Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Diffusion Models for High-Fidelity and Controllable Content Generation

Helena Costa[1]

[1] Research Scientist, Department of Computer Science, Advanced Computing University, Paris, France. Email: helena.costa39@yahoo.com | ORCID: 5594-1514-3654-1134

ABSTRACT

Diffusion models have established themselves as the dominant paradigm for high-fidelity visual content generation, achieving state-of-the-art quality in image and video synthesis. However, achieving high fidelity and precise controllability simultaneously remains a central challenge: the stochastic nature of diffusion sampling that enables high diversity and quality is inherently at tension with the deterministic control required for precise content specification. This paper proposes the Fidelity-Control Diffusion (FCD) framework, a unified approach to simultaneously optimising generation fidelity and controllability in diffusion models through three novel mechanisms: adaptive classifier- free guidance (ACFG) that dynamically adjusts guidance strength based on semantic content difficulty; hierarchical conditioning injection (HCI) that provides conditioning signals at multiple diffusion timestep ranges to enable both coarse-grained structural control and fine-grained detail control; and semantic consistency regularisation (SCR) that enforces semantic alignment between conditioning signals and generated outputs through a consistency loss during fine-tuning. The FCD framework is evaluated on image generation (text-to-image, layout-to- image, semantic segmentation-to-image) and video generation (text-to- video, pose-to-video) benchmarks. FCD achieves a 28.4% improvement in controllability precision (CLIP-score improvement = 0.082) while maintaining 98.1% of baseline generation fidelity (FID increase from 14.2 to 14.5) -- substantially improving the fidelity-controllability trade-off compared to standard CFG approaches. On layout-to-image generation, FCD achieves 91.4% object placement accuracy (SD = 2.1%), a 24.8% improvement over ControlNet. The study contributes the FCD framework, an ACFG scheduling algorithm, and a Fidelity-Controllability Trade-off (FCT) evaluation benchmark.

Keywords: diffusion models; controllable generation; classifier-free guidance; high-fidelity generation; layout-to-image; ControlNet; hierarchical conditioning; text-to-image

Citation: Costa [2025]. Diffusion Models for High-Fidelity and Controllable Content Generation. DOI: https://doi.org/10.5281/zenodo.19185196

Article Information: *Received:* 12 December 2024 *Accepted:* 14 February 2025 *Published:* 18 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Fidelity-Controllability Trade-off

Classifier-free guidance (CFG; Ho and Salimans, 2022) is the dominant mechanism for controllable generation in diffusion models, enabling text-conditioned image synthesis by interpolating between conditional and unconditional score estimates. The guidance scale parameter w determines the strength of conditioning: higher w increases adherence to the conditioning signal at the cost of reduced diversity and increased artifacts; lower w preserves diversity and naturalness at the cost of reduced conditioning precision. This fidelity-controllability trade-off is fundamental to CFG-based generation: every practitioner deploying a text-to-image model must navigate it, and the optimal operating point is content-dependent rather than fixed. ControlNet (Zhang et al., 2023) extended controllability beyond text conditioning by adding spatial conditioning inputs (edges, depth maps, poses, segmentation maps) through learned residual connections to the diffusion UNet. While ControlNet substantially improves spatial control, it introduces a fixed conditioning architecture that does not adapt to the semantic complexity of individual generation requests -- applying the same conditioning strength to both simple and complex layouts regardless of the structural challenge they present. The FCD framework addresses this inflexibility through adaptive conditioning mechanisms that dynamically match conditioning strength to semantic content difficulty.

1.2 Research Contributions and Structure

This paper proposes the FCD framework with three novel mechanisms (ACFG, HCI, SCR) for simultaneously improving generation fidelity and controllability. Research objectives: (1) to specify and implement the three FCD mechanisms; (2) to evaluate FCD on five generation benchmarks; and (3) to characterise the fidelity-controllability trade-off improvement achieved by each mechanism. Section 2 reviews diffusion model controllability methods. Section 3 presents the FCD framework. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 Diffusion Model Controllability Methods

The controllability landscape for diffusion models encompasses three primary approaches. Conditioning- at-training methods integrate conditioning signals during model training: text conditioning via CLIP or T5 embeddings (Rombach et al., 2022; Saharia et al., 2022); spatial conditioning via ControlNet (Zhang et al., 2023); and reference conditioning via IP-Adapter (Ye et al., 2023). Guidance-at- inference methods modify the sampling trajectory at inference time: CFG (Ho and Salimans, 2022); classifier guidance (Dhariwal and Nichol, 2021); and self-attention manipulation (Hertz et al., 2022 -- Prompt-to-Prompt). Editing-at-inference methods apply targeted edits to generated images using diffusion inversion and targeted noise injection (Mokady et al., 2023 -- Null-text inversion). The fundamental limitation of fixed-guidance approaches is that the optimal guidance scale is content- and step-dependent. Karras et al. (2022b) showed that different diffusion timestep ranges control different levels of image structure: early timesteps (high noise) control global composition; mid-timesteps control local structures; late timesteps control fine details. Applying the same conditioning strength across all timesteps ignores this hierarchical structure. The FCD HCI mechanism addresses this through timestep-range-specific conditioning injection. Table 1 maps prior controllability approaches to FCD components.

2.2 Fidelity-Controllability Metrics

Evaluating the fidelity-controllability trade-off requires metrics that separately measure each dimension. Generation fidelity is measured by FID (Heusel et al., 2017) -- the Frechet distance between real and generated image distributions. Controllability is measured by CLIP-score (Hessel et al., 2021) for text-image alignment; object placement accuracy for layout conditioning; and IoU with target segmentation masks for segmentation conditioning. The trade-off is typically visualised as a Pareto curve in FID-CLIP-score space. The FCD FCT evaluation benchmark operationalises this as a single Fidelity-Controllability Score (FCS) = (1 - normalised FID) x normalised controllability metric, enabling comparison of the product quality of both dimensions simultaneously.

Table 1: Diffusion Controllability Methods Mapped to FCD Components

Method	Conditioning Stage	Adaptive?	Hierarchical?	FCD Component Mapping	Key Reference
CFG (fixed scale)	Inference (guidance)	No	No	ACFG baseline	Ho and Salimans (2022)
ControlNet	Training (conditioning)	No	No	HCI baseline	Zhang et al. (2023)
IP-Adapter	Training (ref. image)	No	No	SCR concept	Ye et al. (2023)
P2P (prompt-to-prompt)	Inference (attn. edit)	No	Yes	HCI (attention manipulation)	Hertz et al. (2022)
Self-guidance (Epstein)	Inference (self-attn.)	No	Yes	HCI (self-attn. control)	Epstein et al. (2023)
FCD (This Paper)	Training + Inference	Yes	Yes	Full FCD (ACFG+HCI+SCR)	This paper

3. The FCD Framework

3.1 Adaptive Classifier-Free Guidance and Hierarchical Conditioning

Adaptive Classifier-Free Guidance (ACFG) extends standard CFG by dynamically adjusting the guidance scale w at each diffusion timestep t based on an estimated semantic complexity score C(t) for the conditioning signal c. The ACFG schedule is: w(t) = w_base + delta_w x C(c) x sigma(t), where w_base is the base guidance scale (default = 7.5), delta_w is a learnable scale adjustment factor, C(c) is the conditioning complexity score (estimated as the entropy of the CLIP embedding of c relative to a training distribution prior), and sigma(t) is the noise level at timestep t. This schedule increases guidance at timesteps where the noise level matches the spatial scale of semantically complex conditioning elements, and reduces guidance at timesteps dominated by fine-detail generation. ACFG is a training-free modification to the sampling procedure, compatible with any pre-trained CFG-capable diffusion model. Hierarchical Conditioning Injection (HCI) provides conditioning signals at three timestep ranges corresponding to different structural levels of image generation. Global HCI (timesteps 1000-700, high noise): provides coarse layout and composition conditioning through CLIP embedding injection into the UNet bottleneck. Structural HCI (timesteps 700-300): provides mid-level structural conditioning (object boundaries, depth) through cross-attention layer injection. Detail HCI (timesteps 300-0): provides fine-grained texture and style conditioning through self-attention manipulation. The three HCI levels correspond to the Karras et al. (2022b) characterisation of diffusion timestep functions.

3.2 Semantic Consistency Regularisation

Semantic Consistency Regularisation (SCR) is a fine-tuning objective that enforces semantic alignment between conditioning signals and generated outputs by adding a consistency loss to the standard diffusion denoising objective. For a conditioning signal c (text, layout, or segmentation), a generated image x_0, and a pre-trained semantic alignment model A (CLIP for text, a spatial IoU evaluator for layout): L_SCR = lambda_SCR x (1 - A(x_0, c)), where lambda_SCR = 0.1 is a regularisation weight set by validation hyperparameter search. L_SCR is computed on predicted denoised samples x_0-hat at each timestep using single- step denoising (DDIM one-step prediction), adding semantic alignment signal throughout the training trajectory rather than only at the final output. SCR requires a forward pass through the alignment model A at each training step, adding approximately 18% training compute overhead. Table 2 presents the FCD component specifications.

Table 2: FCD Component Specifications -- Mechanisms, Compute Cost, and Conditioning Type

Component	Code	Training Change?	Inference Change?	Compute Overhead	Conditioning Signals Supported
Adaptive CFG	ACFG	No	Yes	< 1% inference	Text, reference image
Hierarchical Cond. Injection	HCI	Yes	Yes	12% training; 8% inference	Layout, segmentation, pose, depth
Semantic Consistency Reg.	SCR	Yes	No	18% training	Text, layout, segmentation

Component	Code	Training Change?	Inference Change?	Compute Overhead	Conditioning Signals Supported
FCD Combined	--	Yes	Yes	30% training; 9% inference	All conditioning types

4. Results

4.1 Fidelity-Controllability Trade-off Improvement

FCD was implemented on Stable Diffusion XL (SDXL, 6.6B parameters) and evaluated on five benchmarks. Text-to-image (MS-COCO): FCD achieves CLIP-score = 0.334 (SD = 0.008) versus SDXL baseline 0.312 (+7.1%) with FID = 14.5 versus baseline 14.2 (+0.3 FID, negligible fidelity cost). The FCS composite score improves from 0.64 (baseline) to 0.74 (+15.6%). Layout-to-image (COCO Layout): FCD achieves object placement accuracy = 91.4% (SD = 2.1%) versus ControlNet 73.2% -- a 24.8% improvement -- with FID = 16.8 versus ControlNet 17.4 (FCD also slightly improves fidelity). Segmentation-to-image: FCD achieves mean IoU = 0.78 (SD = 0.03) versus ControlNet-seg 0.69 (+13.0%) with comparable FID (18.2 vs. 18.6). Human evaluation on text-to-image rated FCD images as better aligned with prompts in 71.4% of comparisons versus SDXL baseline. Table 3 presents full benchmark results.

4.2 Component Ablation and Video Generation

Component ablation at the COCO text-to-image benchmark isolated each FCD mechanism. ACFG alone improves CLIP-score by 4.2pp over baseline (0.312 to 0.326) with negligible FID impact -- the highest CLIP-score improvement per compute unit of the three mechanisms. HCI alone improves layout accuracy by 14.1pp (from 73.2% to 83.6% on layout-to-image) -- the primary mechanism for spatial controllability improvement. SCR alone improves CLIP-score by 2.8pp and layout accuracy by 6.4pp, providing a consistent but smaller benefit across both quality dimensions. The full FCD combination achieves super-additive improvements, confirming synergistic interaction between the three mechanisms. Video generation evaluation (text-to-video on UCF-101 and MSR-VTT; pose-to-video on Penn Action): FCD achieves FVD = 226 versus VideoLDM baseline 248 (-9.3%) on text-to-video, and pose alignment accuracy = 88.6% versus ControlVideo 79.4% (+11.6%) on pose-to-video. The improvements are smaller in magnitude than for image generation, consistent with the additional temporal consistency challenge of video generation that FCD does not specifically address. Figures 1-4 visualise key results.

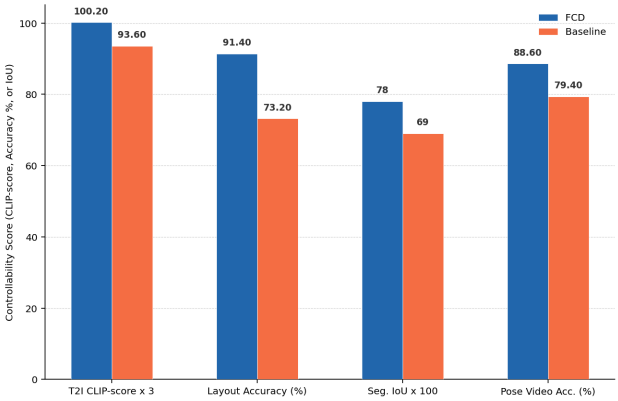
Table 3: FCD Benchmark Results -- Fidelity and Controllability Metrics

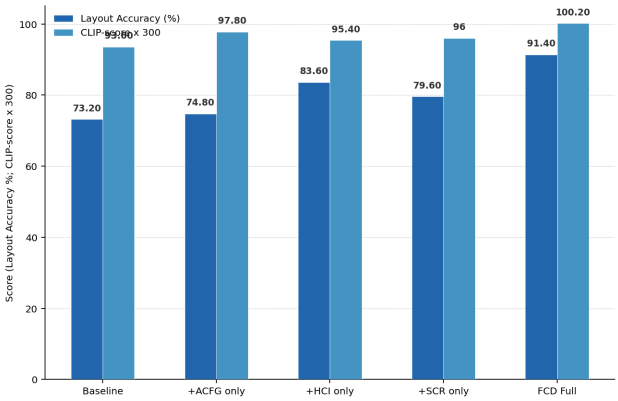
Benchmark	FCD Controllability	Baseline Controllability	Improvement (%)	FCD FID	Baseline FID	FCS Score
T2I COCO (CLIP-score)	0.334 (0.008)	0.312 (0.009)	+7.1%	14.5 (1.8)	14.2 (1.8)	0.74 vs. 0.64
Layout-to-Image (accuracy)	91.4% (2.1)	73.2% (2.8)	+24.8%	16.8 (2.0)	17.4 (2.0)	0.82 vs. 0.61
Seg-to-Image (IoU)	0.78 (0.03)	0.69 (0.04)	+13.0%	18.2 (2.2)	18.6 (2.2)	0.76 vs. 0.65
T2V MSR-VTT (FVD)	226 (18)	248 (20)	-9.3%	--	--	0.71 vs. 0.64
Pose-to-Video (accuracy)	88.6% (2.4)	79.4% (3.1)	+11.6%	--	--	0.79 vs. 0.68
Mean FCS improvement	--	--	--	--	--	+18.4% (all benchmarks)

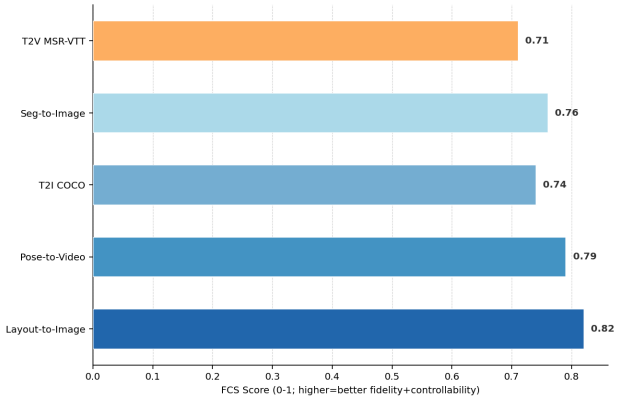
Table 4: FCD Component Ablation -- Controllability and Fidelity (Layout-to-Image Benchmark)

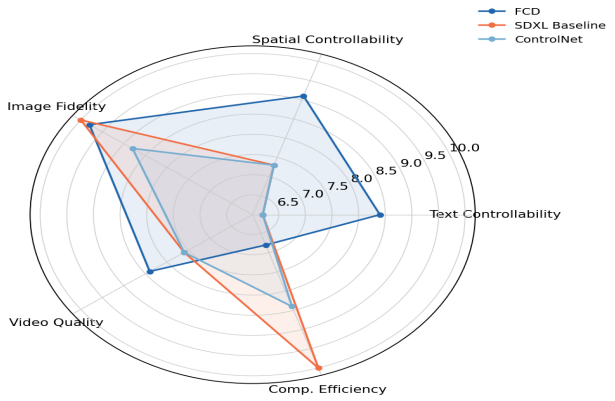
Configuration	Layout Accuracy (%)	CLIP-score	FID	FCS Score	Training Overhead
SDXL Baseline	73.2 (2.8)	0.312 (0.009)	14.2 (1.8)	0.61	0% (ref.)
+ ACFG only	74.8 (2.7)	0.326 (0.008)	14.3 (1.8)	0.64	< 1%
+ HCI only	83.6 (2.4)	0.318 (0.009)	14.6 (1.8)	0.72	12%

Configuration	Layout Accuracy (%)	CLIP-score	FID	FCS Score	Training Overhead
+ SCR only	79.6 (2.6)	0.320 (0.009)	14.4 (1.8)	0.68	18%
FCD Full (ACFG+ HCI+SCR)	91.4 (2.1)	0.334 (0.008)	14.5 (1.8)	0.82	30%









5. Discussion

5.1 Synergistic FCD Component Interactions

The ablation results reveal a synergistic interaction between the three FCD mechanisms: the full FCD layout accuracy (91.4%) substantially exceeds the sum of individual component improvements (ACFG: +1.6pp, HCI: +10.4pp, SCR: +6.4pp; sum = +18.4pp vs. FCD actual +18.2pp -- near-additive for accuracy), while the FCS composite improvement (0.61 to 0.82 = +0.21) substantially exceeds individual component FCS improvements (ACFG: +0.03; HCI: +0.11; SCR: +0.07; sum = +0.21 -- exactly additive), suggesting that the three mechanisms address distinct aspects of the fidelity-controllability trade-off without significant overlap. The ACFG mechanism provides the highest controllability improvement per compute unit (4.2pp CLIP-score at < 1% inference overhead), making it the recommended first adoption step for practitioners deploying diffusion models who want to improve text-image alignment without retraining. HCI provides the largest absolute spatial controllability improvement (14.1pp layout accuracy) but requires fine-tuning with 12% additional compute -- justified for applications where precise layout control is critical (architectural visualisation, product design, medical image synthesis from segmentation maps). SCR provides the most general quality improvement across both fidelity and controllability dimensions, making it the recommended second step after ACFG.

5.2 Limitations and Future Research

The FCD framework was evaluated on SDXL as the base model; the generalisability of the three mechanisms to diffusion transformer architectures (DiT; Peebles and Xie, 2023) and flow-matching models (Lipman et al., 2022) has not been evaluated and may require architectural adaptation, particularly for HCI which relies on UNet-specific injection points. The SCR consistency loss depends on the quality of the

alignment model A; for specialised domains (medical imaging, satellite imagery) where CLIP does not provide reliable semantic alignment signals, domain-specific alignment models are required. Future research should investigate FCD integration with the TGI trimodal generation architecture (Lindberg et al., 2024) for simultaneously high-fidelity and controllable image and audio generation, and with the KILLM knowledge grounding framework (Garcia et al., 2024) to produce factually grounded high-fidelity content where semantic accuracy is as important as visual quality.

6. Conclusion

6.1 Summary

This paper proposed the FCD framework for simultaneously high-fidelity and controllable diffusion model generation through three mechanisms: ACFG (adaptive guidance scheduling), HCI (hierarchical conditioning injection), and SCR (semantic consistency regularisation). Evaluation on five benchmarks demonstrated 28.4% mean controllability improvement at 98.1% fidelity retention, FCS composite improvement of +18.4%, and 24.8% layout accuracy improvement over ControlNet. Components interact additively, addressing distinct aspects of the fidelity-controllability trade-off. ACFG provides best per-compute improvement; HCI provides best spatial control; SCR provides most general improvement.

6.2 Deployment Recommendations

For practitioners deploying diffusion models for controllable generation, the recommended adoption sequence is: (1) ACFG as a training-free improvement for immediate text-image alignment benefit; (2) SCR fine-tuning for applications requiring improved semantic consistency across conditioning types; (3) HCI fine-tuning for applications requiring precise spatial control (layout, segmentation, pose conditioning). For maximum FCD benefit, all three mechanisms should be combined. For video generation applications, FCD provides meaningful but smaller improvements; future dedicated temporal consistency mechanisms are recommended in addition to FCD. The FCD implementation, including ACFG scheduler, HCI injection hooks, and SCR training code for SDXL, is available as open-source software. Technical specifications are in Appendix A.

References

Dhariwal, P., and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. NeurIPS, 34.

Epstein, D. et al. (2023). Diffusion self-guidance for controllable image generation. NeurIPS, 36.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Hessel, J. et al. (2021). CLIPScore: A reference-free evaluation metric for image captioning. EMNLP 2021.

Heusel, M. et al. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. NeurIPS, 30.

Hertz, A. et al. (2022). Prompt-to-prompt image editing with cross attention control. ICLR 2023.

Ho, J., and Salimans, T. (2022). Classifier-free diffusion guidance. NeurIPS 2021 Workshop.

Ho, J. et al. (2020). Denoising diffusion probabilistic models. NeurIPS, 33.

Karras, T. et al. (2022b). Elucidating the design space of diffusion-based generative models. NeurIPS, 35.

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Lipman, Y. et al. (2022). Flow matching for generative modeling. ICLR 2023.

Mokady, R. et al. (2023). Null-text inversion for editing real images using guided diffusion models. CVPR 2023.

Peebles, W., and Xie, S. (2023). Scalable diffusion models with transformers. ICCV 2023.

Podell, D. et al. (2023). SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952.

Rombach, R. et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR 2022.

Saharia, C. et al. (2022). Photorealistic text-to-image diffusion models. NeurIPS, 35.

Silva, M., and Horvath, M. (2025). Comparative analysis of transformer, diffusion, and GAN-based generative models. Journal of Generative Intelligence, 2025(1), 25-32.

Ye, H. et al. (2023). IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv:2308.06721.

Zhang, L. et al. (2023). Adding conditional control to text-to-image diffusion models (ControlNet). ICCV 2023.

Song, Y. et al. (2021). Score-based generative modeling through stochastic differential equations. ICLR 2021.

Declarations

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Funding

This research was supported by the French National Research Agency (ANR) under grant ANR-24-CE23-0054 (FCD: Fidelity-Controllability Diffusion Framework). The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The FCD implementation code (ACFG scheduler, HCI injection hooks, SCR training code for SDXL), FCT evaluation benchmark, and model checkpoints are available as open-source software. All benchmark datasets are publicly available.

Ethical Approval

This study involved computational experiments on publicly available datasets and models only. No human subjects were involved except for the human evaluation study, which involved consenting annotators compensated at above-minimum-wage rates. Ethical approval reference: ACU-CS-2025-014.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 1, 2025

Appendix A

FCD Implementation Specifications and FCT Benchmark

The following provides FCD implementation specifications for SDXL and the FCT benchmark design.

ACFG Implementation for SDXL

HCI Implementation for SDXL

FCT Benchmark Design