cairo.thws
Center for Artificial Intelligence
Technical University of
Applied Sciences
Würzburg-Schweinfurt

UNIVERSITÄT
BIELEFELD

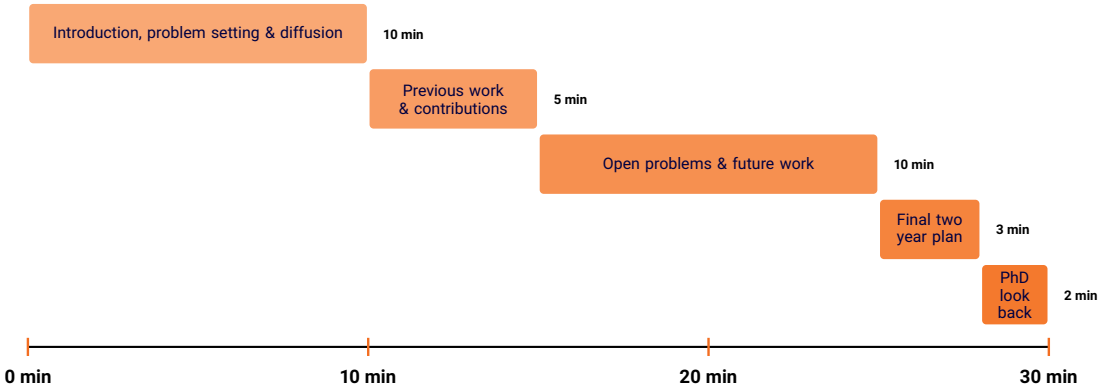
Training-Free Guidance in Diffusion Models: From Classifier Gradients to Future Control Mechanisms

Philipp Vaeth (philipp.vaeth@thws.de)

17-18 June 2026

DMML Workshop Presentation / PhD Exam

Overview and road map



Bayesian setting: sampling from a complex posterior

With $\boldsymbol{x} \sim p_{\text{data}}$ and $\boldsymbol{y}$ as a conditioning variable (label, class, measurement,...), we want to draw samples by:

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{desired posterior}} \propto \underbrace{p_\theta(\boldsymbol{x})}_{\text{learned prior}} \underbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}_{\text{test-time likelihood}} .$$

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x} \mid \boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y} \mid \boldsymbol{x}) .$$

Why not learn the posterior directly?

- **Complex conditioning:** The condition $\boldsymbol{y}$ may be unknown during training or change after training; the likelihood $p(\boldsymbol{y} \mid \boldsymbol{x})$ can be expensive to compute (for example for forward models or simulations).
- **Multiple conditions:** Multiple independent likelihood functions can be applied during sampling:

  $$\nabla_{\boldsymbol{x}} \log p\left(\boldsymbol{x} \mid y_1, \ldots, y_k\right) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \sum_{i=1}^{k} \nabla_{\boldsymbol{x}} \log p\left(y_i \mid \boldsymbol{x}\right) \quad \Longleftrightarrow \quad y_1 \perp y_2 \ldots \perp y_k \mid \boldsymbol{x}$$

- **Flexibility & Efficiency:** The prior $p_\theta(\boldsymbol{x})$ is amortized over many conditions; the likelihood $p(\boldsymbol{y} \mid \boldsymbol{x})$ is amortized over different prior models; plug-and-play.

Diffusion models: score-based overview

Forward SDE

$$\mathrm{d}\boldsymbol{x}_t = f_t(\boldsymbol{x}_t)\,\mathrm{d}t + g_t\,\mathrm{d}\boldsymbol{w}_t, \qquad \boldsymbol{x}_0 \sim p_0, \quad \boldsymbol{x}_T \approx \mathcal{N}(0, \mathbf{I}) \ .$$

Reverse SDE

$$\mathrm{d}\boldsymbol{x}_t = \left[f_t(\boldsymbol{x}_t) - g_t^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t + g_t\,\mathrm{d}\bar{\boldsymbol{w}}_t, \qquad T \to 0, \quad s_\theta(\boldsymbol{x}, t) \approx \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) \ .$$

Denoising score matching objective

$$\theta^* = \arg\min_\theta \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{x}_t}\left[\lambda(t)\left\|s_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log p_{0t}(\boldsymbol{x}_t \mid \boldsymbol{x}_0)\right\|_2^2\right], \qquad \boldsymbol{x}_t \sim \mathcal{N}\left(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t^2 I\right) \ .$$

Sources: Ho et al., 2020; Song et al., 2021a,b; Karras et al., 2022; Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023

Bayesian guidance: prior score to posterior score

Prior reverse SDE

$$\mathrm{d}\boldsymbol{x}_t = \left[f_t(\boldsymbol{x}_t) - g_t^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] \mathrm{d}t + g_t \, \mathrm{d}\bar{\boldsymbol{w}}_t \quad .$$

Score decomposition

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{y}) = \underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)}_{\text{diffusion prior score}} + \underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)}_{\text{test-time likelihood score}} \quad .$$

Guided posterior reverse SDE

$$\mathrm{d}\boldsymbol{x}_t = \left[f_t(\boldsymbol{x}_t) - g_t^2 \left(s_\theta(\boldsymbol{x}, t) + \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)\right)\right] \mathrm{d}t + g_t \, \mathrm{d}\bar{\boldsymbol{w}}_t \quad .$$

Sources: Song et al., 2021b; Chung et al., 2023; Dhariwal and Nichol, 2021

Test-time likelihood: from theory to practical applications

Noise-aware likelihood

$$\nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{y} \mid \boldsymbol{x}_t, t) \quad : \quad \phi^* = \arg\max_\phi \mathbb{E}_{(\boldsymbol{x}_0, y)\sim p_{\text{data}},\, \boldsymbol{x}_t \sim p(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \left[\log p_\phi(\boldsymbol{y} \mid \boldsymbol{x}_t, t)\right] \quad .$$

$$\Rightarrow \quad \begin{aligned} \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t) &\approx \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y} \mid D_\theta(\boldsymbol{x}_t, t) + c), \\ D_\theta(\boldsymbol{x}_t, t) &\approx \mathbb{E}_{p(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)}\left[\boldsymbol{x}_0\right] = \frac{\boldsymbol{x}_t + \sigma_t^2\, s_\theta(\boldsymbol{x}, t)}{\alpha_t} \quad . \end{aligned}$$

Plug-and-play

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t) \quad : \quad \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t) \in \left\{ \log p_\phi(c \mid \cdot),\ \ -\ell(\cdot),\ \ r(\cdot),\ \ -\frac{1}{2\sigma_y^2} \left\| \mathcal{A}(\cdot) - \boldsymbol{y} \right\|_2^2 \right\} \quad .$$

$\Rightarrow$ likelihood model assumptions and setup impact gradient quality.

$\Rightarrow$ requires extensive hyper-parameter searches.

PhD Contributions

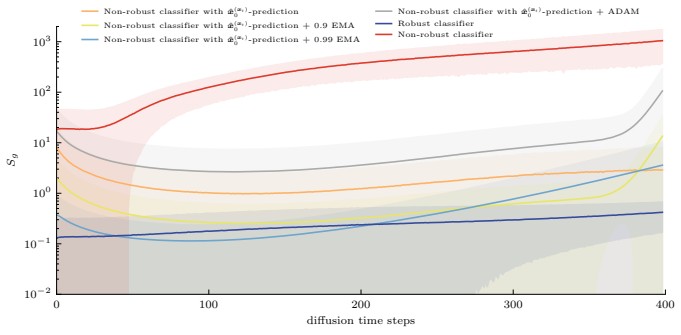
	Paper	Research question	TL;DR
I	DBVCE-eval (ECML Workshop 2023)	How do we evaluate (generated) counterfactual explanations?	Compares evaluation criteria for counterfactual explanations.
II	GradCheck (ArXiv 2024)	How to time dynamics of classifier gradients relate to downstream metrics?	Diagnoses when clean classifiers provide useful or unstable forces.
III	nrCG (ECML 2025)	Can stabilization make non-robust classifiers usable?	Extends classifier guidance with denoising and gradient stabilization.
IV	XAIDIFF (ECML Workshop 2025)	How do counterfactual explanations and generative modeling connect?	Frames explanations probabilistically as guided generative sampling.

Sources: Vaeth et al., 2023, 2024, 2025b,a

Case study II/III: non-robust classifier guidance

Guidance stability metric

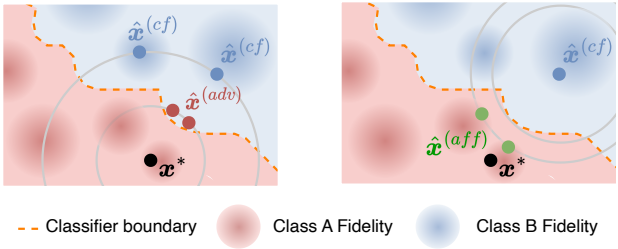
$$S_g\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}\right) = \frac{\left\|\nabla_{\boldsymbol{x}_t} f\left(\boldsymbol{x}_t\right) - \nabla_{\boldsymbol{x}_{t-1}} f\left(\boldsymbol{x}_{t-1}\right)\right\|_2}{\left\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right\|_2}$$



Sources: Vaeth et al., 2024, 2025b

Case study IV: generative explanations by guided sampling

$$\mathrm{d}\boldsymbol{x}_t = \left[f_t(\boldsymbol{x}_t) - g_t^2 \left(\underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)}_{\text{fidelity}} + \gamma_t \underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)}_{\text{validity}} + \beta_t \underbrace{\left\| \hat{\boldsymbol{x}}_0^{(\boldsymbol{x}_t)} - \boldsymbol{x}^* \right\|_1}_{\text{closeness}} \right) \right] \mathrm{d}t + \underbrace{g_t \, \mathrm{d}\bar{\boldsymbol{w}}_t}_{\text{diversity}} .$$



Sources: Vaeth et al., 2025a

Open questions for future research

1. **Approximation gap** between $\nabla_{\boldsymbol{x}_t} \log p(c \mid \boldsymbol{x}_t)$ and $\nabla_{\boldsymbol{x}_t} \log p(c \mid \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t))$
2. **Likelihood model** properties, differentiability and complexity
3. **Dependent conditions** in multi-condition guidance
4. (optional) **Guidance weight/schedule** γ_t
5. (optional) **Posterior calibration** of diversity, likelihood and prior consistency

Open problem: $\hat{x}_0$-prediction / Tweedie Estimation / Mean Guidance

Goal

Approximate the unavailable time-aware likelihood gradient $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)$, interpreting gradient reliability as variance reduction:

$$\mathrm{Var}\left[\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)\right] \;<\; \mathrm{Var}\left[\nabla_{\boldsymbol{x}_t} \log p(c \mid \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t) + \Delta_t)\right] \;<\; \mathrm{Var}\left[\nabla_{\boldsymbol{x}_t} \log p(c \mid \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t))\right] \quad .$$

Fundamental challenge

Off-manifold clean estimates amplify likelihood-gradient noise:

$$p_0\left(\cdot\right) \downarrow \quad \Longrightarrow \quad \mathrm{Var}\left[\nabla \log p(\boldsymbol{y} \mid \cdot)\right] \uparrow \quad .$$

RQ1

Can we close the approximation gap $\nabla_{\boldsymbol{x}_t} \log p(c \mid \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)) \;\approx\; \nabla_{\boldsymbol{x}_t} \log p(c \mid \boldsymbol{x}_t)$?

Open problem: likelihood model

Goal
Semantically meaningful guidance gradients from arbitrary likelihood models $\nabla_{\boldsymbol{x}} \log p_\phi(\boldsymbol{y} \mid \boldsymbol{x})$.

Fundamental challenge
Semantic signal in guidance gradient depends on:

- Complexity of the model
- Training objective
- Inductive biases
- Differentiability, discreteness, Lipschitzness

RQ2
Which likelihood model properties lead to stable, semantic and calibrated gradients?

Sources: Dhariwal and Nichol, 2021; Bansal et al., 2023; Vaeth et al., 2024, 2025b; Chen et al., 2026

Open problem: multiple not conditionally-independent conditions

Goal
Guide by multiple possibly correlated conditions

$$\nabla_{\boldsymbol{x}} \log p\left(\boldsymbol{x} \mid \boldsymbol{y}_1, \boldsymbol{y}_2\right) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \log p\left(\boldsymbol{y}_1 \mid \boldsymbol{x}\right) + \nabla_{\boldsymbol{x}} \log p\left(\boldsymbol{y}_2 \mid \boldsymbol{x}\right) + \underbrace{\nabla_{\boldsymbol{x}} \log \frac{p\left(\boldsymbol{y}_1 \mid \boldsymbol{x}, \boldsymbol{y}_2\right)}{p\left(\boldsymbol{y}_1 \mid \boldsymbol{x}\right)}}_{\text{dependence residual}} .$$

Fundamental challenge

- What are the implications of dropping the dependence residual?
- How could we estimate or learn the dependence residual?

RQ3
How can we guide by multiple correlated conditions?

Sources: Bansal et al., 2023; Vaeth et al., 2025a; Wang et al., 2025; Del Moral, 2004

Open problem: principled guidance weights

Goal

A principled and efficiently optimized guidance weight:

$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{y}) = s_\theta(\boldsymbol{x}, t) + \gamma(t)\, \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t).$

Fundamental challenge

Optimal hyper-parameters require extensive grid/beam searches but settings vary across:

- downstream tasks and data sets
- likelihood model, prior model and solver
- conditioning targets

RQ4

Can $\gamma(t)$ be derived independent of the a) conditioning variable, b) solver, c) prior and likelihood model, d) data set, and e) tasks?

Sources: Ye et al., 2024

Open problem: calibrated posterior diversity

Goal
Prior realism, high likelihood and posterior diversity.

$$\mathrm{d}\boldsymbol{x}_t = \left[f_t(\boldsymbol{x}_t) - g_t^2 \left(\underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)}_{\text{prior realism}} + \gamma_t \underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y} \mid \boldsymbol{x}_t)}_{\text{high likelihood}} \right) \right] \mathrm{d}t + \underbrace{g_t \, \mathrm{d}\bar{\boldsymbol{w}}_t}_{\text{diversity}} .$$

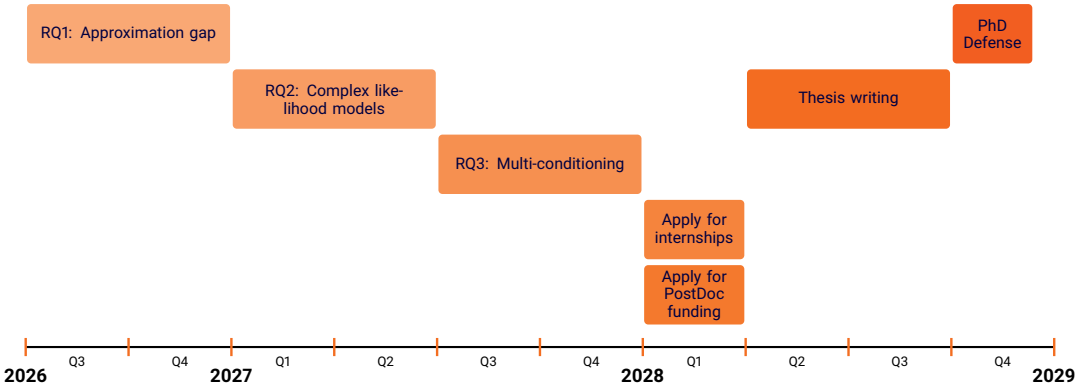
Fundamental challenge

- Prior realism and high likelihood can increase jointly when target labels are well-represented, while often trading-off posterior diversity.
- Guidance has no explicit diversity objective; diversity mainly enters through $g_t \, \mathrm{d}\bar{\boldsymbol{w}}_t$.

RQ5

Can training-free samplers achieve prior realism and high likelihood while preserving posterior diversity?

Future Plan


RQ1: Approximation gap
RQ2: Complex like-lihood models
RQ3: Multi-conditioning
Apply for internships
Apply for PostDoc funding
Thesis writing
PhD Defense
Q3
Q4
Q1
Q2
Q3
Q4
Q1
Q2
Q3
Q4
2026
2027
2028
2029

PhD retrospective: where I stand now

What went well

1. **Personal growth** from zero math background to generative modeling, Diffusion, SDEs, optimization, (differential) geometry, bounds and proofs.
2. **Research seniority** in the research center, reviewing for conferences, (co-)supervising students, being involved in projects, getting to a broader overview of ML, developing better intuitions.

What I want to improve

1. **Research focus** to go deep when things get hard not pivot (cg > cfg > metrics > memorization > privacy/unlearning > quantization).
2. **Faster prototyping** to quickly judge ideas and either move on or go deep (volumes).
3. **Trust in own ideas** instead of reading 20 papers on the topic and dismissing it because "its solved" (cg).
4. **Pragmatism over perfectionism**

cairo.thws
Center for Artificial Intelligence
Technical University of
Applied Sciences
Würzburg-Schweinfurt

UNIVERSITÄT
BIELEFELD

Training-Free Guidance in Diffusion Models: From Classifier Gradients to Future Control Mechanisms

Philipp Vaeth (philipp.vaeth@thws.de)

17-18 June 2026

DMML Workshop Presentation / PhD Exam

References i

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *ICLR*, 2023.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. *CVPR Workshop*, 2023.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. *ICLR*, 2026.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *ICLR*, 2023.

Pierre Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

References ii

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021b.

Philipp Vaeth, Alexander M. Fruehwald, Benjamin Paassen, and Magda Gregorova. Diffusion-based visual counterfactual explanations – towards systematic quantitative evaluation. *ECML PKDD XKDD Workshop*, 2023. URL `https://arxiv.org/abs/2308.06100`.

References iii

Philipp Vaeth, Alexander M. Fruehwald, Benjamin Paassen, and Magda Gregorova. Gradcheck: Analyzing classifier guidance gradients for conditional diffusion sampling. *arXiv:2406.17399*, 2024. URL https://arxiv.org/abs/2406.17399.

Philipp Vaeth, Alexander M. Fruehwald, Benjamin Paassen, and Magda Gregorova. Generative example-based explanations: Bridging the gap between generative modeling and explainability. *ECML PKDD XKDD Workshop*, 2025a. URL https://arxiv.org/abs/2410.20890.

Philipp Vaeth, Dibyanshu Kumar, Benjamin Paassen, and Magda Gregorova. Diffusion classifier guidance for non-robust classifiers. *ECML PKDD*, 2025b. URL https://arxiv.org/abs/2507.00687.

Haoxuan Wang, Jinlong Peng, Qingdong He, Hao Yang, Ying Jin, Jiafu Wu, Xiaobin Hu, Yanjie Pan, Zhenye Gan, Mingmin Chi, Bo Peng, and Yabiao Wang. Unicombine: Unified multi-conditional combination with diffusion transformer. *ICCV*, 2025.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *NeurIPS*, 2024.