MULTI: Disentangling Camera Lens, Sensor, View, and Domain for Novel Image Generation

Sonali Godavarthy[1,3 *], Matthias Neuwirth-Trapp[1,2 *], Tim-Felix Faasch[1], Maarten Bieshaar[1], Michael Moeller[3], Danda Pani Paudel[4]

[1]Bosch Research, [2]ETH Zürich, [3]University of Siegen, [4]INSAIT, Sofia University ‘‘St. Kliment Ohridsk‘‘

* Equal Contribution

Motivation

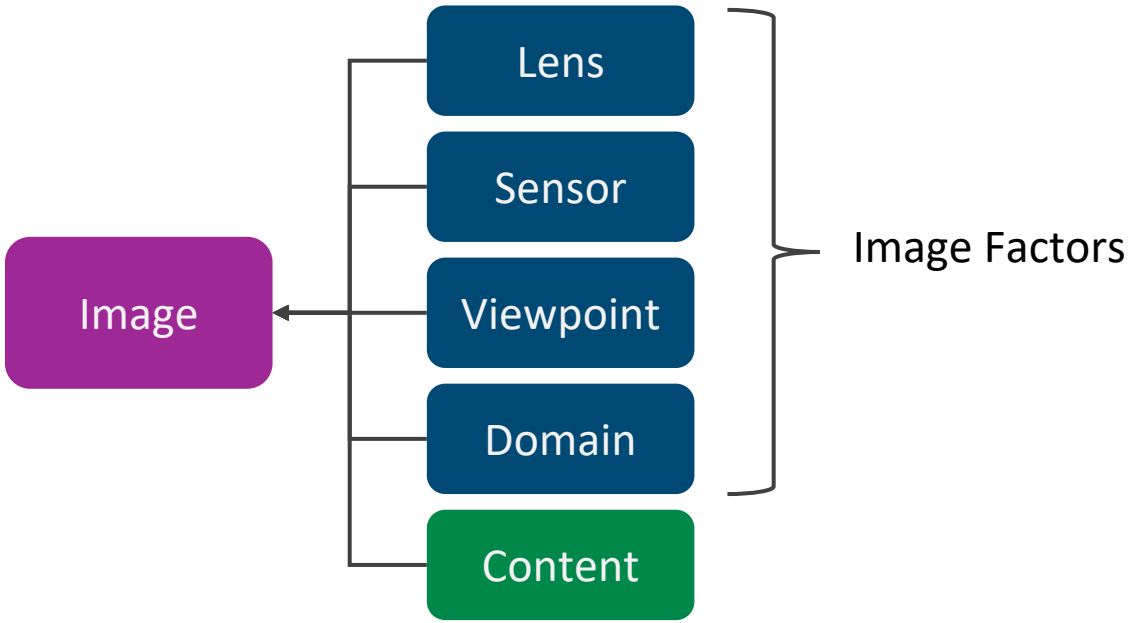



“Draw a thermal image, with fisheye lens from a drone view of a synthetic environment.”

Human artist

GenAI

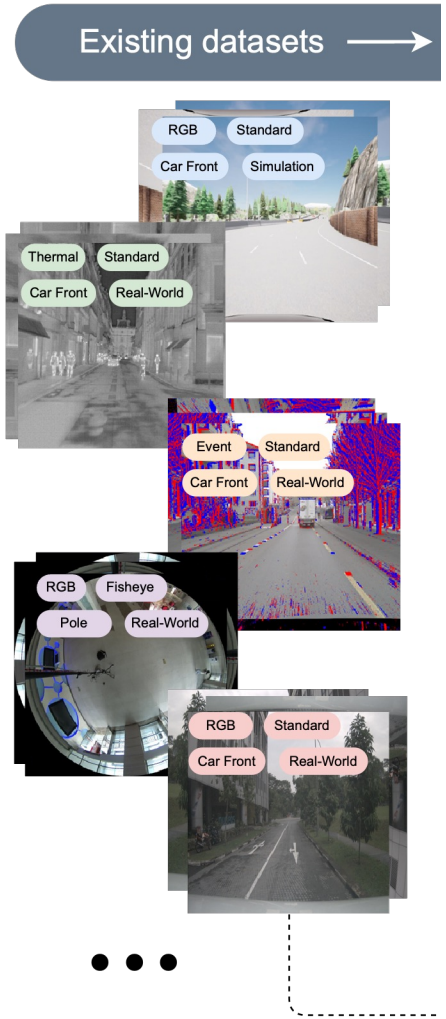
Who will succeed?



Goal: Find a way to extract the image factors from different images to combine them in novel ways.

ETH zürich INSAIT Universität Siegen BOSCH

Goal



Taken different datasets, we **extract the entangled image factors** and enable generation of new images with new combinations.

Further control can be achieved through **conditioning with depth and canny-edge maps**.

Contribution

Our work makes the following new **contributions**

New Problem
Disentangling Image Factors

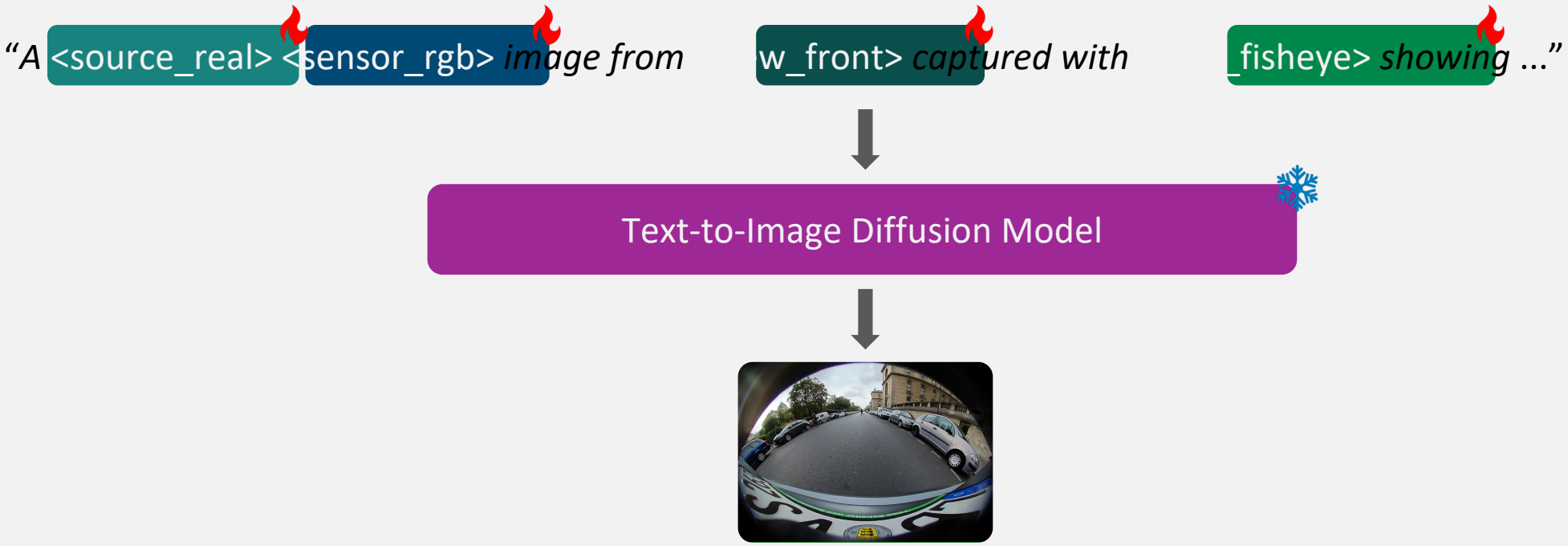
New Method
MULTI (Multi-factor disentanglement through Textual Inversion)

New Evaluation
Benchmark + FAA (Factor Alignment Accuracy)

Related Works: Existing methods (e.g. DreamBoth, InspirationTree) focus on disentangling only on the image content.

Method

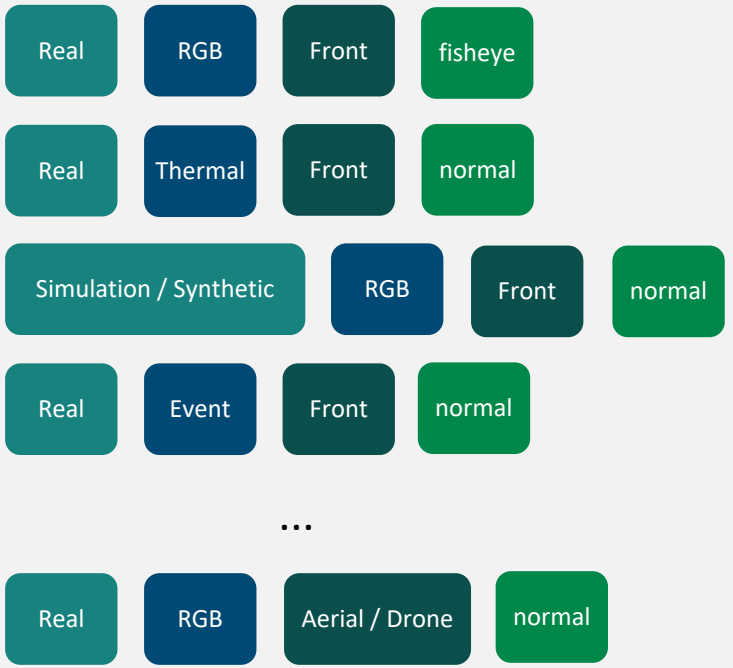
Basis: **Textual Inversion** with frozen Diffusion Model



However: This does not disentangle the factors.

Method

Combine **overlapping and non-overlapping factors** into single **batch**.



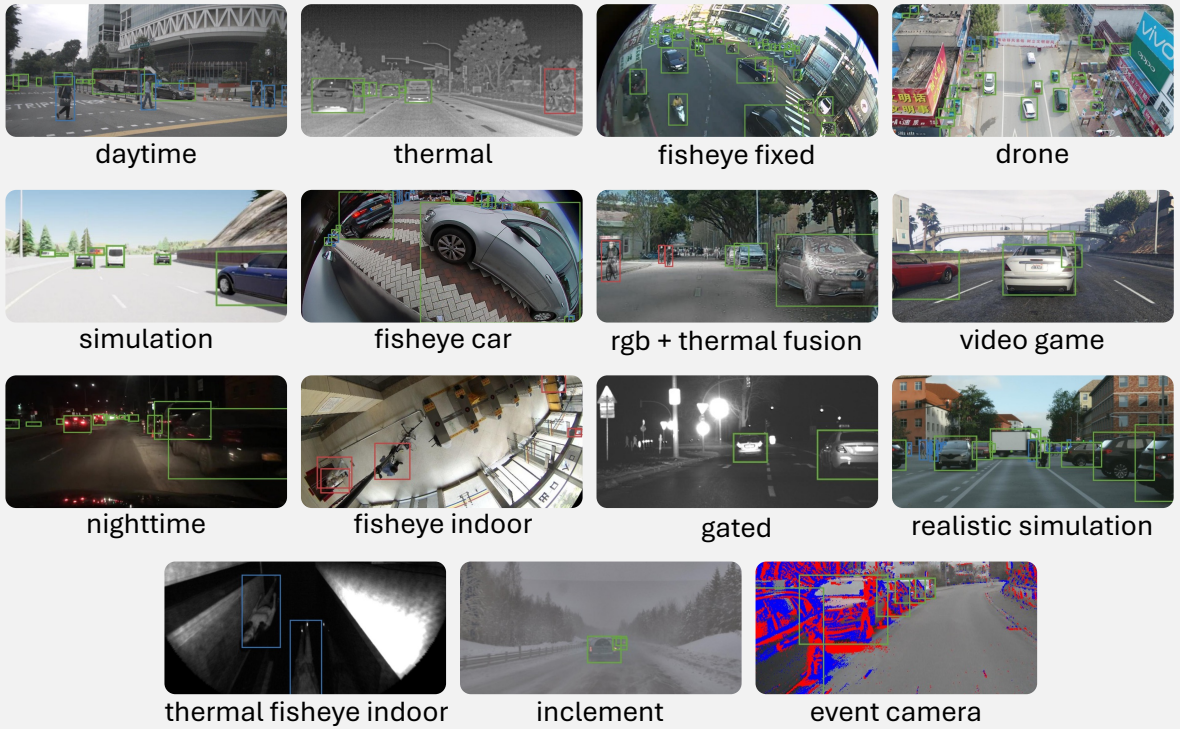
Only factor embeddings are trained

Details on the **individualization factors to specific datasets** can be found in the paper

- Same **factors are shared across datasets** and images
- Batching allows the training to **identify the similarities**
- Initiation of factor embedding from text

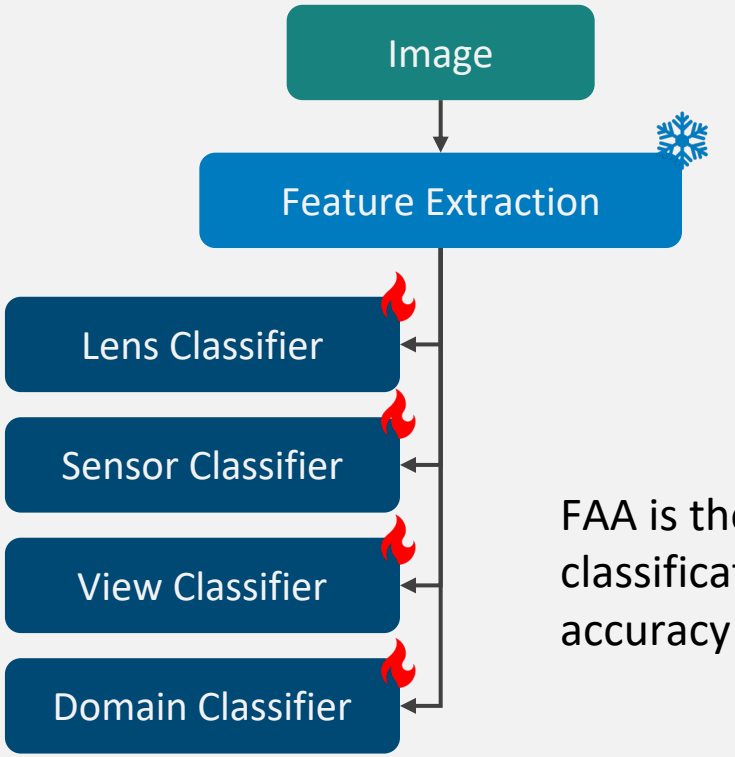
Benchmark and Measuring Accuracy

RICO benchmark from Incremental Object Detection adapt to make **DF-RICO**



Neuwirth-Trapp, Matthias, et al. "*RICO: Two Realistic Benchmarks and an In-Depth Analysis for Incremental Learning in Object Detection*." *ICCV Workshops*. 2025.

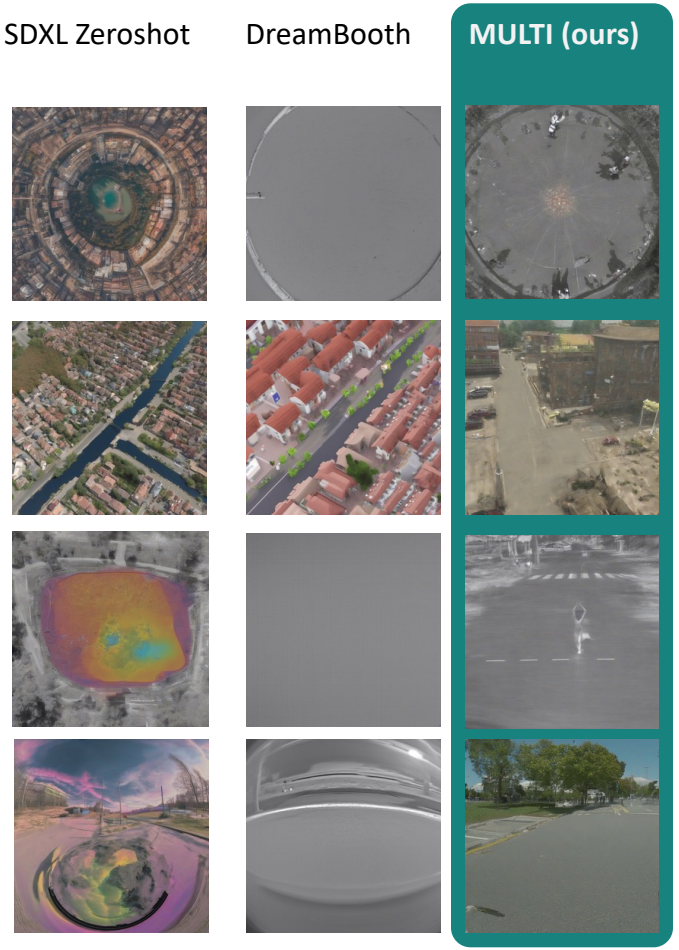
Factor Alignment Accuracy



FAA is the mean classification accuracy

Results: Qualitative

Domain	Sensor	Viewpoint	Lens
Real	Thermal (flir)	Drone	Fisheye
Simulation	RGB	Drone (loaf)	Normal
Real	Thermal	Pole	Normal
Real	RGB-Thermal	Front	Fisheye



ControlNets can be used for **Image-to-Image generation**

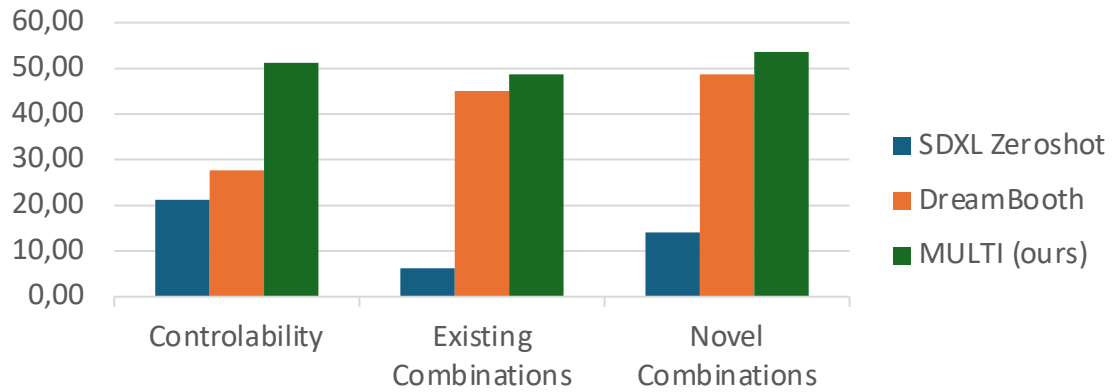
Results: Quantitative and User Study

Quantitative Results

Method	CLIP ↑	IS ↑	DS ↑	FAA ↑ (ours)
SDXL Zeroshot	22.01	1.23	0.53	40
DreamBooth	27.13	**1.86**	0.53	42
MULTI (Ours)	**27.32**	1.79	**0.57**	**44**

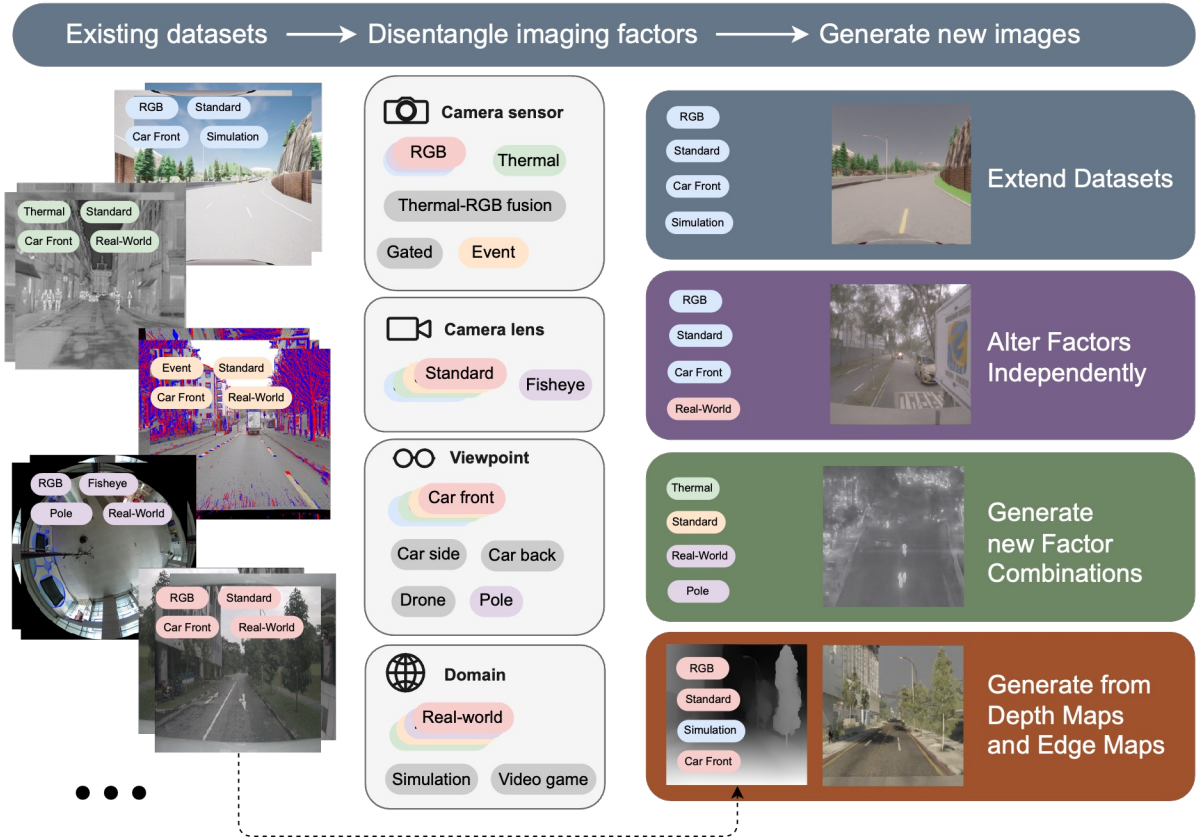
FID <u>cannot</u> be applied for novel combinations, as is there is no distribution to compare to.

MULTI outperforms baselines also in a **user study**



MULTI: Disentangling Camera Lens, Sensor, View, and Domain for Novel Image Generation

Contact
M. Neuwirth-Trapp
mneuwirth@ethz.ch



Scan for the paper
arXiv:2605.12134

Follow-up: *X-MULTI* (accepted to ECCV-W, coming to arXiv soon)

ETH zürich INSAIT Universität Siegen BOSCH

Appendix

Motivation

Scene Content

Text-to-image models already give strong, precise control over what appears in a scene like objects, style, composition.

Capture Conditions

Almost no control exists over how a scene is physically captured like its camera lens, sensor type, and viewpoint.

That gap matters for safety-critical AI: for autonomous driving, the exact lens, sensor, and viewpoint used to capture an image determine whether a vision model trained on it will transfer to the real world.

Problem

Individual datasets tend to fix several factors together, so models never see them varied independently:

Simulation & video-game datasets

→ almost always front-facing view, and with standard lens

Gated & event-camera datasets

→ almost always captured with a standard lens

As a result, generative models overfit to these shortcuts instead of learning each factor independently.

Related Work

Existing Work

- Stable Diffusion XL [1]: Frozen text-to-image backbone
- Textual Inversion (TI) [3]: learns a new token for a concept keeping model frozen
- DreamBooth [2]: learns new concept by training small subset of parameters
- InspirationTree [4]: disentangles concepts using hierarchical tree structure

Focuses on controlling artistic style or image content.

Contributions

- Introducing imaging factor disentanglement as controllable imaging factors
- **MULTI**: two stage learning: general factor → dataset-specific refinement
- **DF-RICO**: first benchmark for imaging-factor disentanglement
- **Factor Alignment Accuracy (FAA):** a new metric to evaluate factor correctness

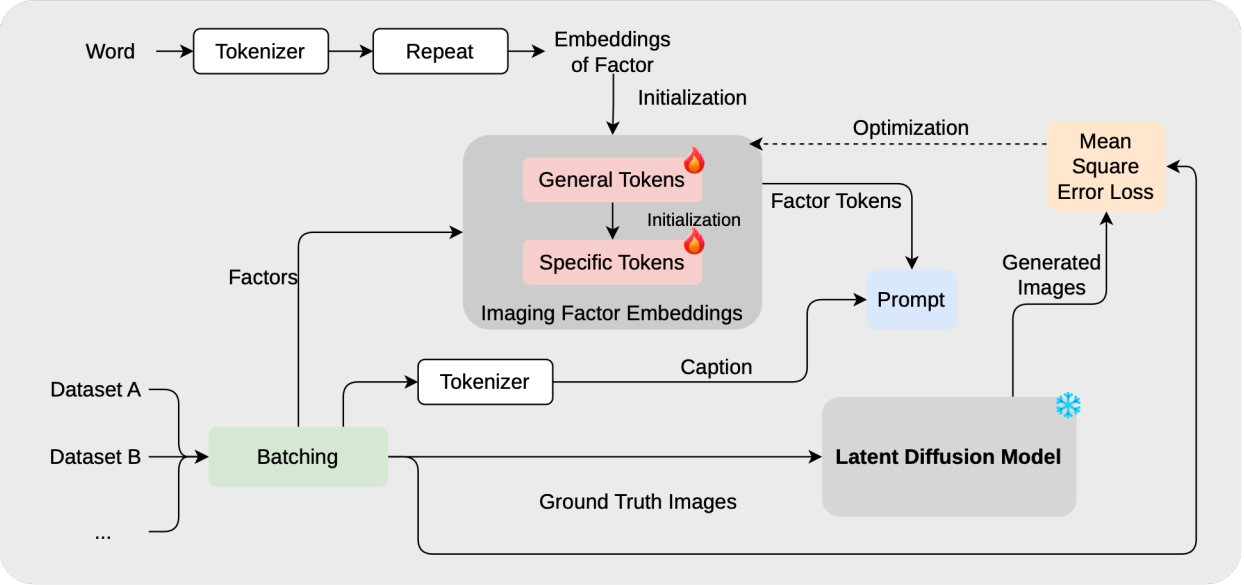
Controls imaging factors like lens type, the sensor with its associated color profile, the viewpoint, and the domain.

The Task
Imaging Factor Disentanglement

Camera Lens	Sensor Type	Viewpoint	Domain Source
Standard · Fisheye	RGB · Thermal · RGB Thermal · Gated · Event	Front · Back · Side · Pole · Drone	Real · Simulation · Video game

***Goal**: control each factor independently to generate novel, out-of-distribution factor combinations.*

Method
MULTI



1 General factor learning

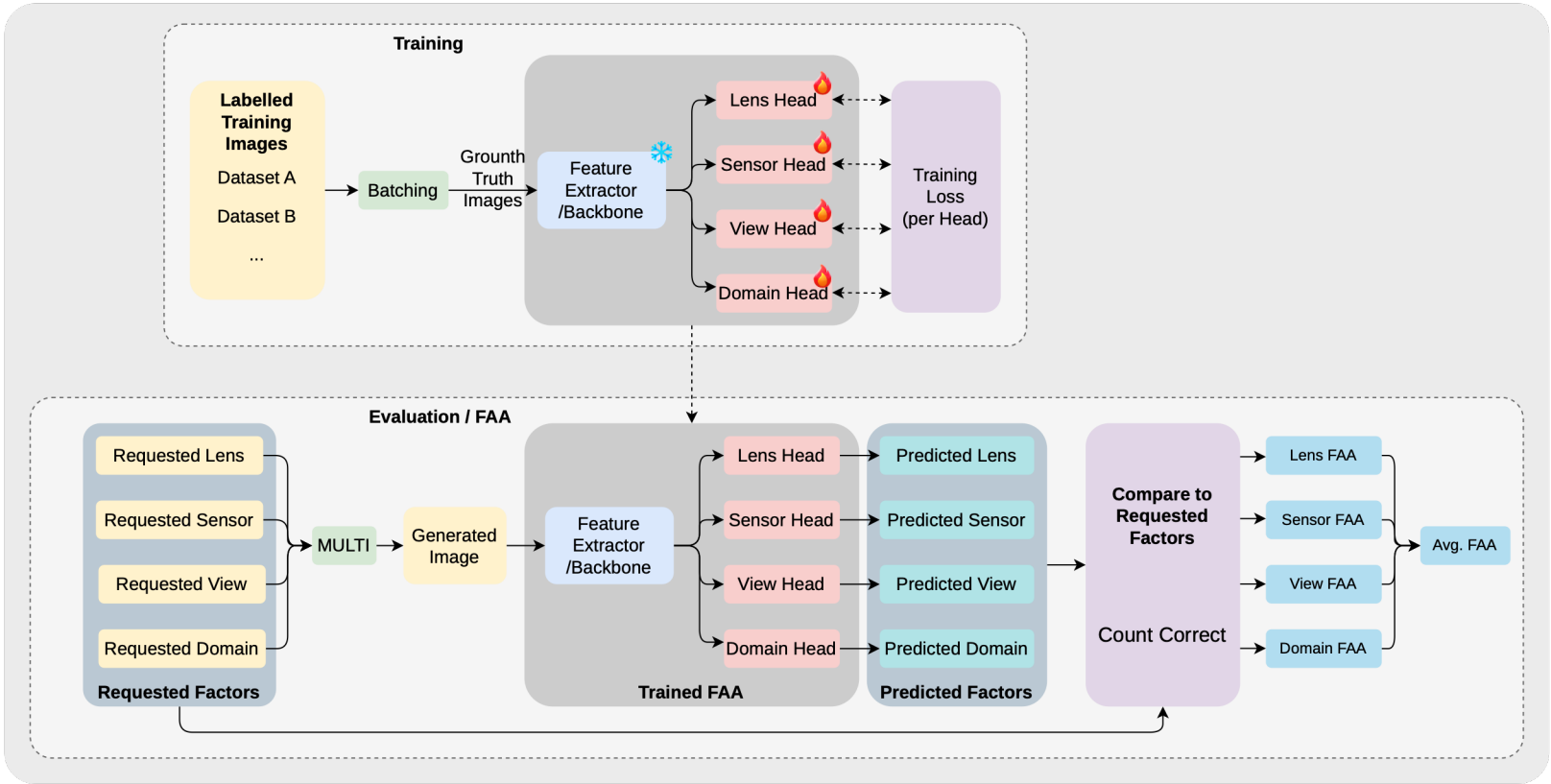
One shared token per factor, trained jointly across all datasets.

2 Dataset-specific refinement

A copy of each token is refined to capture per-dataset nuance.

Each token consists of 'n' learnable vectors

Metric
Factor Alignment Accuracy (FAA)



Standard image-quality metrics (FID, CLIP Score, IS, DS) do not check whether a generated image actually carries the requested imaging factor. FAA does.

FAA reports whether a generated image actually contains the requested factor

Benchmark

DF-RICO: the first benchmark for factor disentanglement

15

datasets combined
Collected from RICO [5] benchmark

4

imaging-factor types

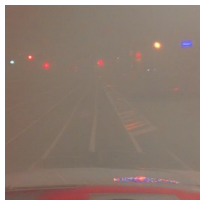
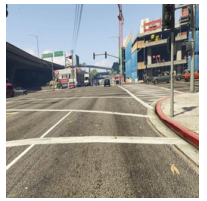
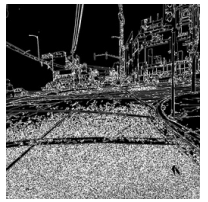
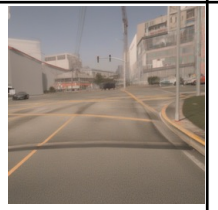
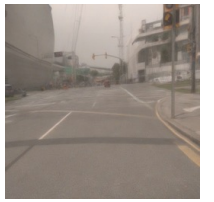
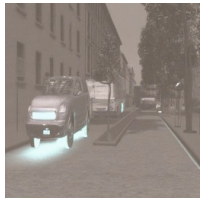
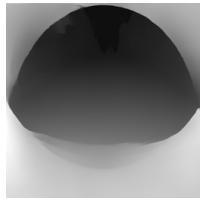
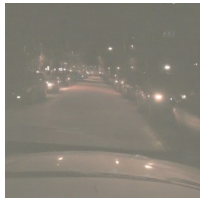
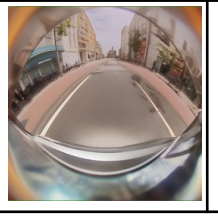
5

evaluation metrics

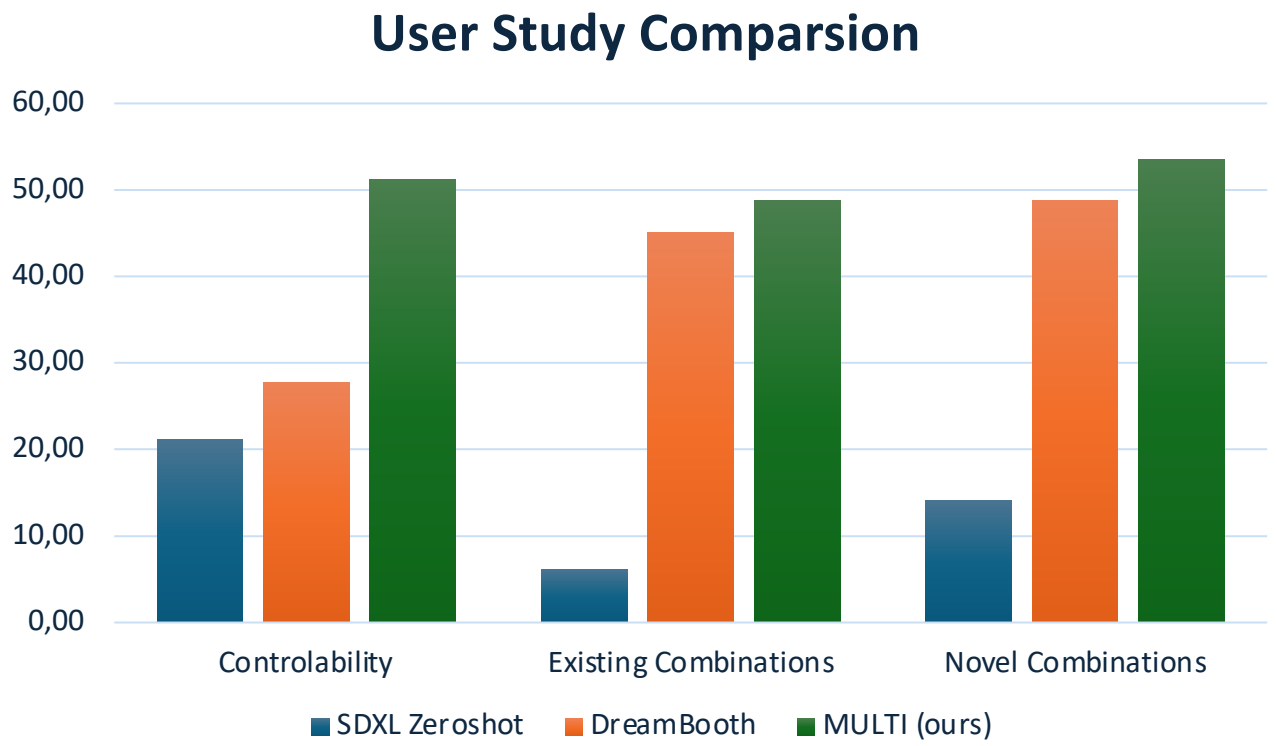
Evaluation metrics: FID, CLIP Score, IS, DS, and Factor Alignment Accuracy (FAA).

Results - Qualitative

Novel Combination (Using Control Maps)

Domain	Sensor	Viewpoint	Lens	Original	Altered Factor	ControlNet Maps	SDXL Zeroshot	DreamBooth	MULTI (ours)
Real	RGB	Front	Normal		RGB -> Event				
Video game	RGB	Front	Normal		Videogame -> Real (nuimages)				
Simulation	RGB	Front	Normal		RGB -> Thermal (flir)				
Real	RGB	Back	Fisheye		RGB -> RGB-Thermal				

Results – User Study



Conclusion

- MULTI learns one text token per imaging factor via two-stage textual inversion, keeping the backbone frozen.
- DF-RICO is the first benchmark for imaging-factor disentanglement, spanning 15 datasets.
- Factor Alignment Accuracy (FAA), our new metric, shows MULTI reaches 44% factor control.
- ControlNet integration extends MULTI to image-to-image factor edits.
- **MULTI is a first step toward fully controllable image generation with lens, sensor, view and domain being freely combined.**

References

[1] Podell, Dustin, et al. "Sdxl: Improving latent diffusion models for high-resolution image synthesis." International Conference on Learning Representations. Vol. 2024. 2024

[2] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[3] Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

[4] Vinker, Yael, et al. "Concept decomposition for visual exploration and inspiration." ACM Transactions on Graphics (TOG) 42.6 (2023): 1-13.

[5] Neuwirth-Trapp, Matthias, et al. "RICO: Two Realistic Benchmarks and an In-Depth Analysis for Incremental Learning in Object Detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.

Appendix

DF-RICO: full dataset composition

Dataset	Sensor	Viewpoint	Lens	Domain
nuImages & WoodScape	RGB	Back / Front / Side	Normal	Real
Teledyne FLIR	Thermal	Front	Normal	Real
Fisheye8K	RGB	Pole	Fisheye	Real
VisDrone	RGB	Drone	Normal	Real
SHIFT & Synscapes	RGB	Front	Normal	Simulation
SMOD	RGB-Thermal	Front	Normal	Real
Sim10k	RGB	Front	Normal	Video game
BDD100K & DENSE	RGB	Front	Normal	Real
LOAF	RGB	Drone	Fisheye	Real
DENSE	Gated	Front	Normal	Real
TIMo	Thermal	Pole	Fisheye	Real
DSEC	Event	Front	Normal	Real

Appendix
Training objective: details

- Backbone: a frozen, pretrained text-to-image diffusion model (Stable Diffusion XL).
- Learnable parameters: only the embedding vectors of the factor tokens each associated with n learnable vectors, following the Textual Inversion.
- Stage 1 - general factor learning: images from all datasets are shown together and each factor token is updated via the standard diffusion reconstruction loss.
- Stage 2 - dataset-specific refinement: a copy of each Stage-1 token is initialized and further optimized on its own dataset, to absorb dataset-specific visual nuance.
- Inference: learned tokens are inserted into a single prompt, ControlNet can be used to add spatial conditioning for image-to-image edits.