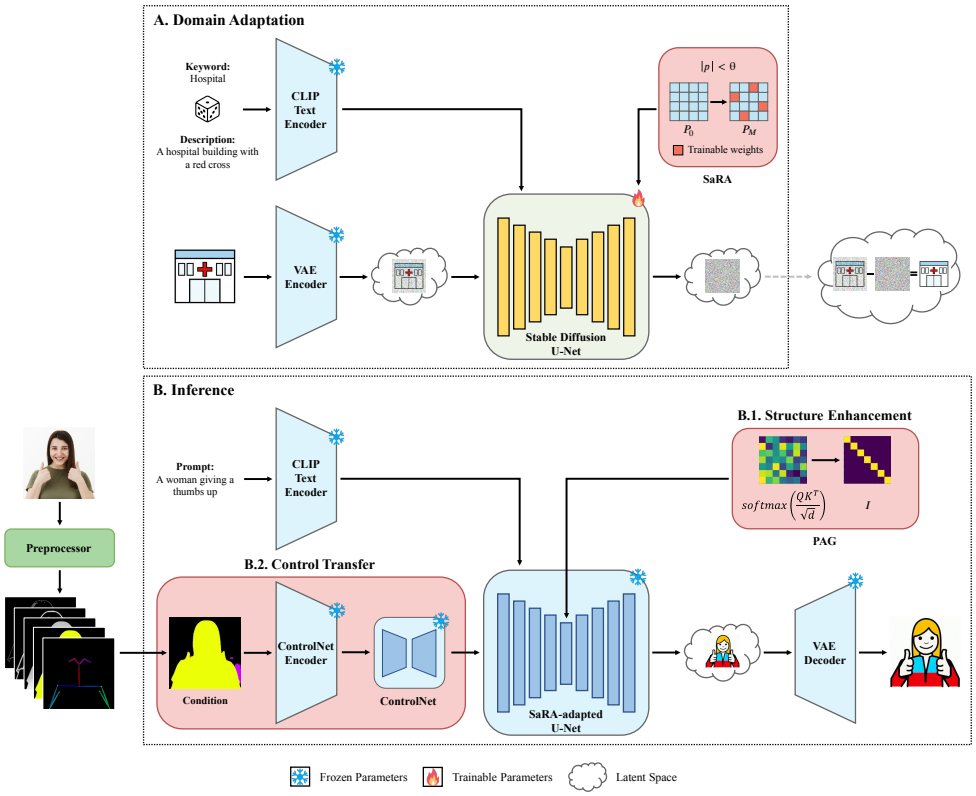

Figure 1: Overview of the proposed PictoSD.

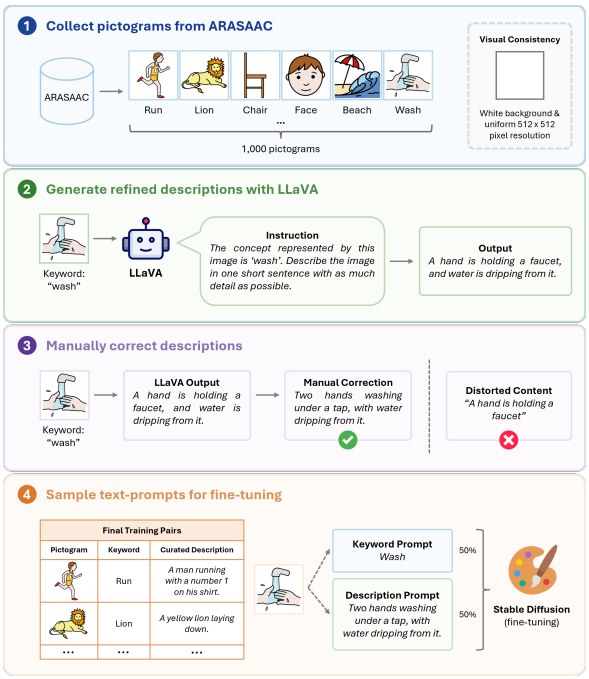

Figure 2: Data preparation procedure for adapting Stable Diffusion to the pictogram domain with SaRA.

terpretable. To address this issue, we integrate Perturbed-Attention Guidance (PAG) (Ahn et al., 2024). PAG modifies the denoising process as:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, c) + \omega(\epsilon_\theta(z_t, c) - \hat{\epsilon}_\theta(z_t, c)). \tag{3}$$

We further apply a Linear Adaptive Scaling (LAS) strategy:

$$\omega_t = \omega_{\text{init}} - s(1000 - t), \tag{4}$$

where ω_{init} is the initial guidance scale and s is the adaptive scaling coefficient. This mechanism helps produce pictograms with cleaner structures and fewer visual artifacts while preserving semantic consistency.

3.3 Image-Guided Control Transfer

Text prompts alone often provide limited control over the spatial composition of generated images (Ruiz et al., 2023; Mou et al., 2024). For pictogram generation, this becomes particularly important, since many concepts require specific spatial arrangements to be easily interpreted. Therefore, we introduce explicit control through ControlNet (Zhang et al., 2023).

ControlNet modifies the output of the U-Net through:

$$y_c = \mathcal{F}(c_t; \Theta) + \mathcal{Z}(\mathcal{F}(c_t + \mathcal{Z}(c; \Theta_{z_1}); \Theta_c); \Theta_{z_2}). \tag{5}$$

This enables finer-grained spatial control during generation. Thus, PictoSD can combine textual intent with visual structure: a keyword or detailed prompt specifies what concept should be represented, while an input image guides how that concept should be visually arranged.

We evaluate our strategy using multiple spatial input conditions, including edge, canny, semantic segmentation, depth, and keypose maps.

4 Experimental Setup

We implement PictoSD with HuggingFace diffusers library (von Platen et al., 2022), and adopt the pre-trained SD v1.5 as our base T2I model. We fine-tune the parameters of its U-Net for 10,000 steps with a batch size of 4 using as training data the 1,000 pictograms collected from ARASAAC and preprocessed following the methodology described in Section 3.1. We utilize AdamW optimizer (Loshchilov and Hutter, 2017) with a cosine learning rate scheduler, where, based on the findings of Hu et al. (2025), the initial learning rate lr is computed according to the threshold θ assigned for SaRA:

$$lr = 10^{-3} \times e^{-350\theta}. \tag{6}$$

For our experiments, we set $\theta = 9 \times 10^{-3}$. During domain adaptation, we recaption the images of the pictogram dataset by adding the prefix '*Pictogram style*' to their corresponding text prompts. In the inference stage, we use the DDIM sampler with 50 steps and CFG set to 7.5. Both training and sampling are performed on a single NVIDIA Titan RTX with 24 GB of memory. Pre-trained ControlNet modules are not fine-tuned, and PAG is applied only to the mid-block of U-Net, following the implementation of Ahn et al. (2024).

Adopting the practice of some works (Zhang et al., 2023; Mou et al., 2024), we calculate the Fréchet Inception Distance (FID) (Heusel et al., 2017) to measure style consistency, and the CLIP Score (Radford et al., 2021) to measure text fidelity. Because FID and CLIP Score tend to exhibit opposite behaviours when T2I models are fine-tuned for specific tasks, we also calculate the Visual-Linguistic Harmony Index (VLHI) (Hu et al., 2025), which measures both style consistency and prompt alignment by normalizing FID and CLIP scores.

5 Experiments

5.1 Main Results: Low-Resource Adaptation and Structure Enhancement

Table 1 shows that for both SaRA and full fine-tuning, the addition of PAG significantly improves the VLHI metric, suggesting that PAG helps to better fit the pictogram distribution (FID) and preserve the prior knowledge of the pre-trained model (CLIP Score). On the other hand, we find that SaRA outperforms full fine-tuning in terms of the VLHI metric regardless of whether or not PAG is included. This validates the effectiveness of SaRA in adapting a pre-trained model to the pictogram domain while preserving semantic consistency under limited training data. For instance, SaRA requires optimizing only 220M parameters, whereas full fine-tuning requires training 860M parameters.

Method	Guidance	FID ↓	CLIP ↑	VLHI ↑
Full Fine-tuning	✗	<u>79.48</u>	28.10	0.68
	PAG	**78.39**	28.16	<u>1.08</u>
SaRA	✗	81.83	**28.87**	1.00
	PAG	80.72	<u>28.80</u>	**1.24**

Table 1: Quantitative comparison of SaRA and full fine-tuning ± PAG on 500 samples. Best and second-best values are bold and underlined, respectively.

The qualitative results in Figure 3 support the quantitative findings. The base SD model without adaptation is unable to accurately reproduce the pictogram style on its own, creating samples with irrelevant details. However, with domain adaptation, considering both full fine-tuning and SaRA, the style consistency improves significantly, allowing the model to produce images with refined structures and simple color schemes. Compared with ARASAAC pictograms, the samples generated by the adapted models have a higher level of detail, containing easily recognizable objects that facilitate the identification of the illustrated concept.

Moreover, it can be seen that the quality of the samples produced by SaRA is quite similar to that of full fine-tuning, despite SaRA being a much less demanding approach in terms of trainable parameters. Figure 4 shows the impact of PAG on the model adapted by SaRA. The application of PAG results in semantically consistent structures with clearer edges and artifacts removed, notably enhancing the visual quality of the samples. These results suggest that PAG helps maintain the simple and interpretable visual structures required for pictogram generation.

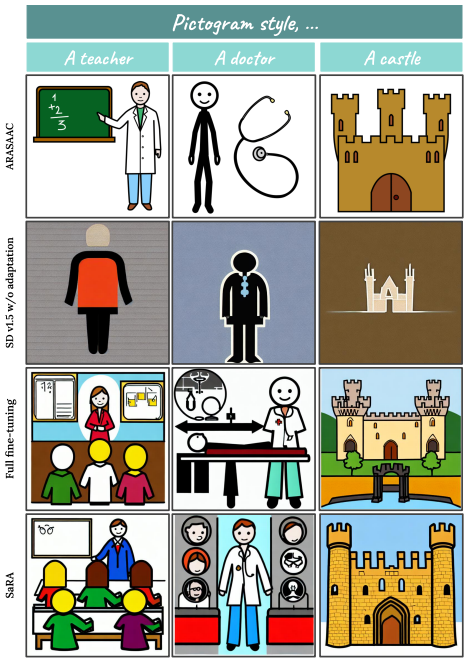

Figure 3: Comparison between ARASAAC pictograms, the base model (SD v1.5), and adapted versions trained with full fine-tuning and SaRA. Both adaptation techniques generate more semantically coherent pictograms than the base model, and achieve a visual quality that is comparable to ARASAAC pictograms.

5.2 Multimodal Control with Input Images

Figure 5 shows the full potential of PictoSD by integrating SaRA, PAG and ControlNet. After providing a control map, which is produced from a reference image, PictoSD generates new images that preserve the original structure of the control map while maintaining the characteristic style of pictograms. This is important because text alone may not be sufficient to describe the desired visual arrangement of a pictogram, especially for actions, abstract concepts, or culturally grounded scenes. Thus, PictoSD can combine textual intent with image-based spatial guidance to generate pictograms that better reflect the intended concept.

We also explore replacing the ControlNet module with the IP-Adapter (Ye et al., 2023), which allows PictoSD to generate pictograms directly from reference images. As shown in Figure 6, integrating the IP-Adapter allows real-world images, such as photographs, to be converted into pictograms in an automated way, which is especially helpful if the user needs to build a personalized visual vocabulary quickly. This further illustrates the flexibility of PictoSD for multimodal pictogram generation from real-world visual inputs.

5.2.1 Culturally Representative Pictograms

We demonstrate the potential of PictoSD by creating a small collection of pictograms that includes culturally representative concepts from Mexico[1]. To build a diverse evaluation set, we first collect 110 concepts using GPT-4o (OpenAI, 2024): 100 common communication concepts (*e.g.*, good, bad, eat) and 10 concepts related to Mexican culture (*e.g.*, axolotl, taco, tequila). For each concept, we obtain a reference image from the internet and generate its corresponding condition maps. Following the data preparation procedure in Section 3.1, we employ LLaVA (Liu et al., 2023) to describe each reference image and manually correct these descriptions to build curated prompts for pictogram generation. We then apply PictoSD to create four pictograms per pair of curated prompt and condition map. For each concept, we select the three best pictograms and rank them through a survey of 7 annotators, who evaluate their visual quality and style adherence using a 3-star scale (on the website, image filename suffixes indicate annotator ranking). In most cases, annotators showed a clear preference for one pictogram over the remaining alternatives, while only nineteen concepts resulted in ties. The annotator rankings suggest that PictoSD is able to generate visually coherent pictograms that preserve both the intended concept and the characteristic pictogram style across culturally grounded scenarios. As Figure 7 shows, PictoSD is able to represent culturally specific concepts while preserving the pictogram style.

5.3 Ablation Studies

We perform ablation studies to better understand the contribution of semantic supervision, low-resource adaptation, and guidance strategies within PictoSD. In particular, we analyze how different forms of textual supervision affect semantic alignment during adaptation, how the number of trainable parameters influences the preservation of semantic priors, and how guidance strategies impact the structural quality of generated pictograms. Considering the same evaluation scheme of Hu et al. (2025), for all cases the FID, CLIP Score and VLHI metrics are evaluated on 500 samples generated by the fine-tuned model with the corresponding modification.

[1]We release the dataset at anonymous.4open.science/r/picto-dataset-F631

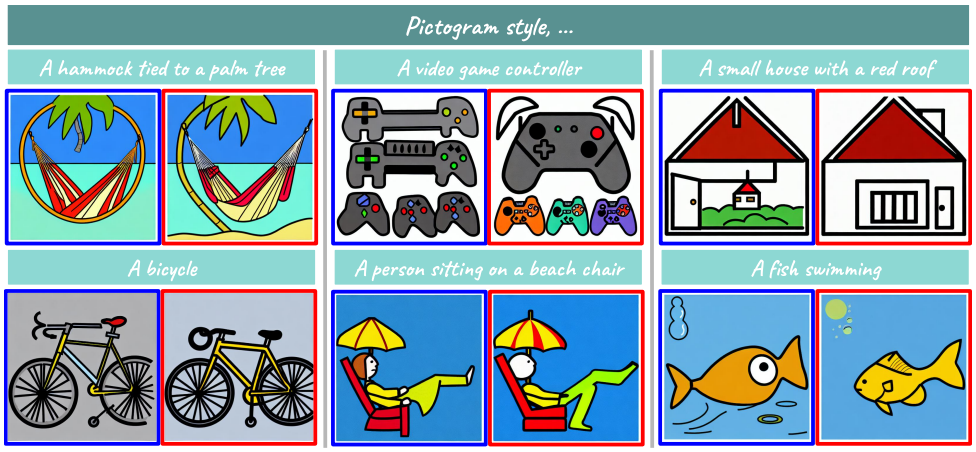

Figure 4: Comparison of pictograms generated by the base model adapted via SaRA without (blue outline) and with (red outline) PAG. Applying PAG leads to more semantically consistent structures and improved visual quality.

Figure 5: Created pictograms with PictoSD. Examples display reference images (orange outline), their corresponding control maps (blue outline), and the pictograms generated by the integration of SaRA, PAG and ControlNet (red outline). PictoSD allows users to produce personalized pictograms from input conditions that reflect desired scenes and structures.

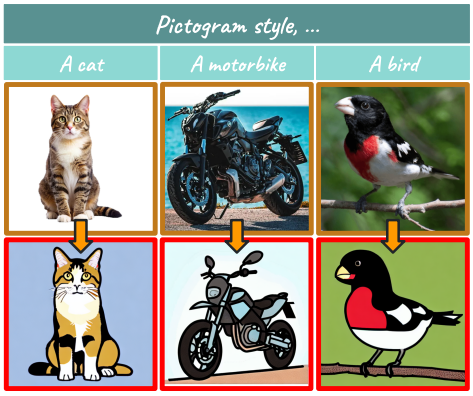

Figure 6: Pictograms generated (red outline) from reference images (orange outline) using PictoSD in conjunction with the IP-Adapter.

5.3.1 Effect of Textual Supervision on Semantic Alignment

We study the impact of different types of textual descriptions on the domain adaptation stage. These include keywords extracted from ARASAAC, short and long descriptions generated by LLaVA (Liu et al., 2023), both manually curated and uncurated, and combinations of keywords with each type of description. By *combinations*, we mean randomly selecting, with equal probability, each pictogram's text prompt from either its keyword or description during model fine-tuning.

As shown in Table 2, combining keywords with any type of description results in better performance compared to using descriptions alone. In particular, when keywords are used together with curated short descriptions, the VLHI metric achieves its highest score. Keywords alone also perform well in terms of VLHI, indicating that the presence of concept words improves both the style and prompt following abilities of the T2I model. These results suggest that combining concise concept-level supervision with richer textual descriptions helps preserve semantic alignment while learning the pictogram style under limited data. Surprisingly, the curated long descriptions show the worst performance among the prompt types, contradicting the premise that longer, more detailed descriptions facilitate the learning of the T2I model. We believe this could be due to a lim-

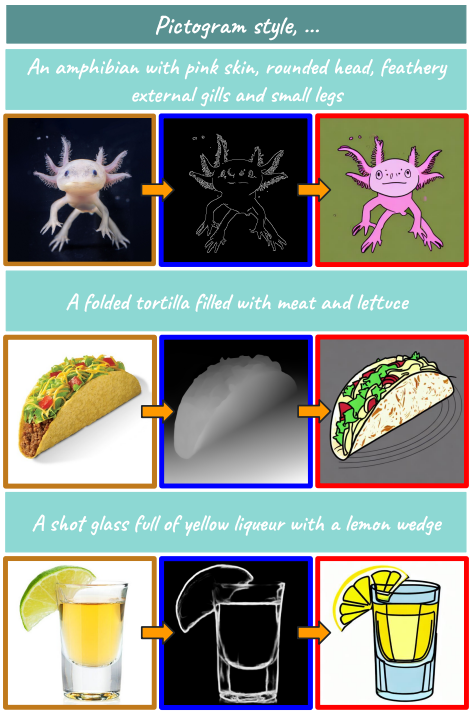

Figure 7: Culturally representative pictograms created with PictoSD. Figures show reference images (orange outline) of Mexican culture, their corresponding control maps (blue outline), and the resulting pictograms produced by PictoSD (red outline).

itation in the implementation of the base model, SD v1.5, which only supports a maximum of 75 tokens per prompt. Additional analyses on trainable parameter selection and guidance strategies are provided in Appendix A.

6 Conclusion

In this work, we propose PictoSD, a low-resource strategy for the automatic generation of pictograms based on a pre-trained Stable Diffusion model. By leveraging SaRA, PictoSD effectively adapts the model to the pictogram domain while preserving semantic consistency under limited training data. In addition, the integration of PAG and ControlNet improves the visual consistency and controllability of the generated pictograms, enabling the combination of textual intent with image-guided spatial structure.

We further demonstrate the potential of PictoSD by creating a collection of 110 generated pictograms, including concepts representative of Mexican culture. To the best of our knowledge, this constitutes the first attempt to create a bank of pictograms representative of Mexico and a step toward more culturally grounded and accessible communication systems for Mexican Augmentative and Alternative Communication (AAC) users. Both quantitative and qualitative results demonstrate the capability of PictoSD to generate visually consistent and semantically coherent pictograms. We believe our work opens new research directions around the use of T2I models to support accessible and culturally grounded visual communication.

Prompt	FID ↓	CLIP ↑	VLHI ↑
K	78.02	26.64	1.50
USD	83.29	26.52	0.59
CSD	84.71	<u>26.70</u>	1.04
ULD	81.23	26.69	1.35
CLD	86.12	26.53	0.37
K + USD	78.62	26.42	0.70
K + CSD	76.52	**26.72**	**1.90**
K + ULD	**75.46**	26.61	<u>1.62</u>
K + CLD	<u>75.63</u>	26.55	1.43

Table 2: Quantitative evaluation of different text prompt types in the model adaptation stage. We compare ARASAAC keywords (K), uncurated (USD) and curated (CSD) short descriptions, and uncurated (ULD) and curated (CLD) long descriptions—each generated by LLaVA—as well as combinations of keywords with each type of description.

Limitations

One limitation of our framework lies in the control transfer stage, where the ControlNet module is sensitive to the conditioning scale. In future work, we plan to explore automatic strategies for selecting this parameter and to evaluate PictoSD through broader user-centered studies in AAC scenarios.

Ethical considerations

Although the pictograms generated by PictoSD are intended to support communication, they are not a substitute for expert-designed pictograms and should be evaluated through user studies before being utilized in real AAC applications. However, the generated pictograms could serve as inspiration for graphic designers and AAC practitioners, making the creative process more flexible and accessible.

Acknowledgements

The pictographic symbols used are property of the Government of Aragón and have been created by Sergio Palao for ARASAAC (http://www.arasaac.org), that distributes them under Creative Commons License BY-NC-SA. During the

preparation of this work, AI assistants were used for grammar correction, orthographic revision, and stylistic polishing of the authors' original content. The manuscript was primarily developed from the authors' own technical material and ideas. The final text was reviewed and edited by the authors, who take full responsibility for its content.

References

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. 2024. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision*, pages 1–17. Springer.

Izadora M.C. Barros, Thaciana S. Alcântara, Alessandra R. Mesquita, Anne Caroline O. Santos, Felipe P. Paixão, and Divaldo P. Lyra Jr. 2014. The use of pictograms in the health care: A literature review. *Research in Social and Administrative Pharmacy*, 10(5):704–719.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*.

Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022a. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems*, volume 35, pages 16664–16678. Curran Associates, Inc.

Yinpeng Chen, Zhiyu Pan, Min Shi, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022b. Design what you desire: Icon generation from orthogonal application and theme labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2536–2546.

Jule Drews, Marco Weißmann, Julian Keil, Frank Dickmann, and Dennis Edler. 2025. A new AI tool for the design of cartographic pictograms (PictoAI) and its potentials for increasing their meaningfulness. *KN-Journal of Cartography and Geographic Information*, 75(1):23–35.

Demi Guo, Alexander Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4884–4896. Association for Computational Linguistics.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. 2023. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. 2025. SaRA: High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *The Thirteenth International Conference on Learning Representations*.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 13753–13773. PMLR.

Jiawei Lin, Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Ting Liu, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. 2024a. IconDM: Text-guided icon set expansion using diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 156–165.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. 2024b. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates Inc.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101v2*.

Ajkel Mino and Gerasimos Spanakis. 2018. LoGAN: Generating logos with a generative adversarial neural network conditioned on color. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 965–970. IEEE.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Daniela Paolieri and Alejandra Marful. 2018. Norms for a pictographic system: The aragonese portal of augmentative/alternative communication (ARASAAC) system. *Frontiers in Psychology*, 9:2538.

Jayr Alencar Pereira, David Macêdo, Cleber Zanchettin, Adriano Lorena Inácio de Oliveira, and Robson do Nascimento Fidalgo. 2022. PictoBERT: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR.

Jose Rodrigo and David Corral. 2013. ARASAAC: Portal aragonés de la comunicación aumentativa y alternativa. Software, herramientas y materiales para la comunicación e inclusión. *Informática na Educação: Teoria & Prática*, 16(2).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510.

Animesh Sinha, Bo Sun, Anmol Kalia, Arantxa Casanova, Elliot Blanchard, David Yan, Winnie Zhang, Tony Nelli, Jiahui Chen, Hardik Shah, Licheng Yu, Mitesh Kumar Singh, Ankit Ramchandani, Maziar Sanjabi, Sonal Gupta, Amy Bearman, and Dhruv Mahajan. 2024. Text-to-sticker: Style tailoring latent diffusion models for human expression. In *European Conference on Computer Vision*, pages 36–52. Springer.

Céline Vaschalde, Pauline Trial, Emmanuelle Esperança-Rodier, Didier Schwab, and Benjamin Lecouteux. 2018. Automatic pictogram generation from speech to help the implementation of a mediated communication. In *Conference on Barrier-free Communication*.

Agustín Vicente, Natàlia Barbarroja, and Elena Castroviejo. 2024. Linguistic, concept and symbolic composition in adults with minimal receptive vocabulary. *Clinical Linguistics & Phonetics*, 38(2):155–171.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`.

Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. 2023. IconShop: Text-guided vector icon synthesis with autoregressive transformers. *ACM Transactions on Graphics (TOG)*, 42(6):1–14.

Hongyi Yang, Chengqi Xue, Xiaoying Yang, and Han Yang. 2021. Icon generation based on generative adversarial networks. *Applied Sciences*, 11(17):7890.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721v1*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

A Additional Ablation Studies

A.1 Effect of Trainable Parameter Selection in Low-Resource Adaptation

Selecting an appropriate threshold when employing SaRA is crucial to effectively adapt the T2I model to the pictogram domain. Therefore, we investigate the effect of varying the threshold in the domain adaptation process. To this end, we test various thresholds set from 4×10^{-5} to 2×10^{-2}, resulting in trainable parameter sizes between 1M and 415M, from a total of 860M parameters.

The results are shown in Table 3. We find that moderate thresholds (in the range of 1×10^{-3} and 9×10^{-3}) provide the best balance between image quality and prompt alignment. Notably, we obtain the best VLHI at a threshold of 9×10^{-3}. These findings indicate that adapting only a moderate subset of parameters is sufficient to learn the pictogram style while preserving the semantic priors of the original T2I model. However, when the threshold

is relatively small (below 8×10^{-4}) or very high (such as 2×10^{-2}), model performance degrades, *i.e.*, the VLHI metric becomes worse. These observations are also corroborated by the T2I samples presented in Figure 8.

Threshold	Params	FID ↓	CLIP ↑	VLHI ↑
4×10^{-5}	1M	85.42	28.16	0.26
2×10^{-4}	5M	85.47	28.21	0.31
4×10^{-4}	10M	82.22	28.18	0.89
8×10^{-4}	20M	83.10	28.30	0.88
1×10^{-3}	25M	82.98	28.20	0.76
2×10^{-3}	50M	82.06	28.51	1.36
4×10^{-3}	100M	**81.33**	28.41	1.38
9×10^{-3}	220M	81.83	**28.87**	**1.90**
2×10^{-2}	415M	86.58	28.13	0.00

Table 3: Comparison results of SaRA at different threshold values. We report FID, CLIP and VLHI scores, along with the number of trainable parameters.

A.2 Effect of Guidance Strategies on Pictogram Structure Preservation

We also evaluate the effectiveness of the LAS strategy by comparing it with two other guidance scaling methods. The first is Classifier-Free Guidance with Rescale (CFGR) (Lin et al., 2024b), a rescaling technique to address the image over-exposure problem commonly faced by CFG. The second is a Norm-based Scaling (NS) approach that we propose to dynamically adjust the intensity of guidance in PAG during sampling.

This method consists of applying a rescaling factor to the guidance scale, which is proportional to the magnitude of the difference between the original noise prediction $\epsilon_\theta(x_t)$ and its perturbed variant $\hat{\epsilon}_\theta(x_t)$. At each denoising step t, the guidance weight is computed using the following equation:

$$\omega_t = \omega_{\text{init}} \times \frac{||\epsilon_\theta(x_t) - \hat{\epsilon}_\theta(x_t)||}{\max\limits_{t \leq r \leq T} ||\epsilon_\theta(x_r) - \hat{\epsilon}_\theta(x_r)||}. \quad (7)$$

As before, ω_{init} represents the initial guidance scale. In addition, we assess the combined effect of CFGR with each of the two scaling methods for PAG.

As shown in Table 4, LAS strategy outperforms the other guidance scaling techniques. The NS approach, on the other hand, shows the worst performance among all alternatives. Moreover, although combining CFGR with LAS or NS leads to better prompt alignment, it harms the sample quality. These observations suggest that progressively reducing guidance intensity during sampling helps preserve the simple and interpretable visual structures required for pictogram generation.

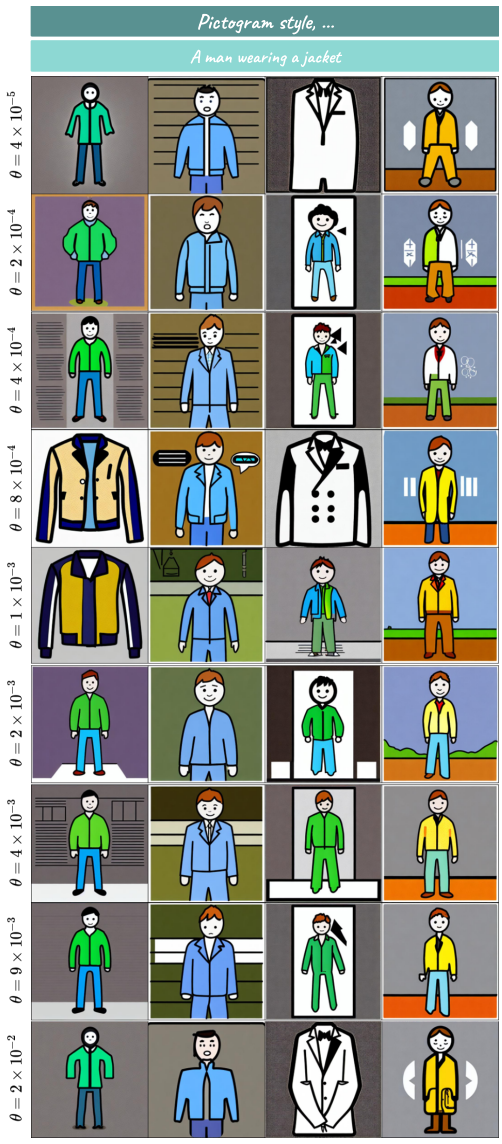

Figure 8: Comparison of samples generated by the base model adapted with SaRA at different thresholds.

Method	FID ↓	CLIP ↑	VLHI ↑
LAS	**80.62**	28.81	**1.30**
NS	82.16	28.75	0.32
CFGR	82.38	**28.93**	1.22
LAS + CFGR	82.65	28.87	0.78
NS + CFGR	82.91	28.85	0.56

Table 4: Quantitative comparison between different guidance scaling techniques. We report results for Linear Adaptive Scaling (LAS), Norm Scaling (NS), Classifier-Free Guidance with Rescale (CFGR), and combinations of the latter with the other two techniques.