AI-Based Representation: Diffusion Models Fine-tuning as a Way of Transformative Operative *Èkphrasis*

Enrico Pupi

Abstract

This study is positioned at the intersection of architectural representation and GenAI (generative artificial intelligence), critically and systematically exploring the transformative potential of fine-tuning DMs (diffusion models). Starting from acknowledging the intrinsic limitations of pure prompt engineering, this research posits that fine-tuning emerges as a contemporary form of transformative operative *èkphrasis*. This approach enables personalised conditioning in the generation of architectural imagery. This contribution critically evaluates their respective potentials and limitations through a comparative analysis of prominent methodologies such as *DreamBooth*, *Textual Inversion*, *LoRA,* and *Hypernetworks*. It also highlights the inherent challenges stemming from the two-dimensional nature of the generated visual representations. Emphasising the necessity for developing evaluation metrics calibrated explicitly for the architectural domain, this research outlines pertinent future research directions. These encompass seamless integration with BIM/CAD environments, the exploration of 3D model generation, and empirical validation through operational experimentation. Fine-tuning is positioned as an enabling technological tool for innovation within architectural representation.

Keywords
Fine-tuning, architectural representation, Diffusion Models, Generative Artificial Intelligence, *èkphrasis*.

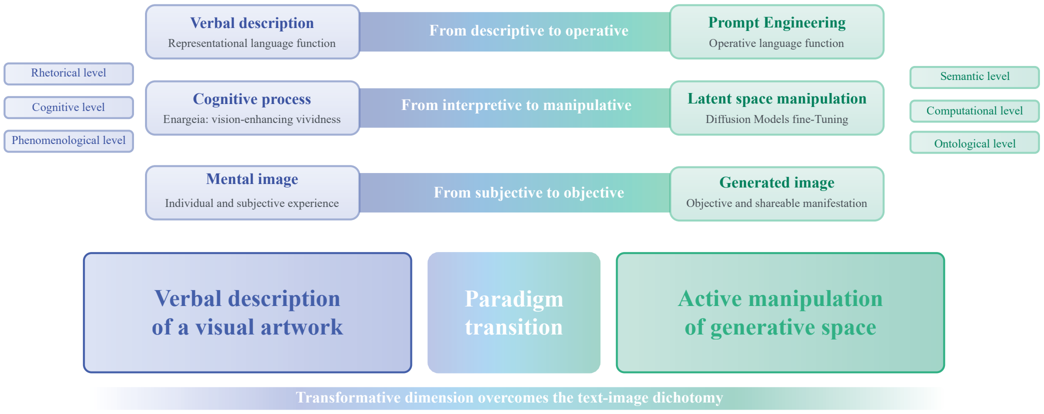

Verbal description
Representational language function
From descriptive to operative
Prompt Engineering
Operative language function
Rhetorical level
Cognitive level
Phenomenological level
Cognitive process
Energeia: vision-enhancing vividness
From interpretive to manipulative
Latent space manipulation
Diffusion Models fine-Tuning
Semantic level
Computational level
Ontological level
Mental image
Individual and subjective experience
From subjective to objective
Generated image
Objective and shareable manifestation
Verbal description of a visual artwork
Paradigm transition
Active manipulation of generative space
Transformative dimension overcomes the text-image dichotomy

Diagram contextualising diffusion models to illustrate the paradigm shift from descriptive verbal ékphrasis of visual artworks to transformative operative ékphrasis in the era of Generative AI. It underscores the transition from descriptive to manipulative cognitive processes (elaboration by the author).

doi.org/10.3280/oa-1430-c920

Introduction

The concept of *èkphrasis* (or *écfrasis*; also *ékphrasis*), traditionally understood as the verbal description of a visual work of art, has evolved complexly over time. Originally, *èkphrasis* was a rhetorical exercise to evoke vivid and sensory descriptions in oratory, transforming listeners into engaged spectators [Webb 1999]. Subsequently, the term shifted to designate the literary representation of a work of art, whether real or imagined. In this sense, *èkphrasis* is configured as a second-order representation, wherein an autonomous image of the world becomes, in turn, the object of representation within a verbal framework, generating a dialectic between equivalence and competition between language and image [Yacobi 2013]. However, a further evolution has propelled *èkphrasis* from a representational to a performative function [Bajohr 2024], focusing on the action of the text rather than the imitation of the image, emphasising the literary response and the causal relationship between text and visual work.

The advent of GenAI (Generative Artificial Intelligence) has further expanded the boundaries of *èkphrasis*, projecting it into the digital domain [Scorzin 2024]. Here emerges operative *èkphrasis*, in which text describes and actively generates images through computational operations. This manifests in two modalities [Bajohr 2024]: sequential operative *èkphrasis*, where code produces an image in which pragmatics prevails over semantics, and connectionist operative *èkphrasis*, typical of multimodal models, where text and image collapse into a shared and multidimensional representational space. As observed by Hannes Bajohr, one might speak of a multimodal pictorial third [Louvel 2018], meaning resides at a higher level, beyond the word and the image. In this instance, semantics prevails over pragmatics, as the model learns the semantic link between textual descriptions and visual manifestations, enabling the generation of images from textual prompts. This scenario suggests a transcendence of the text-image dichotomy, opening towards a novel ontology of *èkphrasis*.

Prompt engineering, the practice of formulating precise instructions to guide GenAI, is configured as a new declination of operative *èkphrasis*: the prompt functions as text, the AI system as a generator, and the generated image is the result. Although substantial differences exist in terms of outcomes (subjectivity of mental experience versus objectivity of the digital image), the analogy is reinforced by considering the rhetorical function of *èkphrasis*, aimed at evoking vivid images, or *enargeia* [Verdicchio 2024].

Within this context, the present contribution intends to analyse how fine-tuning techniques for Diffusion Models (DMs) represent a further contemporary evolution of fundamental importance, particularly in the architectural domain. While comparative analyses have been conducted and benchmarks defined to evaluate the efficacy of fine-tuning techniques in other fields, such as digital character generation [Martini *et al.* 2024], a comprehensive and systematic study for the discipline of architectural representation is still lacking.

State-of-the-art and research scope

The landscape of GenAI has witnessed a progressive and significant evolution. Early implementations, predicated on Variational Autoencoders (VAEs) [Kingma, Welling 2019] and Generative Adversarial Networks (GANs) [Goodfellow *et al.* 2014] while achieving remarkable capabilities in generating visual outputs exhibit distinct characteristics and limitations. VAEs, despite their rapid processing and suitability for broader exploratory endeavours, tend to produce images characterised by blurriness and a lack of precision. GANs are distinguished by their speed and proficiency in generating precise outputs; however, they often prove more efficacious in generating specific instances and less versatile across domains. The subsequent introduction of DMs marked a paradigm shift [Dhariwal, Nichol 2021]. Although comparatively slower, these models amalgamate the precision of GANs with enhanced versatility in exploring the generative space through an iterative process of progressive noise addition and subsequent denoising. They have demonstrated notable capabilities in generating images of exceedingly high quality, affording more effective control over the generation process. They have rapidly become state-of-the-art in the domain of text-to-image generation.

Notwithstanding the advancements achieved by DMs, recent research elucidates that an approach predicated solely on prompt engineering entails significant limitations [Elsharif *et al.* 2025]. When deployed in architectural design, language-based models exhibit consistent challenges in translating complex concepts into efficacious textual instructions [Bolojan *et al.* 2022].

Natural language's inherently ambiguous and polysemous nature complicates translating architectural concepts, spatial relationships, and stylistic nuances into unequivocal and effective textual instructions for DMs. Despite the vast datasets available [Schuhmann *et al.* 2022], the reliance on the model's capacity to correctly interpret the prompt –often influenced by biases inherent in the training data and constrained by the generality of its visual vocabulary– can lead to over-interpretation or under-interpretation of requests, or the generation of visually incoherent artefacts.

The necessity for formulating exceedingly detailed and iterative texts to attain acceptable control over generation reveals the inadequacy of pure prompt engineering as a sufficiently precise instrument for the specific exigencies of architectural design and representation.

In this context, fine-tuning techniques are critical for augmenting the level of conditioning in the inference process and overcoming the limitations of prompt engineering, thereby achieving elevated control in a generation. Fine-tuning consists of adapting a pre-trained model on a specific and targeted dataset, enabling it to learn concepts, styles, and details characteristic of the architectural domain. Through exposure to a dataset of architectural images and text-image pairs, the model acquires a more profound comprehension of architecture's visual and conceptual language. This process facilitates the infusion of domain knowledge into the model, allowing for the generation of images that adhere more faithfully to user intentions and prompt specifications, reducing ambiguity and enhancing the stylistic and conceptual coherence of the output.

Fine-tuning methods analysis

Within the domain of image generation via DMs, fine-tuning constitutes a pivotal nexus in the interaction between language and image, serving as a fundamental lever for adapting these instruments to domain-specific applications. The analysis of principal fine-tuning methodologies is predicated upon a comprehensive review of pertinent scientific literature and critical observations, with particular emphasis on potential practical applications. Currently, comprehensive experimentations within the specific field of architectural representation are lacking. Consequently, the objective is to frame these techniques through a discipline-specific lens, evaluating their applicability and implications for the discipline of representation:

- *DreamBooth* [Ruiz *et al.* 2023]. Operating through a direct modification of the network's internal weights, this technique is distinguished by its capacity to deeply integrate novel concepts through the association of a unique identifier with a specific set of representative images (fig. 1). Its strength lies in the ability to maximise representational fidelity, enabling precise manipulation of model parameters that translates into considerable architectural potential. This technique can excel in preserving specific architectural details, learning distinctive styles, and faithfully reproducing characteristic construction elements, accurately capturing complex stylistic nuances. However, the implementation of *DreamBooth* presents significant challenges, including the elevated computational requirements for training, which may limit its accessibility, and the considerable size of the resulting model, which has implications for storage and sharing. Furthermore, a significant risk of overfitting exists when utilising limited training datasets, necessitating a careful balance between the specificity of the learned concept and the model's capacity for generalisation.
- *Textual Inversion* [Gal *et al.* 2022]. In this instance, a differentiated and non-invasive approach is adopted, concentrating on the optimisation of a novel vector embedding within the latent space of the pre-trained model (fig. 2). Rather than altering the original weights of the model, this technique creates a new embedding vector that represents the desired concept, associating it with a unique identifier. The architectural potential of *Textual Inversion* may manifest in its capacity to represent abstract architectural concepts, codify composi-

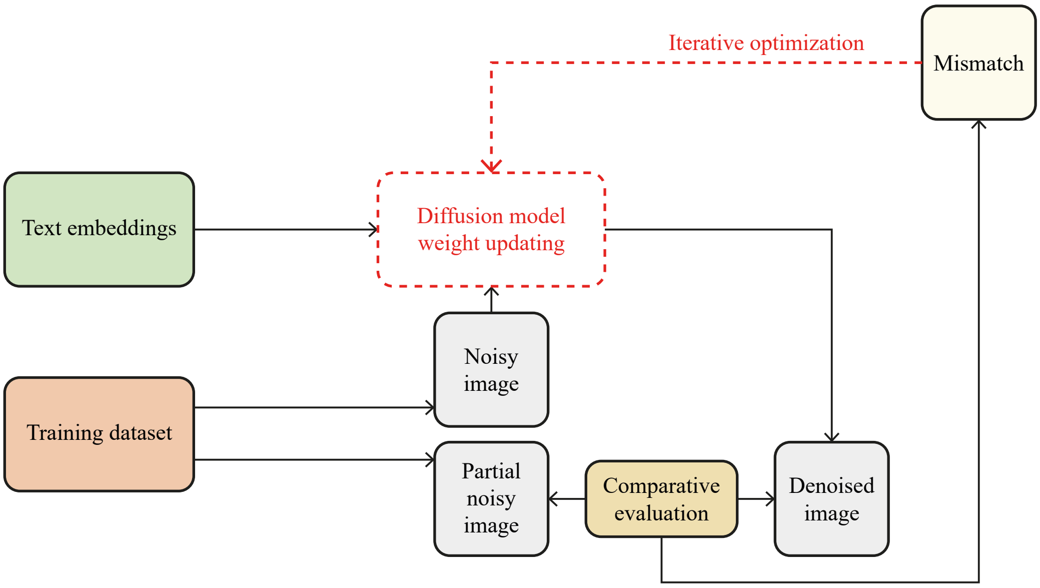


Fig. 1. Schematic diagram of the *DreamBooth* fine-tuning process, illustrating the direct modification of diffusion model weights through iterative optimisation predicated on text embeddings and a representative image training dataset (elaboration by the author).

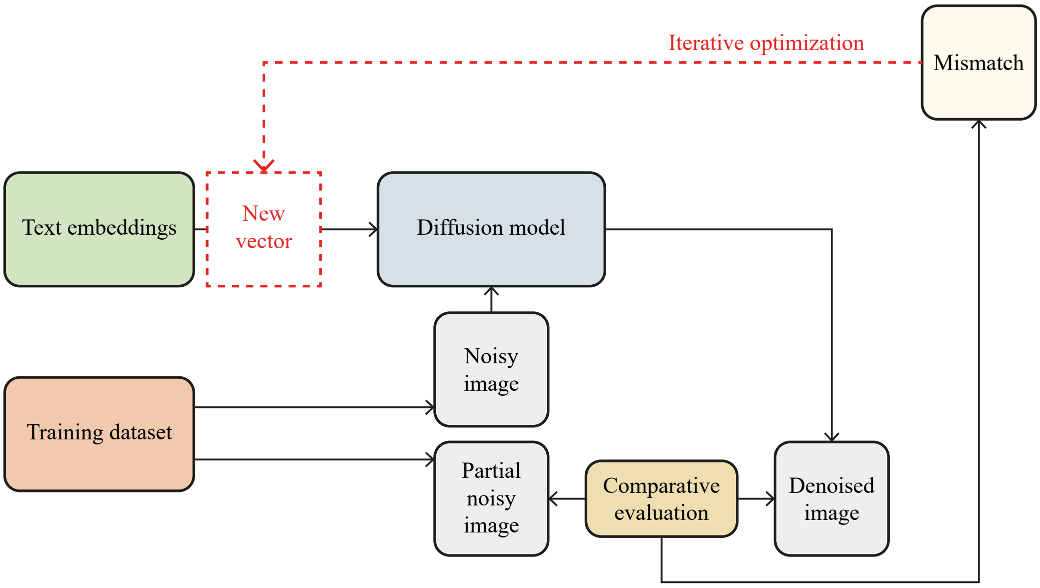


Fig. 2. Schematic diagram of the *Textual Inversion* process, focused on optimising a novel vector embedding within the latent space of a pre-trained diffusion model (elaboration by the author).

tional principles, and integrate spatial qualities. Its inherent nature allows for generating outputs with a contained size, facilitating sharing and implementation. Despite these advantages, *Textual Inversion* may exhibit diminished precision in reproducing specific details compared to *DreamBooth* and frequently necessitates more sophisticated prompt engineering to guide the model in correctly activating the learned embedding, potentially introducing ambiguities in representing particularly complex concepts.

- *Low-Rank Adaptation - LoRA* [Hu *et al.* 2021]. This technique, among the most prevalent, introduces an efficient approach to fine-tuning by appending low-rank matrices to the existing weights of the model (fig. 3), distinguishing itself through its computational efficiency in training and the production of additional layers of contained dimensions. *LoRA* is the original

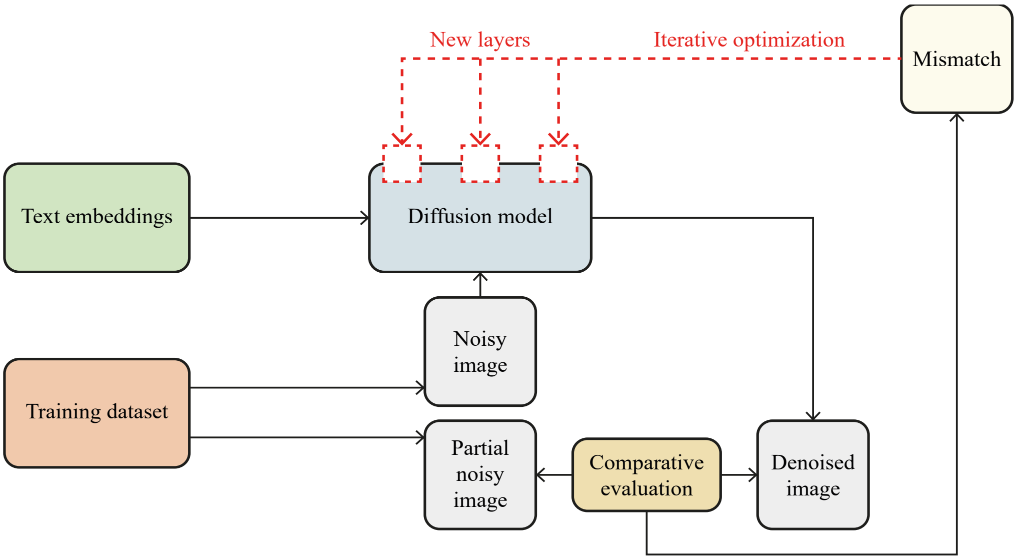


Fig. 3. Schematic diagram of the LoRA (Low-Rank Adaptation) fine-tuning process, highlighting the augmentation of the pre-existing network of the diffusion model with new layers (elaboration by the author).

method, but a similar technique known as *LyCORIS* may exhibit greater expressivity as it captures a more significant number of details using the same training images. Fine-tuning via *LoRA* offers notable architectural potential, enabling rapid adaptation to diverse architectural styles and flexibility in combining stylistic features, achieving an effective equilibrium between precision and efficiency. The capacity to combine and superimpose multiple *LoRAs* opens avenues to complex and nuanced representational scenarios. However, it is important to consider that *LoRA* may present a considerably lower precision than *DreamBooth* in reproducing adaptive architectural details and necessitates the meticulous calibration of training parameters.

- *Hypernetworks* [Ha *et al.* 2016]. This method constitutes an indirect strategy for fine-tuning, wherein an auxiliary neural network generates weights for select layers of the primary model (Fig. 4). This approach offers compelling architectural potential for the exploration of a broad spectrum of design variants and the generation of multiple stylistic interpretations,

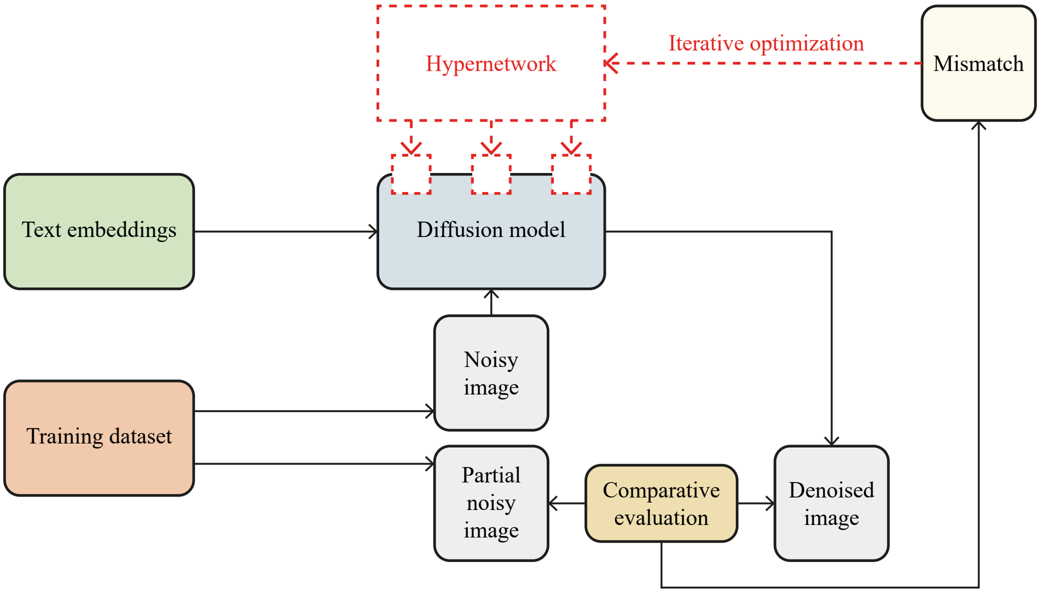


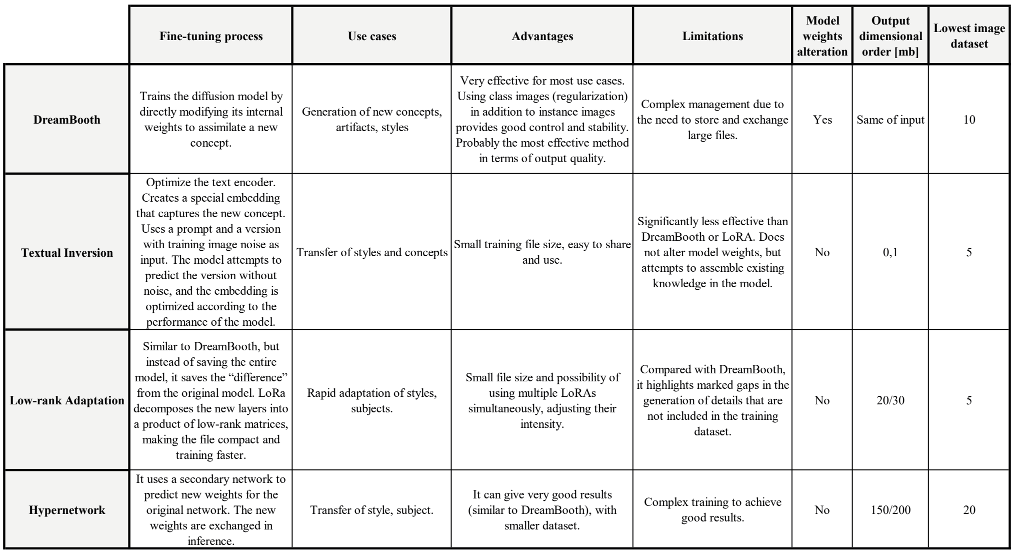
Fig. 4. Schematic diagram of the *Hypernetworks* fine-tuning process. Illustrative schema of fine-tuning via *Hypernetworks*, which employs an auxiliary neural network to generate a portion of the weights for the primary diffusion model (elaboration by the author).

enabling a dynamic manipulation of architectural characteristics. Despite the theoretical flexibility in weight manipulation, *Hypernetworks* exhibits diminished efficiency compared to *LoRA* regarding required computational resources and potentially less consistent result stability. Implementing *Hypernetworks* entails significant complexity, rendering them less accessible for practical and immediate application than other techniques.

Comparative-decisional framework

Analysing the diverse fine-tuning methodologies reveals an intrinsically complex system that necessitates a considered interpretation. While the comparative assessment provides a synoptic overview of the interrelationships between precision, computational efficiency, dimensional footprint, and flexibility (Fig. 5), it is deemed important to acknowledge a tabular representation's inherent limitations in capturing each technique's nuanced complexity. Although a fundamental guiding criterion, the precision/efficiency ratio risks oversimplifying the operative peculiarities.

The analysis conducted thus far constitutes a valuable instrument for preliminary orientation; however, it should not be construed as a univocal prescriptive guide. The selection of the most appropriate methodology for architectural representation mandates an evaluation that, in addition to quantitative comparison, considers the specific qualitative exigencies of the representation's subject matter, the resources available, and the intrinsic nature of the architectural concept intended for visualisation.



	Fine-tuning process	Use cases	Advantages	Limitations	Model weights alteration	Output dimensional order [mb]	Lowest image dataset
DreamBooth	Trains the diffusion model by directly modifying its internal weights to assimilate a new concept.	Generation of new concepts, artifacts, styles	Very effective for most use cases. Using class images (regularization) in addition to instance images provides good control and stability. Probably the most effective method in terms of output quality.	Complex management due to the need to store and exchange large files.	Yes	Same of input	10
Textual Inversion	Optimize the text encoder. Creates a special embedding that captures the new concept. Uses a prompt and a version with training image noise as input. The model attempts to predict the version without noise, and the embedding is optimized according to the performance of the model.	Transfer of styles and concepts	Small training file size, easy to share and use.	Significantly less effective than DreamBooth or LoRA. Does not alter model weights, but attempts to assemble existing knowledge in the model.	No	0,1	5
Low-rank Adaptation	Similar to DreamBooth, but instead of saving the entire model, it saves the "difference" from the original model. LoRa decomposes the new layers into a product of low-rank matrices, making the file compact and training faster.	Rapid adaptation of styles, subjects.	Small file size and possibility of using multiple LoRAs simultaneously, adjusting their intensity.	Compared with DreamBooth, it highlights marked gaps in the generation of details that are not included in the training dataset.	No	20/30	5
Hypernetwork	It uses a secondary network to predict new weights for the original network. The new weights are exchanged in inference.	Transfer of style, subject.	It can give very good results (similar to DreamBooth), with smaller dataset.	Complex training to achieve good results.	No	150/200	20

Fig. 5. Tabular representation synthesising the principal characteristics and trade-offs of the fine-tuning methodologies *DreamBooth*, Textual Inversion, *LoRA*, and *Hypernetworks* (elaboration by the author).

LoRA fine-tuning experiment on authorial design language

To bridge the gap between theoretical analysis and practical application, an empirical experiment was conducted to investigate fine-tuning as a tangible form of operative *èkphrasis*. The objective was to move beyond the general vocabulary of pre-trained models by embedding a specific, authorial design language into the model, thereby transforming it into a bespoke generative tool.

The experiment involved fine-tuning the *Flux-dev* model (weights are released under a non-commercial license by 'Black Forest Lab') using a curated dataset of 25 high-resolution renders from the author's personal architectural portfolio (fig. 6).

Among the available fine-tuning techniques, *LoRA* was chosen for this experiment for three strategic reasons. Firstly, its computational efficiency and the small footprint of the resulting files make it a highly practical and accessible method for designers. Secondly, *LoRA*'s approach of creating a separate adaptation layer offers a perfect balance between specialisation and flexibility. These features make *LoRA* the ideal candidate to demonstrate how a designer can build a library of authorial styles or concepts, treating them as modular generative assets. The *LoRA* model was trained using the dataset, associating the authorial style with the trigger word. Fine-tuning was accomplished through the *Fluxgym Web UI* (fig. 7), which combines a frontend derived from AI-Toolkit with a backend based on *Kohya Scripts*. Microsoft's *Florence-2* visual foundational model was used for the dataset annotation phase. *LoRA* training was conducted with the following parameters: 20 GB of VRAM, 10 repeat trains per image, 16 max train epochs, and dataset images resized to 512 pixels. Total training time: 1 hour, 17 minutes and 53 seconds; average time per iteration: 1.17 seconds; average loss: 0.3. All computational tasks were performed on a workstation with an Intel(R) Core(TM) i9-14900K CPU, 128GB of RAM, and an NVIDIA GeForce RTX 4090 GPU with 24GB GDDR6 VRAM.

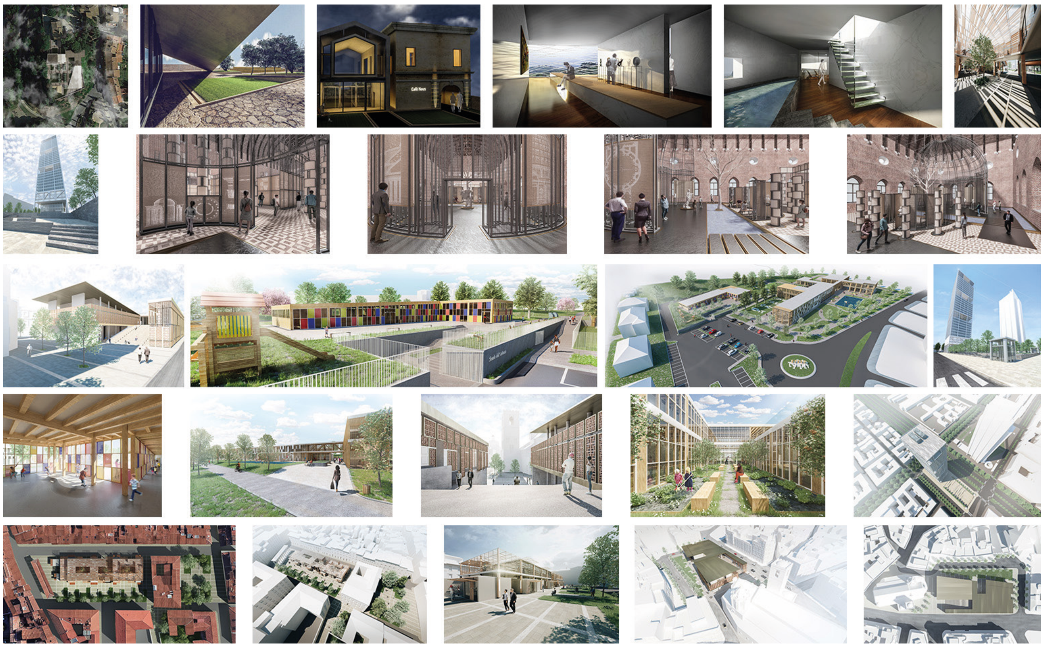
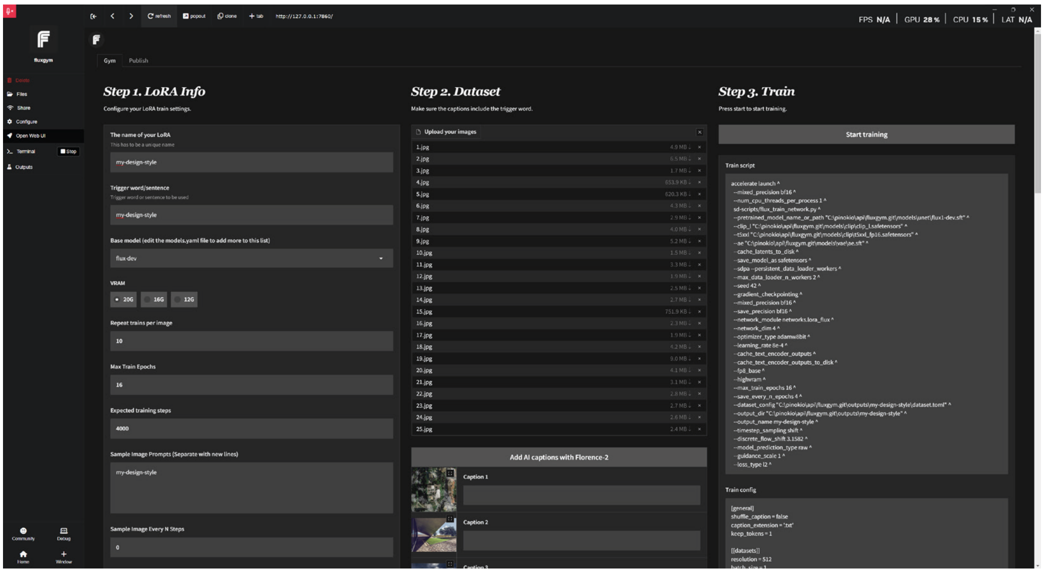


Fig. 6. The curated dataset for the *LoRA* experiment. The images were selected to represent a coherent authorial design language, characterised by specific material palettes, lighting conditions, and spatial compositions (elaboration by the author).



Step 1. LoRA Info

Configure your LoRA train settings.

The name of your LoRA
This has to be a unique name
my-design-style

Trigger word/sentence
Trigger word or sentence to be used
my-design-style

Base model (edit the models.yaml file to add more to this list)
flux-dev

VRAM
● 20G ○ 16G ○ 12G

Repeat trains per image
10

Max Train Epochs
16

Expected training steps
4000

Sample Image Prompts (Separate with new lines)
my-design-style

Sample Image Every N Steps
0

Step 2. Dataset

Make sure the captions include the trigger word.

Upload your images

```
1.jpg   4.9 MB
2.jpg   6.5 MB
3.jpg   1.7 MB
4.jpg   653.9 KB
5.jpg   620.3 KB
6.jpg   4.3 MB
7.jpg   2.9 MB
8.jpg   4.0 MB
9.jpg   5.2 MB
10.jpg  1.5 MB
11.jpg  3.3 MB
12.jpg  1.9 MB
13.jpg  2.5 MB
14.jpg  2.7 MB
15.jpg  751.9 KB
16.jpg  2.3 MB
17.jpg  1.9 MB
18.jpg  4.2 MB
19.jpg  9.0 MB
20.jpg  4.1 MB
21.jpg  3.1 MB
22.jpg  2.8 MB
23.jpg  2.7 MB
24.jpg  2.6 MB
25.jpg  2.4 MB
```

Add AI captions with Florence-2

Caption 1

Caption 2

Caption 3

Step 3. Train

Press start to start training.

Start training

Train script

```
accelerate launch ^
  --mixed_precision bf16 ^
  --num_cpu_threads_per_process 1 ^
  sd-scripts/flux_train_network.py ^
  --pretrained_model_name_or_path "C:\pinokio\api\fluxgym.git\models\unet\flux1-dev.sft" ^
  --clip_l "C:\pinokio\api\fluxgym.git\models\clip\clip_l.safetensors" ^
  --t5xxl "C:\pinokio\api\fluxgym.git\models\clip\t5xxl_fp16.safetensors" ^
  --ae "C:\pinokio\api\fluxgym.git\models\vae\ae.sft" ^
  --cache_latents_to_disk ^
  --save_model_as safetensors ^
  --sdpa --persistent_data_loader_workers ^
  --max_data_loader_n_workers 2 ^
  --seed 42 ^
  --gradient_checkpointing ^
  --mixed_precision bf16 ^
  --save_precision bf16 ^
  --network_module networks.lora_flux ^
  --network_dim 4 ^
  --optimizer_type adamw8bit ^
  --learning_rate 8e-4 ^
  --cache_text_encoder_outputs ^
  --cache_text_encoder_outputs_to_disk ^
  --fp8_base ^
  --highvram ^
  --max_train_epochs 16 ^
  --save_every_n_epochs 4 ^
  --dataset_config "C:\pinokio\api\fluxgym.git\outputs\my-design-style\dataset.toml" ^
  --output_dir "C:\pinokio\api\fluxgym.git\outputs\my-design-style" ^
  --output_name my-design-style ^
  --timestep_sampling shift ^
  --discrete_flow_shift 3.1582 ^
  --model_prediction_type raw ^
  --guidance_scale 1 ^
  --loss_type l2 ^
```

Train config

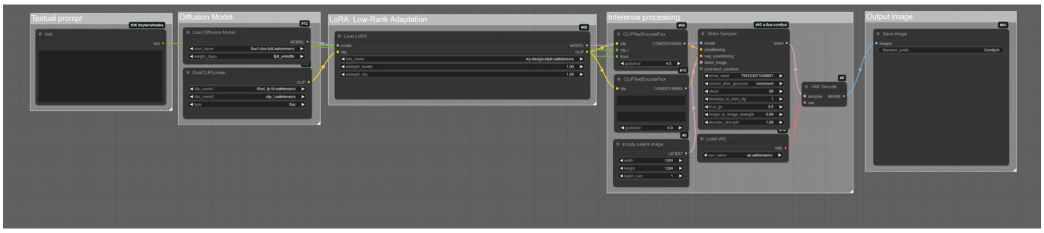
```
[general]
shuffle_caption = false
caption_extension = '.txt'
keep_tokens = 1

[[datasets]]
resolution = 512
batch_size = 1
```

Fig. 7. The user interface of Fluxgym Web UI, used for the *LoRA* training process. The screenshot shows the configuration of key parameters, including the base model (*Flux-dev*), the trigger word (my-design-style), VRAM allocation, and the number of training epochs (elaboration by the author).

ComfyUI, a VPL (Visual Programming Language) environment, was used for the inference phase, with a workflow exploiting the *XLabs-AI* sampler, with 50 steps. Average processing time per image: 1 minute 10 seconds; *LoRA* used: epoch 16 (37.9 mb); strength_model and strength_clip: 1.0 (fig. 8).

The results demonstrate a significant qualitative leap from standard prompt engineering (figs. 9,10). The fine-tuned DM via *LoRA* is able to generate new architectural scenes that not only replicate the surface aesthetics of the training data but also successfully interpret the underlying design principles of the authorial language. By applying the my-design-style trigger to new typologies not present in the original dataset, the model proves its capacity for meaningful generalisation.


Textual prompt
#36 tinyterraNodes
text
text

Diffusion Model
#12
Load Diffusion Model
MODEL
unet_name flux1-dev-fp8.safetensors
weight_dtype fp8_e4m3fn
DualCLIPLoader
CLIP
clip_name1 t5xxl_fp16.safetensors
clip_name2 clip_l.safetensors
type flux

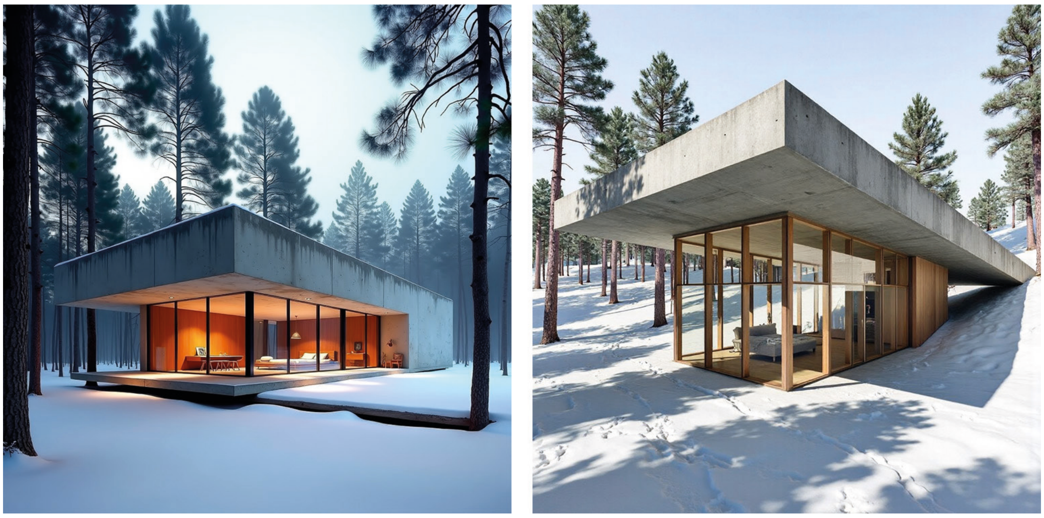
LoRA: Low-Rank Adaptation
#66
Load LoRA
model
clip
MODEL
CLIP
lora_name my-design-style.safetensors
strength_model 1.00
strength_clip 1.00

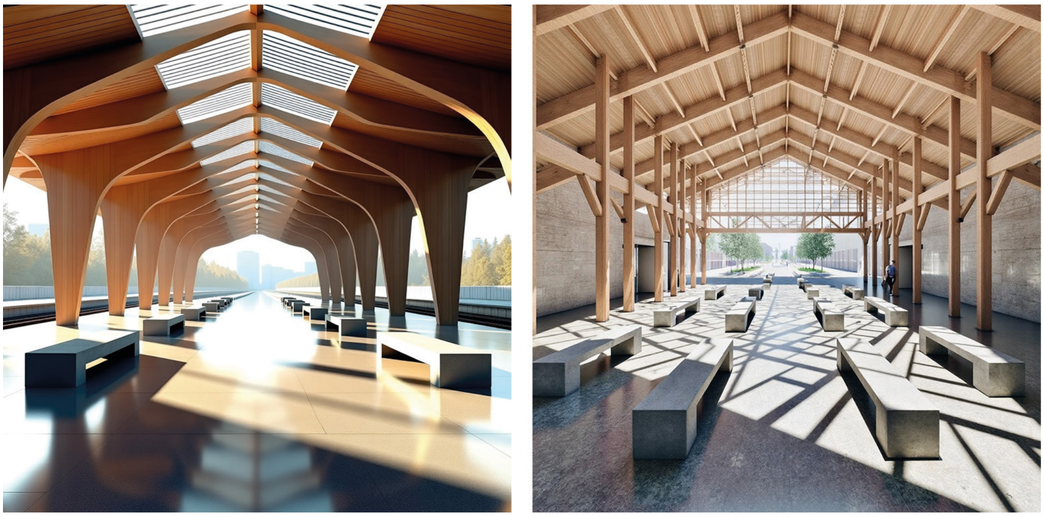
Inference processing
#69
CLIPTextEncodeFlux
clip
clip_l
t5xxl
CONDITIONING
guidance 4.0
#70
CLIPTextEncodeFlux
clip
CONDITIONING
guidance 4.0
#5
Empty Latent Image
LATENT
width 1024
height 1024
batch_size 1
#45 x-flux-comfyui
Xlabs Sampler
model
conditioning
neg_conditioning
latent_image
controlnet_condition
latent
noise_seed 791223071038997
control_after_generate increment
steps 50
timestep_to_start_cfg 1
true_gs 3.5
image_to_image_strength 0.00
denoise_strength 1.00
Load VAE
VAE
vae_name ae.safetensors
#8
VAE Decode
samples
vae
IMAGE

Output image
#64
Save Image
images
filename_prefix ComfyUI

Fig. 8. The inference workflow designed in *ComfyUI*. This visual programming environment was used to generate the test images, showing the sequence of nodes for loading the base model, the fine-tuned *LoRA*, and applying the *XLabs-AI* sampler for the diffusion process (elaboration by the author).

The process transcends a simple instructional dialogue; it becomes a formative act where the 'text' (the curated dataset of the author's work) does not merely describe, but actively reshapes the generative capabilities of the model. The designer, acting as both curator and trainer, sculpts the latent space, embedding their own creative sensibility into the tool. This shifts the paradigm from simply using a generative DM to actively authoring its expressive potential.


Comparative testing the model's ability to apply the authorial style. On the left, the image was generated using the base Flux-dev model. On the right, the image was generated with the LoRA activated, demonstrating a superior adherence to the learned material and spatial language (elaboration by the author).



Generalization testing on a public interior, focusing on structural expression and atmospheric lighting. The base model's output on the left shows a lack of specific stylistic coherence. On the right, the LoRA-enhanced output successfully translates the signature use of lighting (elaboration by the author).

Implications in the manipulation of latent space

Fine-tuning techniques transcend the function of merely connecting verbal description and visual control, giving rise to innovative possibilities in the relationship between language and image within the realm of architectural representation. Whereas prompt engineering operates within the confines of a pre-existing latent space, constrained by the interpretation of the pre-trained model, fine-tuning introduces the capacity to exert an active influence upon this space. As demonstrated by the *LoRA* training experiment, a transition occurs from the description of an image emerging from a predefined visual lexicon to intervention upon the intrinsic structure of the model, refining its sensitivity and orienting its generative capabilities towards a specific architectural language.
From this perspective, the concept of *èkphrasis* in the GenAI era is subject to significant remodelling. It evolves from a connectionist operative *èkphrasis* [Bajohr 2024], wherein text and image converge in a shared representational space, to a form of transformative operative *èkphrasis*. The user is no longer solely configured as an external agent formulating instructions to elicit a visual response, but rather becomes an active manipulator of the latent space. This process, exemplified by the curation of the authorial dataset and the calibration of training parameters, constitutes an act that shapes the DM's expressive potentialities. The fine-tuning process itself assumes an ekphrastic valence, wherein the language of data and configurations directly impacts the modalities of image generation. The fine-tuned model, like the one trained in the experiment, constitutes not merely an instrument but a specifically oriented medium, a more precise and sensitive interpreter of the visual language of architecture. This evolution implies a redefinition of roles and competencies, with the user assuming more direct and conscious control over the creative process, transcending the dichotomy between verbal and visual dimensions through a conjoint action upon the generative space.

Limits and potentials of the analysed techniques

Despite their undeniable transformative potential, fine-tuning techniques are configured as sophisticated instruments that inherit and amplify the complexities intrinsic to DMs. The dependence on the quantity and, more critically, the quality of training data remains an inescapable epistemological constraint [Amershi *et al.* 2019]. An insufficient dataset, or one characterised by inherent biases, would not only compromise the representational fidelity of the model but could insidiously convey stylistic or typological preconceptions, limiting the model's capacity to generalise effectively. Therefore, exploring diverse fine-tuning techniques must be oriented towards a profound understanding of their epistemological affordances, transcending a mere performance-based comparison. Furthermore, the challenge of objective evaluation must be addressed through the development of evaluative protocols situated within the domain of architectural discipline [Ceconello *et al.* 2023]. It is necessary to transcend mere pixel-wise fidelity or superficial stylistic coherence and to conceive metrics that interrogate functional relevance, spatial quality, typological innovation, and even the critical capacity of the generative output. However, it is pertinent to acknowledge that many of the limitations identified thus far may derive from a constraint intrinsic to the two-dimensional nature of the generated image. Architecture, by its essence, manifests in three-dimensional space, and its reduction to a planar representation entails inevitable losses of information and spatial experience. Depth, volumes, complex spatial relationships, and the synesthetic perception of architectural space cannot always be fully captured in a two-dimensional medium.

In this perspective, the emergence of techniques for 3D mesh generation from images based on DMs represents a highly compelling research frontier [Li *et al.* 2024]. Tools such as *Tripo AI*, *Tencent Hunyuan3D*-2.0, *Stable Point Aware 3D* and the more recent *PartPacker* enable overcoming the two-dimensional limitation, paving the way for a spatial architectural representation.The possibility of generating coherent and semantically rich 3D models from textual or visual inputs implies unprecedented opportunities, from volumetric and spatial exploration in the ideation phase to advanced virtual prototyping and even the integration of generative models into BIM workflows or XR (eXtended Reality) environments. The future of generative architectural representation may reside in this transition from image generation to space generation, transforming current limitations into a fertile ground for experimentation and disciplinary innovation.

Conclusions

The intellectual exploration undertaken in this work has demonstrated that the fine-tuning of DMs represents a salient evolutionary step for architectural representation, transcending the intrinsic limitations of pure prompt engineering.This technique is configured as an efficacious bridge between verbal description and visual control, defining a form of transformative operative *èkphrasis* in which the user becomes an active shaper of the model's latent space. For professional practice, fine-tuning paves the way for the development of contextualised design tools capable of learning and adapting to the designer's specific needs, potentially integrating with increasing fluidity within established workflows [Gammal 2024]. For research, fine-tuning is a potent epistemological accelerator, offering a framework for investigating the nature of architectural representation, the relationship between the verbal and the visual, and the implications of GenAI in the creative process.

The issue of economic and technical accessibility to computational resources for fine-tuning, often mediated by paid online tool models, warrants careful consideration to avert new barriers to innovation and experimentation. Furthermore, exploring alternative techniques, such as AttGANs (Attentive Generative Adversarial Networks), may offer complementary or alternative pathways to address specific challenges in generative architectural representation [Del Campo 2021].

Future research directions entail a thorough investigation of the seamless integration of fine-tuned DMs with widely employed BIM/CAD software to explore modalities of deep and bidirectional interoperability that can streamline workflows and maximise the operability of generative tools within the context of architectural representation. Looking beyond the two-dimensional limitation of the image, an area of significant interest is the exploration of 3D models and XR environment generation, opening new frontiers for interactive and immersive spatial representation. Finally, empirical validation and operational experimentation are important for corroborating the theoretical validity and practical efficacy of fine-tuning through rigorous studies and experimentations in real design contexts, utilising instruments such as kohya scripts to identify guidelines and parameters of measurable effectiveness.

The exponential multiplication of possibilities offered by the field of GenAI, coupled with critical reflection and a robust foundation of empirical research, holds the potential to re-

define the relationship between human creativity and computational capacity, opening new perspectives for innovation and epistemological advancement in architecture.
Proficient command of these tools does not necessarily require specialised computer science expertise; multiple approaches and interfaces allow access to significant control over inference processes, moving beyond a merely superficial utilisation. These innovative technological instruments' comprehension and conscious manipulation derive from the ability to orchestrate critical thinking and technical acumen.
It is deemed fundamental that this technological evolution be accompanied by a fertile synergy between technical skills and humanistic sensibilities to ensure that the future of architectural representation in the GenAI era is efficient, performative, ethically conscious, and deeply meaningful for architectural culture.

Reference List

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E. (2019). Guidelines for Human-AI Interaction. In F. Floyd Mueller, P. Kyburz, J.R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, I. Shklovski (Eds.) *Proceedings of the CHI Conference on Human Factors. In Computing Systems*. Glasgow, Scotland Uk, 4-9 May 2019, pp. 1–13. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233.

Bajohr, H. (2024). Operative ekphrasis: the collapse of the text/image distinction in multimodal AI. In *Word & Image*, 40(2), pp. 77–90. https://doi.org/10.1080/02666286.2024.2330335.

Bolojan, D., Vermisso, E., Yousif, S. (2022). Is language all we need? A query into architectural semantics using a multimodal generative workflow. In J. Van Ameijde, N. Gardner, K. Hoon Hyun, D. Luo, U.Sheth (eds.) *Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*. Sydney, Australia 9-15 April 2022,vol. 1, pp. 353–362. Hong Kong: CCAADRIA.

Ceconello, M., Spallazzo, D., Sciannamè, M. (2019). Design and AI: prospects for dialogue. In *Convergências*, XII(23), pp.1-6. http://convergencias.esart.ipcb.pt/?p=article&id=350.

Del Campo, M. (2021). Architecture, language and AI: Language, attentional generative adversarial networks (AttnGAN) and architecture design. In *Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia* CCAADRIA, vol. 1, pp. 211-220. Hong Kong: CADRIA.

Dhariwal, P., Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J. Wortman Vaughan (Eds.) *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Online, December 6-14, 2021, vol. 34, article 672. Red Hook, New York: Curran Associates Inc.

Elsharif, W., Alzubaidi, M., She, J., Agus, M. (2025). Visualizing Ambiguity: Analyzing Linguistic Ambiguity Resolution in Text-to-Image Models. In *Computers*, 14(1), p. 19. https://doi.org/10.3390/computers14010019.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). *An image is worth one word: Personalizing text-to-image generation using textual inversion*. https://arxiv.org/abs/2208.01618.

Gammal, Y. O. E. (2024). The "Cognitive" Architectural Design Process and Its Problem with Recent Artificial Intelligence Applications. In *Engineering and Applied Sciences*, 9(5), pp. 83-105. https://doi.org/10.11648/j.eas.20240905.11.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. Montréal, Québec, Canada, December 8-134, 2014, vol. 2, pp. 2672-2680. Cambridge, MA: MIT Press. https://doi.org/10.5555/2969033.2969125.

Ha, D., Dai, A., Le, Q.V. (2016). *HyperNetworks*. In arXiv. https://arxiv.org/abs/1609.09106.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021). *Lora: Low-rank adaptation of large language models*. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685.

Kingma, D. P., Welling, M. (2019). An Introduction to Variational Autoencoders. In *Foundations and Trends® in Machine Learning,* 12(4), pp. 307-392. https://doi.org/10.1561/2200000056.

Louvel, L. (2018). *The Pictorial Third: An Essay into Intermedial Criticism*. New York: Routledge. https://doi.org/10.4324/9780429485992.

Li, C., Zhang, C., Cho, J., Waghwase, A., Lee, L.-H., Rameau, F., Yang, Y., Bae, S.-H., & Hong, C. S. (2024). *Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era*. arXiv:2305.06131. https://arxiv.org/abs/2305.06131.

Martini, L., Iacono, S., Zolezzi, D., Vercelli, G. V. (2024). Advancing Persistent Character Generation: Comparative Analysis of Fine-Tuning Techniques for Diffusion Models. In *AI*, 5(4), pp. 1779-1792. https://doi.org/10.3390/ai5040088.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K. (2023). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* Vancouver, Canada, June 18-22, 2023, pp. 22500-22510. Los Alamitos, CA: IEEE Computer Society. https://doi.org/10.1109/CVPR52729.2023.02155.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Casanueva, I., Masse, B., Rommel, S., Luther, F., Yang, T. B., Sanh, V., Miller, J., Wortsman, M., Brooker, P., Mullis, H., Brew, C., Jitsev, J. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278-25294.

Scorzin, P. C. (2024). From Descriptive Storytelling to Digital Image Generation with Ai. In *Studi di Estetica*, 28, pp. 21-39.

Verdicchio, M. (2024). Ekphrasis and prompt engineering. A comparison in the era of generative AI. In Studi di Estetica, 28(28), pp. 59-78. https://dx.doi.org/10.7413/1825864661.

Webb, R. (1999). Ekphrasis Ancient and Modern: The invention of a genre. In *Word and Image*, 15(1), pp. 7-18 https://hal.science/hal-01781223.

Yacobi, T. (2013). Ekphrastic Double Exposure and the Museum Book of poetry. In *Poetics Today,* 34(1-2), pp. 1-52. https://doi.org/10.1215/03335372-1894487.

Author
Enrico Pupi , Politecnico di Torino, enrico.pupi@polito.it

To cite this chapter: Enrico Pupi (2025). AI-based Representation: Diffusion Models Fine-tuning as a Way of Transformative Operative *Èkphrasis*. In L. Carlevaris *et al.* (Eds.). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation.* Proceedings of the 46th International Conference of Representation Disciplines Teachers. Milano: FrancoAngeli, pp. 3185-3196. DOI: 10.3280/oa-1430-c920.

ISBN 9788835182412