Yandex Research

From Noise to Narrative: The Evolution of Visual Generative Models

Sergey Kastryulin

Research Scientist

Agenda

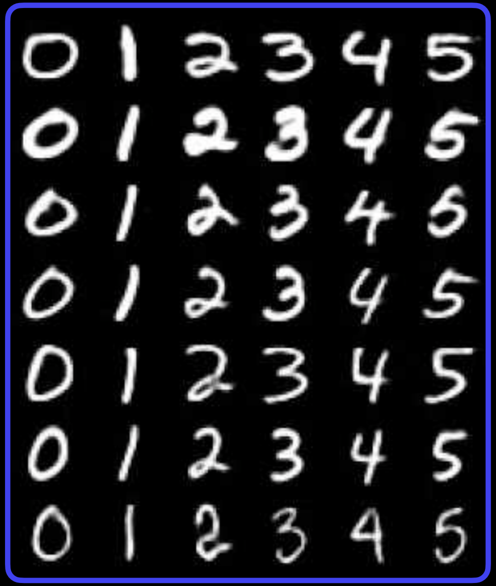
01 An overview of conditioning

02 Multi-modal generative models

03 (Open) questions

04 Inference-time compute scaling

Visual Generative Modeling



Unconditional

Visual Generative Modeling

Unconditional

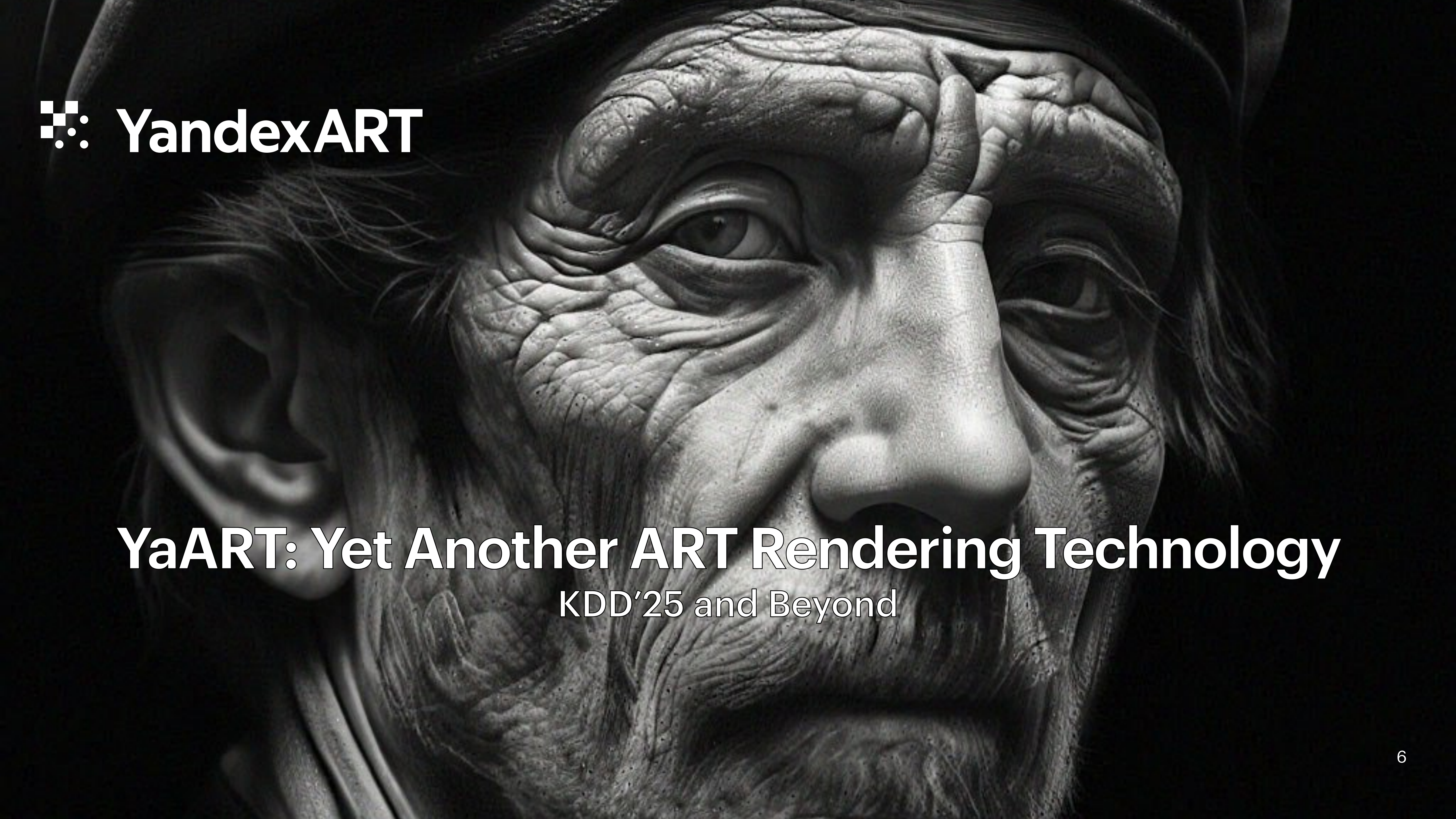
Class-conditional

Visual Generative Modeling

Unconditional

Class-conditional

Text-conditional



YandexART

YaART: Yet Another ART Rendering Technology

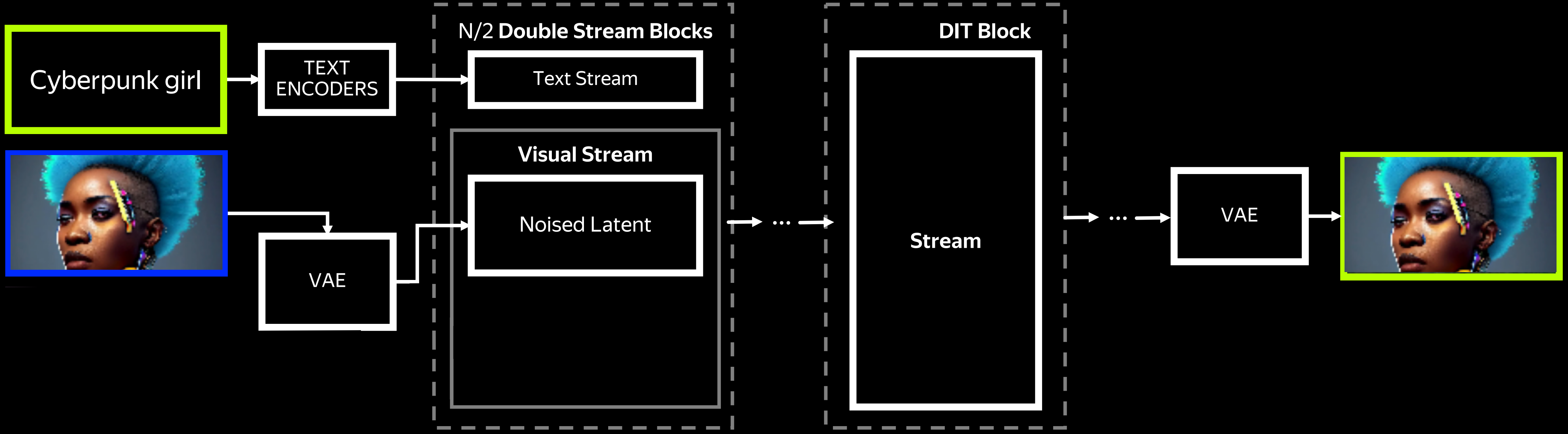
KDD’25 and Beyond

YandexART

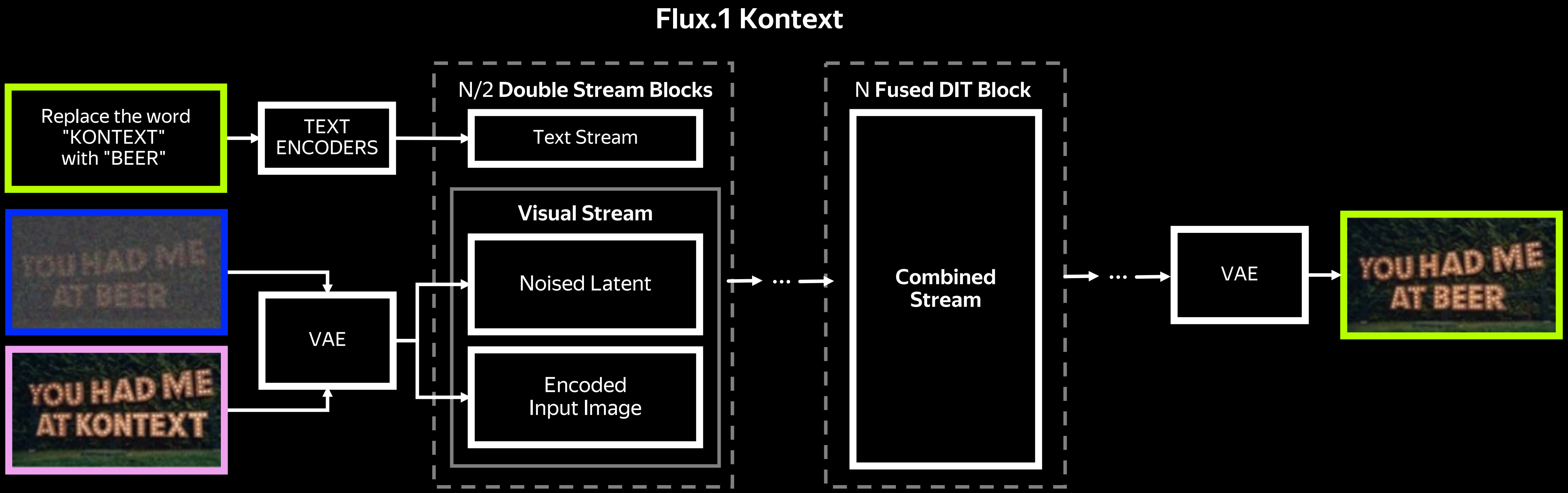
NeurIPS’25
Fall into ML — 16:00, poster #39)

Conditioning on text

In general



Conditioning on text and image



S. Batifol et. al., "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space", 2025

Conditioning on text and image

Restore photo

Add sports costume

Yandex Research

In reality, we want a dialog

In reality we want a dialog

Why?

01 Generate new content

02 Edit existing content

03 Ask questions

04 Work with several images

05 Support context

06 Bonuses from different modalities*

* J. Zhang et. al., “Are Unified Vision-Language Models Necessary: Generalization Across Understanding and Generation”, 2025

Dialog means unification

Unify two main generative worlds

- Inherently continuous (visual): images, video, 3D
- Inherently discrete (textual): text, code, math

Dialog means unification

Unify two main generative worlds

- Inherently continuous (visual): images, video, 3D
- Inherently discrete (textual): text, code, math

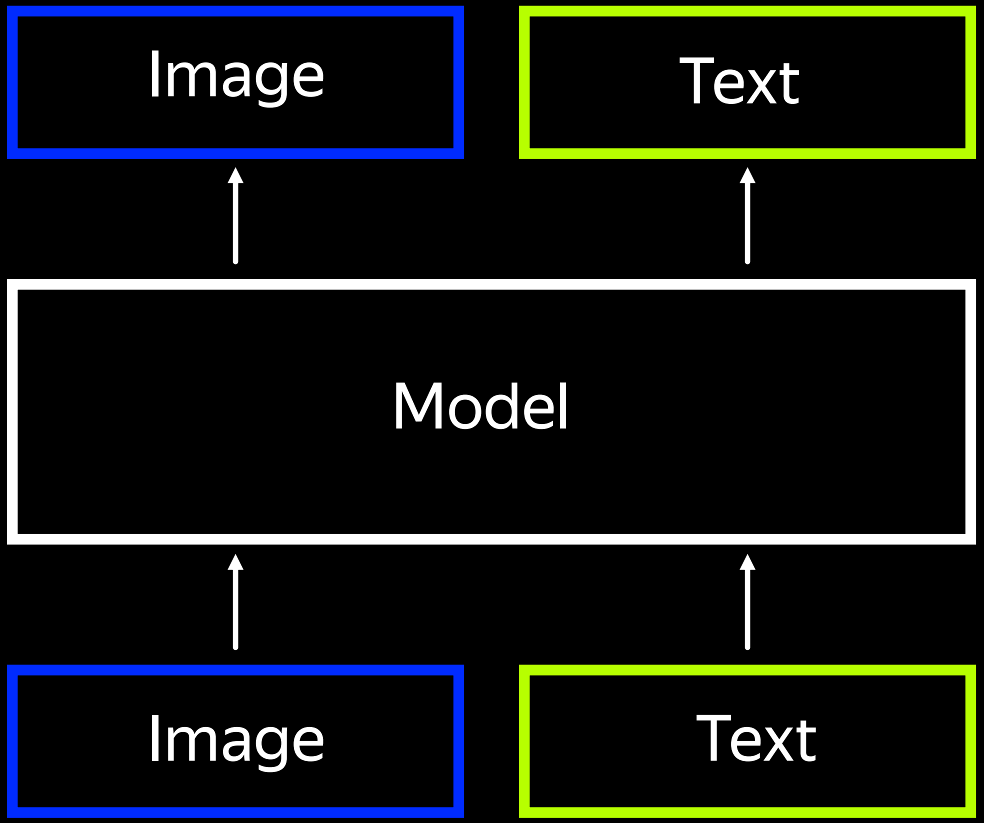
Currently we can model

- text-to-text — LLM
- text-to-image — all we discussed above
- image-to-text — VLM
- text+image-to-image — Editing

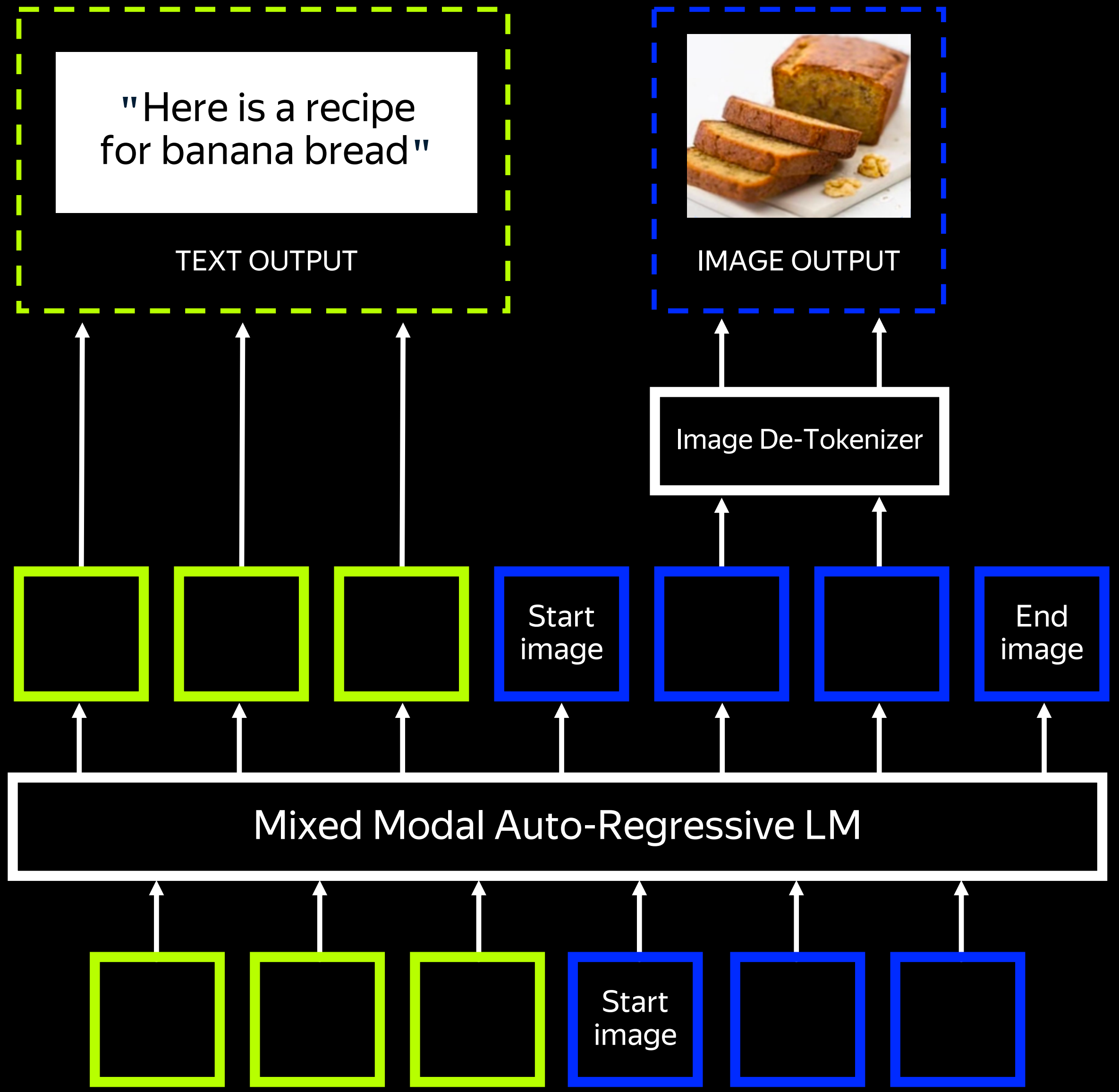
Yandex Research

Towards Unified Generative Models

What kind of model would help us?

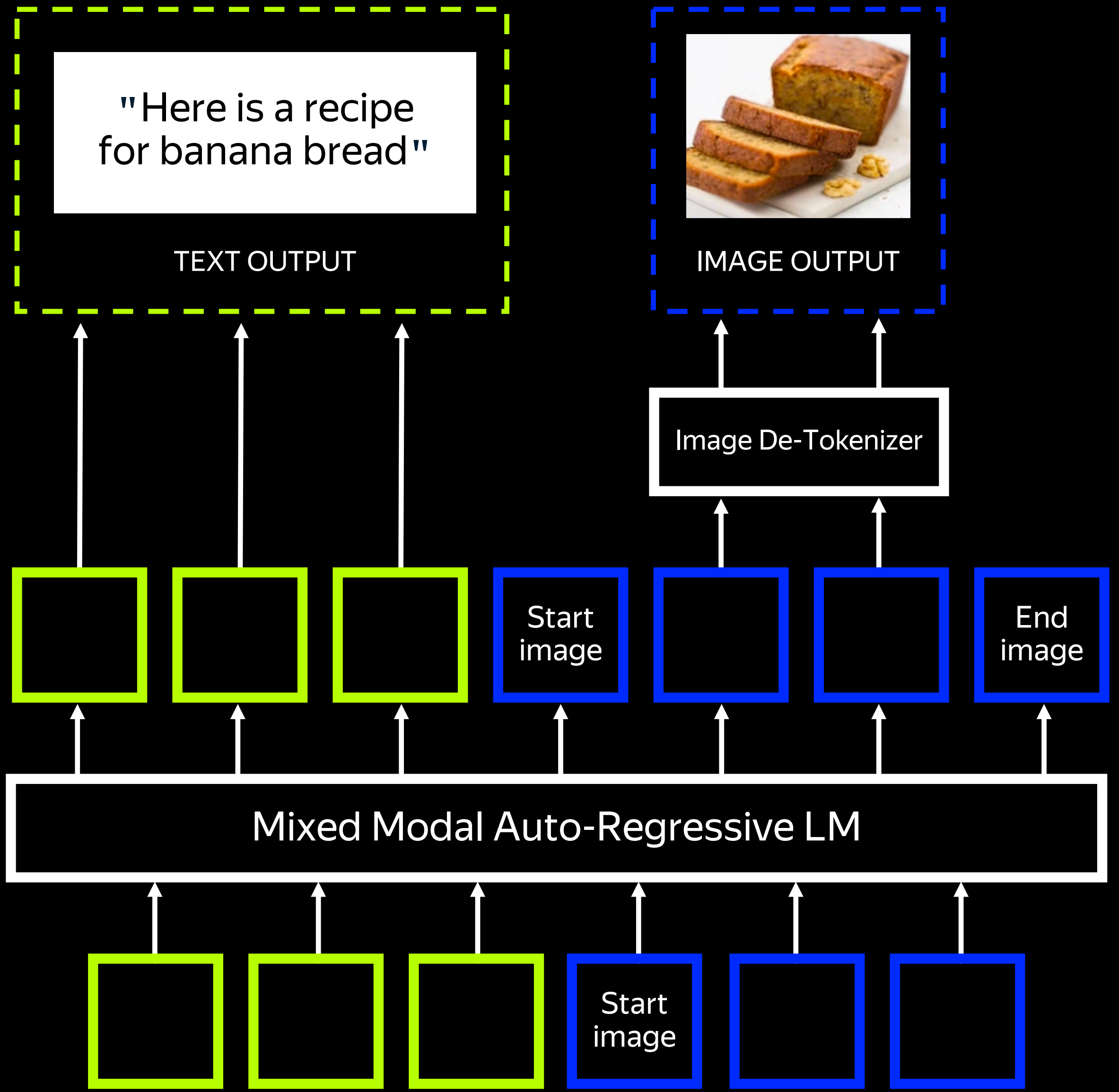


Naive approach



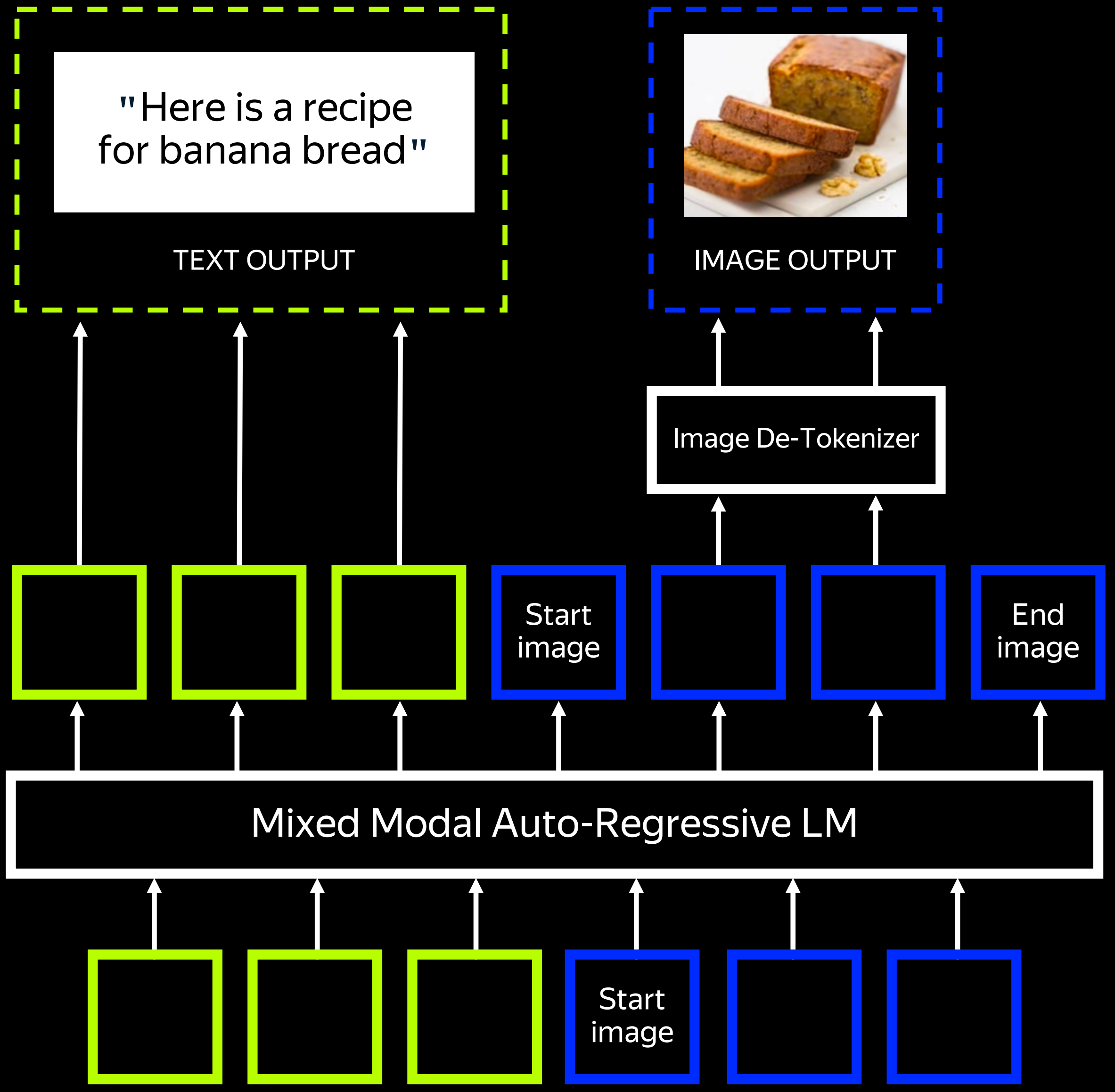
Naive approach

- ⊕ Uniform implementation
- ⊖ Raster order is not native to images
- ⊖ Discretisation ruins image quality



Naive approach

- Uniform implementation
- Next-token prediction is not native to images
- Discretisation ruins image quality

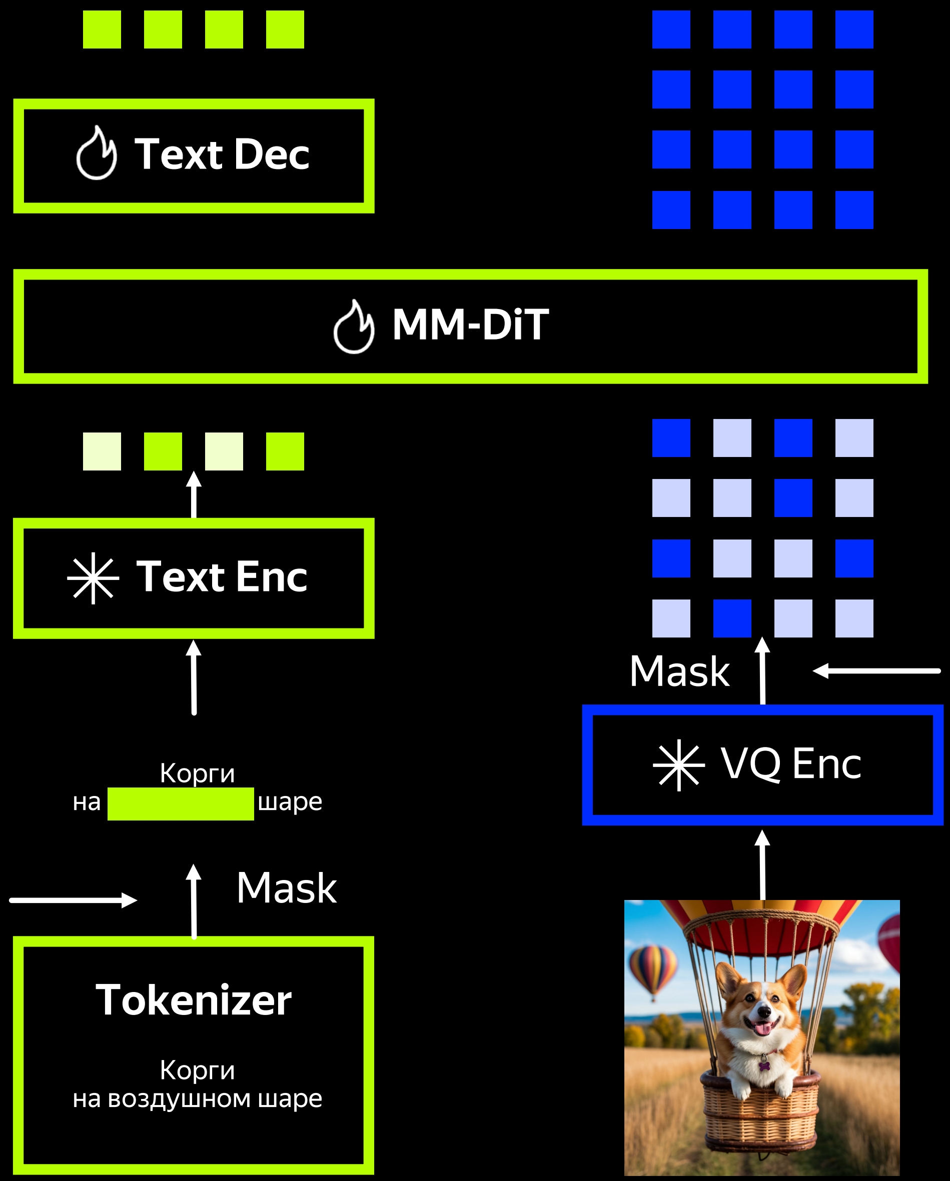


sic!

A. Ramesh et. al., “Zero-Shot Text-to-Image Generation” (aka DALL-E 1), 2021

Muddit

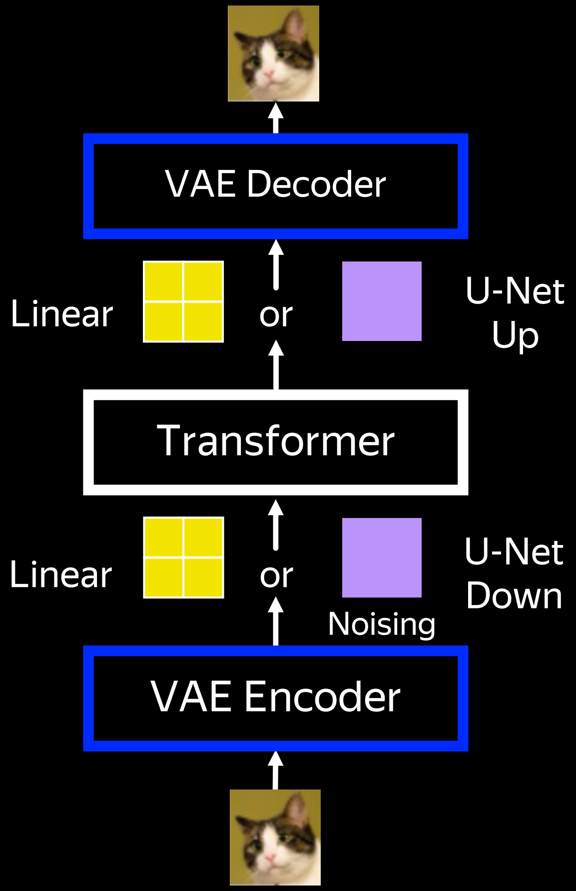
- Uniform implementation (+)
- Discrete diffusion for text data* (−)
- Discrete tokens for image data (−)

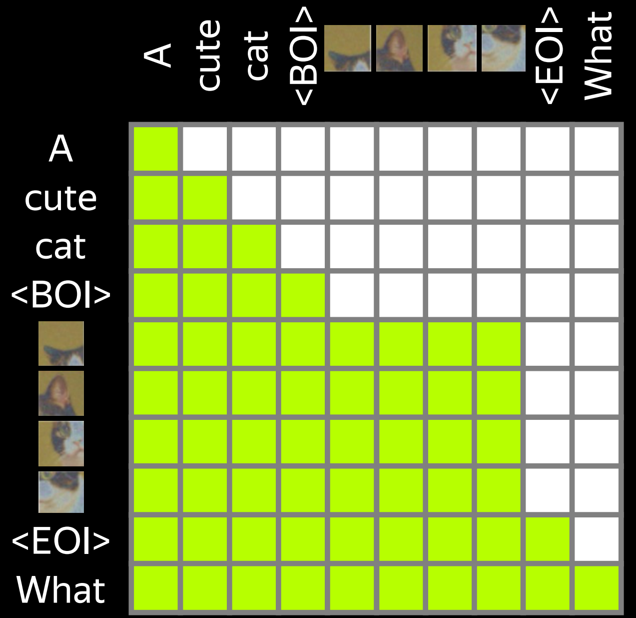


Q. Shi et. al., “Muddit: Liberating Generation Beyond Text-to-Image with a Unified Discrete Diffusion Model”, 2025

Transfusion

- Unified representation in a single transformer
- Causal attn for text, bidirectional for images
- Continuous encoding of images



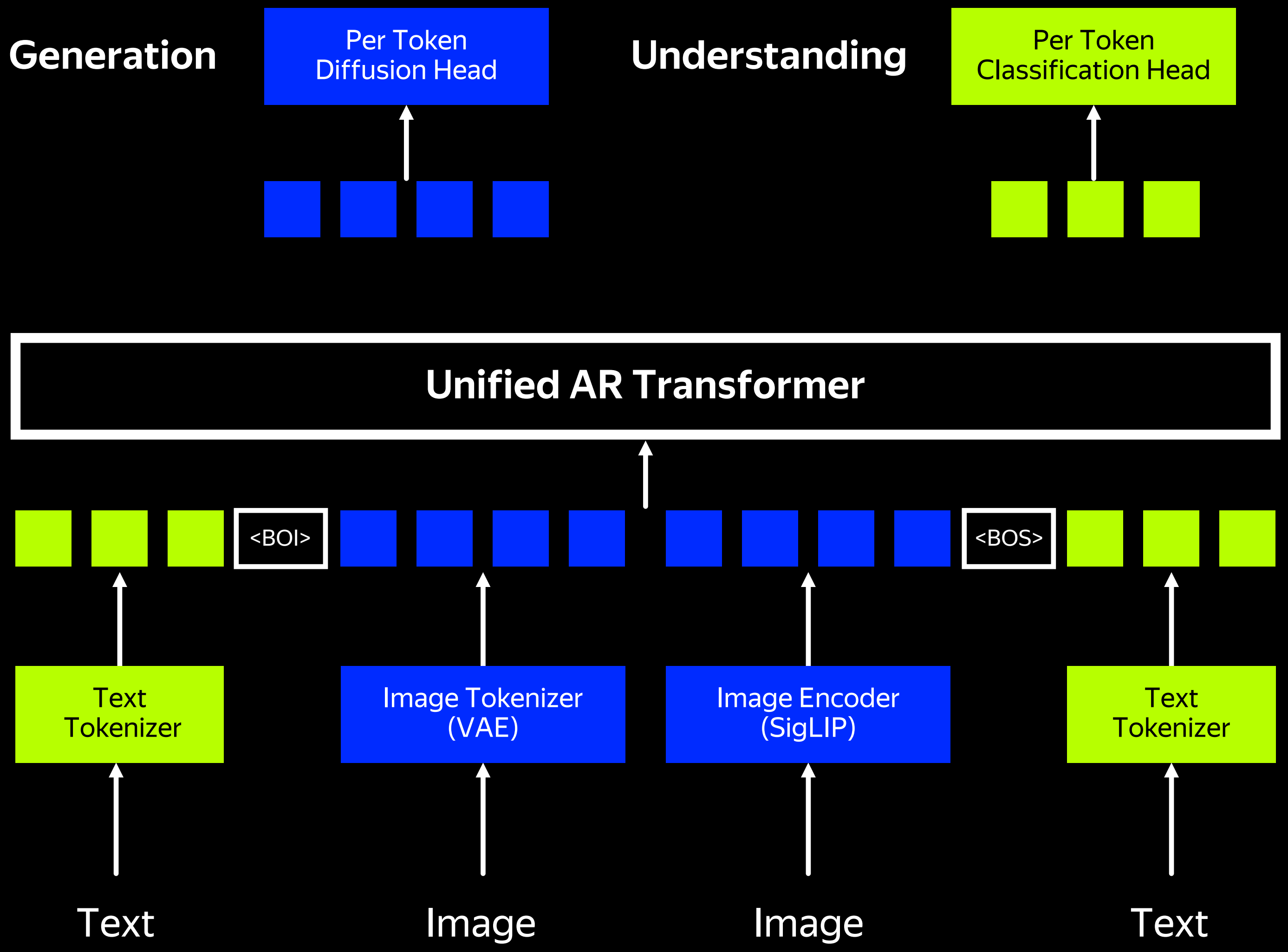


$$\mathcal{L}_{Transfusion} = \mathcal{L}_{LM} + \lambda \times \mathcal{L}_{DDPM}$$

C. Zhou et. al., “Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model”, 2024

MAR, (Uni)Fluid

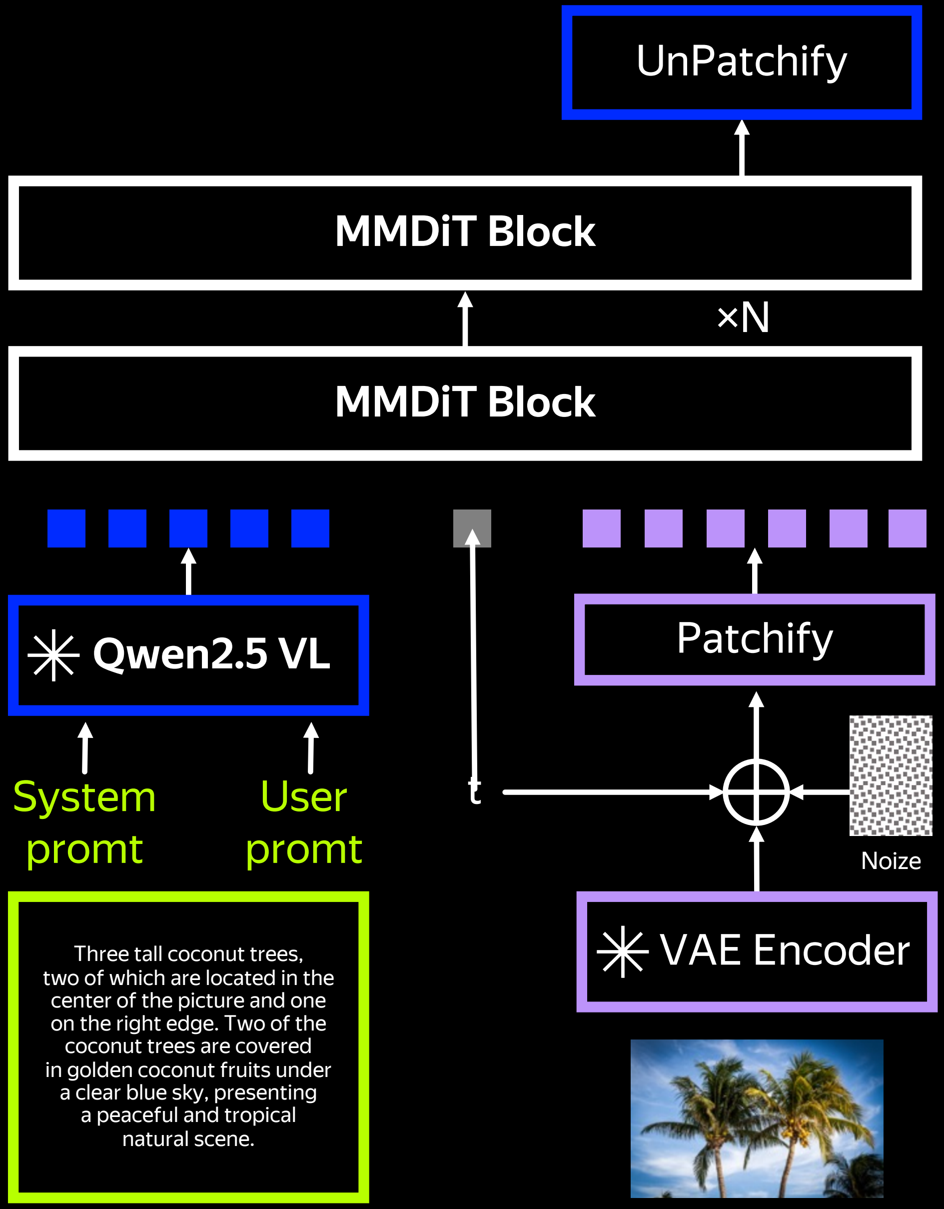
- (+) Use pre-trained image encoders (VAEs)
- (+) Continuous representation of images
- (−) Long sampling



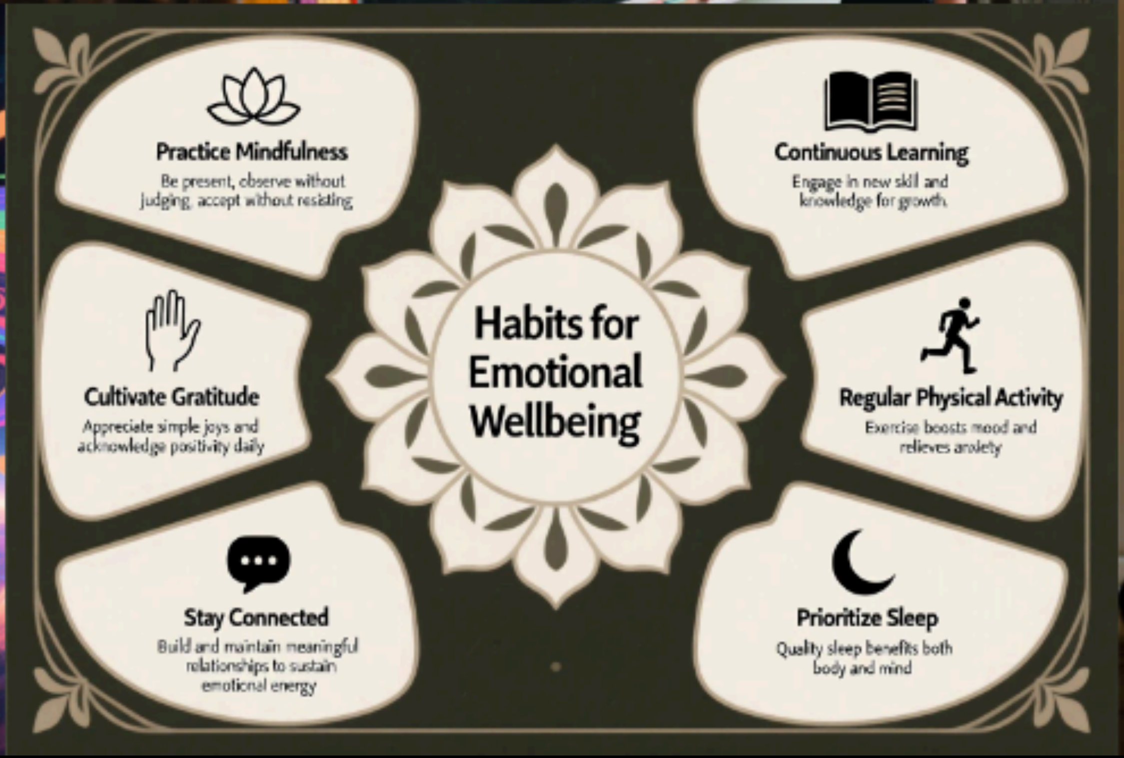
T. Li et. al., “Autoregressive Image Generation without Vector Quantization”, 2024

NexusGen | Qwen Image

- ⊕ Re-use pre-trained diffusion denoiser
- ⊕ Should be cheaper to train, right?*
- ⊖ Hard to align with good quality



C. Wu et. al., “Qwen-Image Technical Report”, 2025



C. Wu et. al., “Qwen-Image Technical Report”, 2025

How hard is it to train the connector?

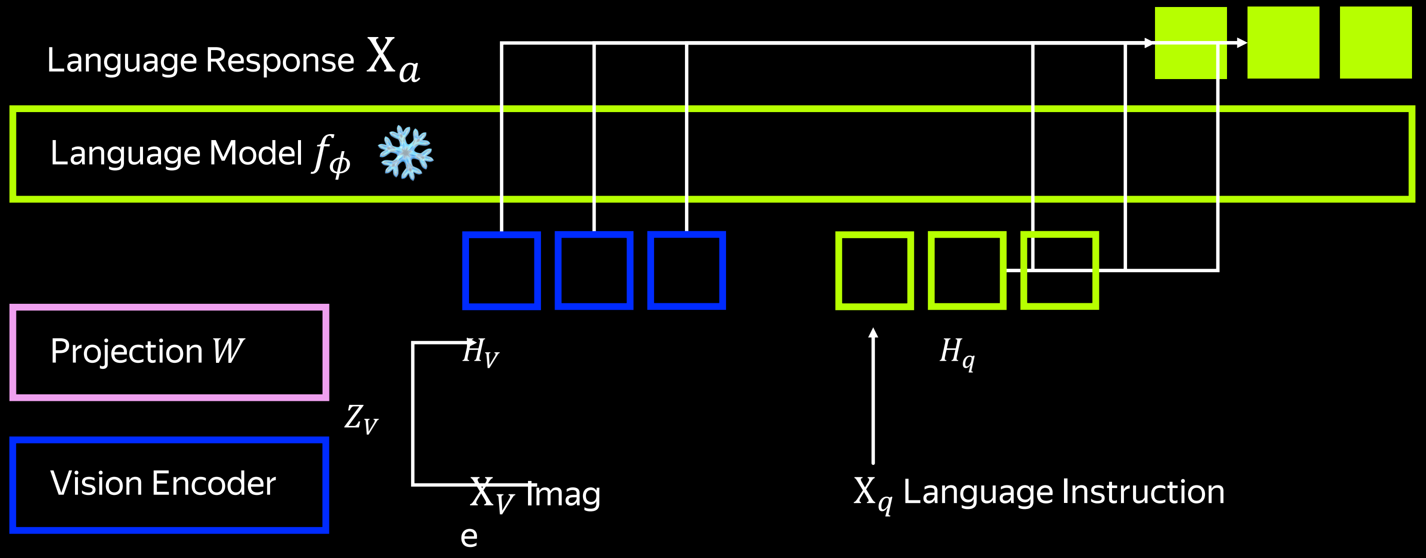
"a pair of cherries, dressed in a delicate ballerina outfit"

- Images generated with **NexusGen** have low quality
- Did not train from scratch

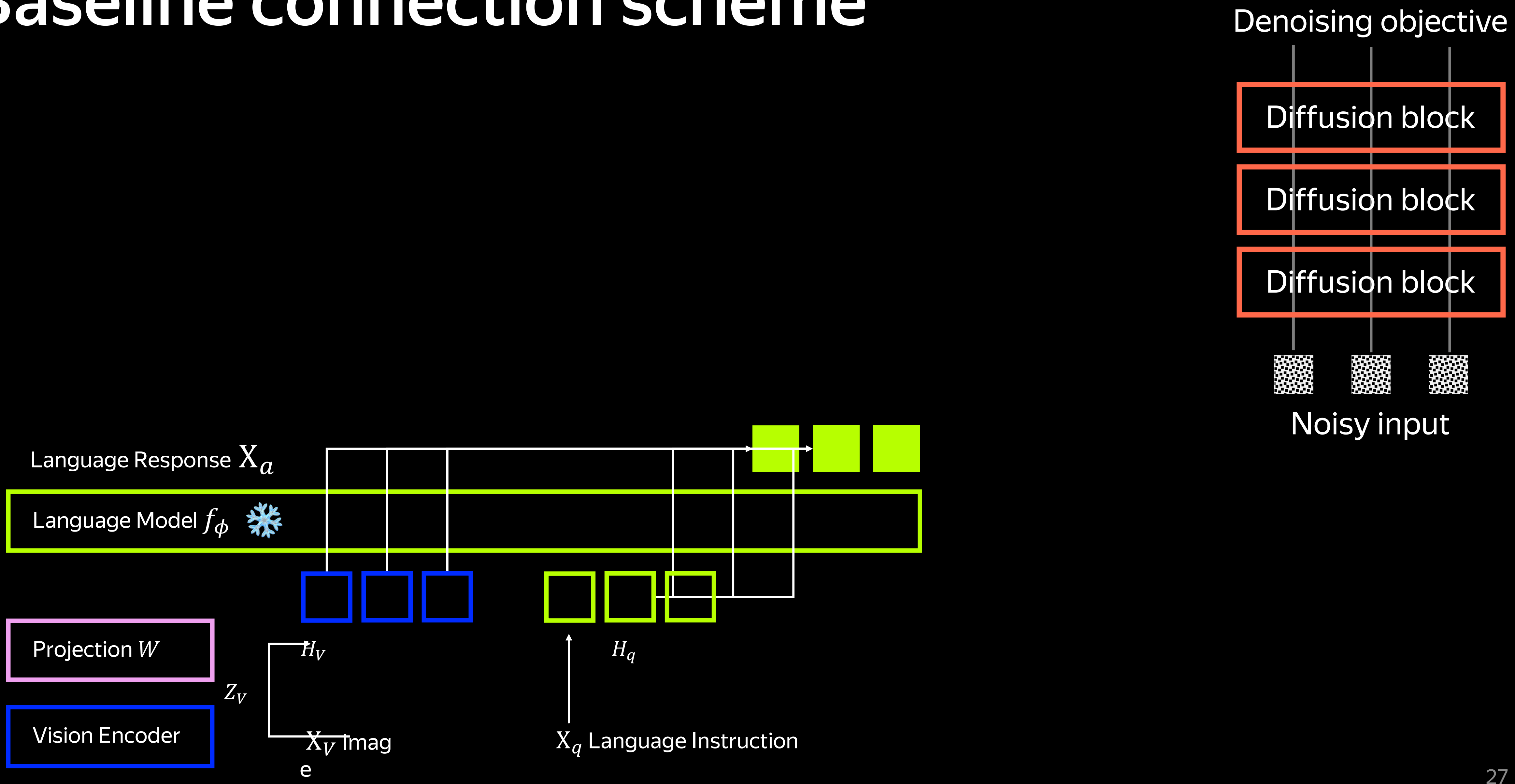
- Images generated with **Qwen Image** are much better
- Trained from scratch

* Very much a hand-waving argumentation, lots of other variables are not aligned between setups

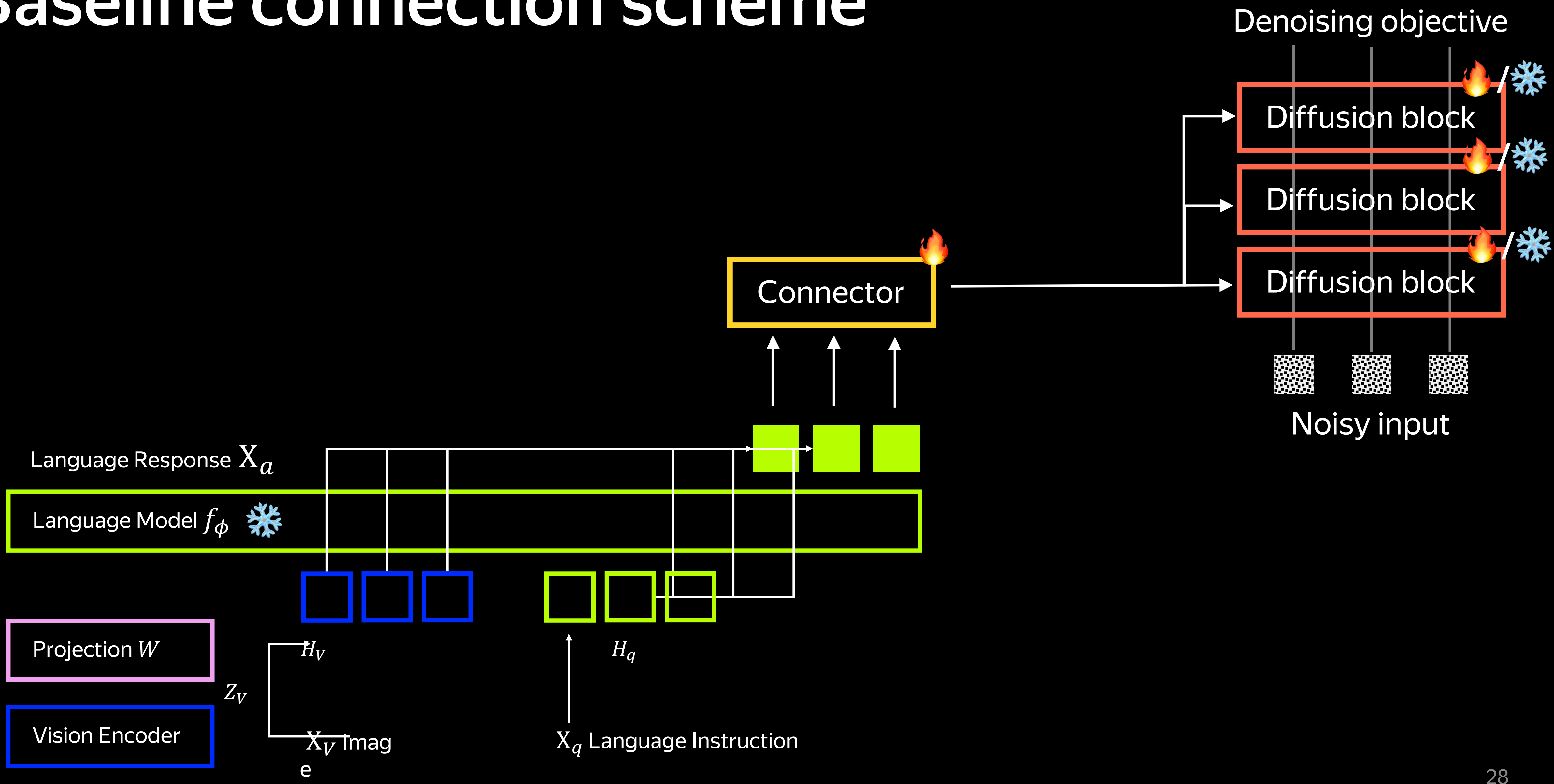
Baseline connection scheme



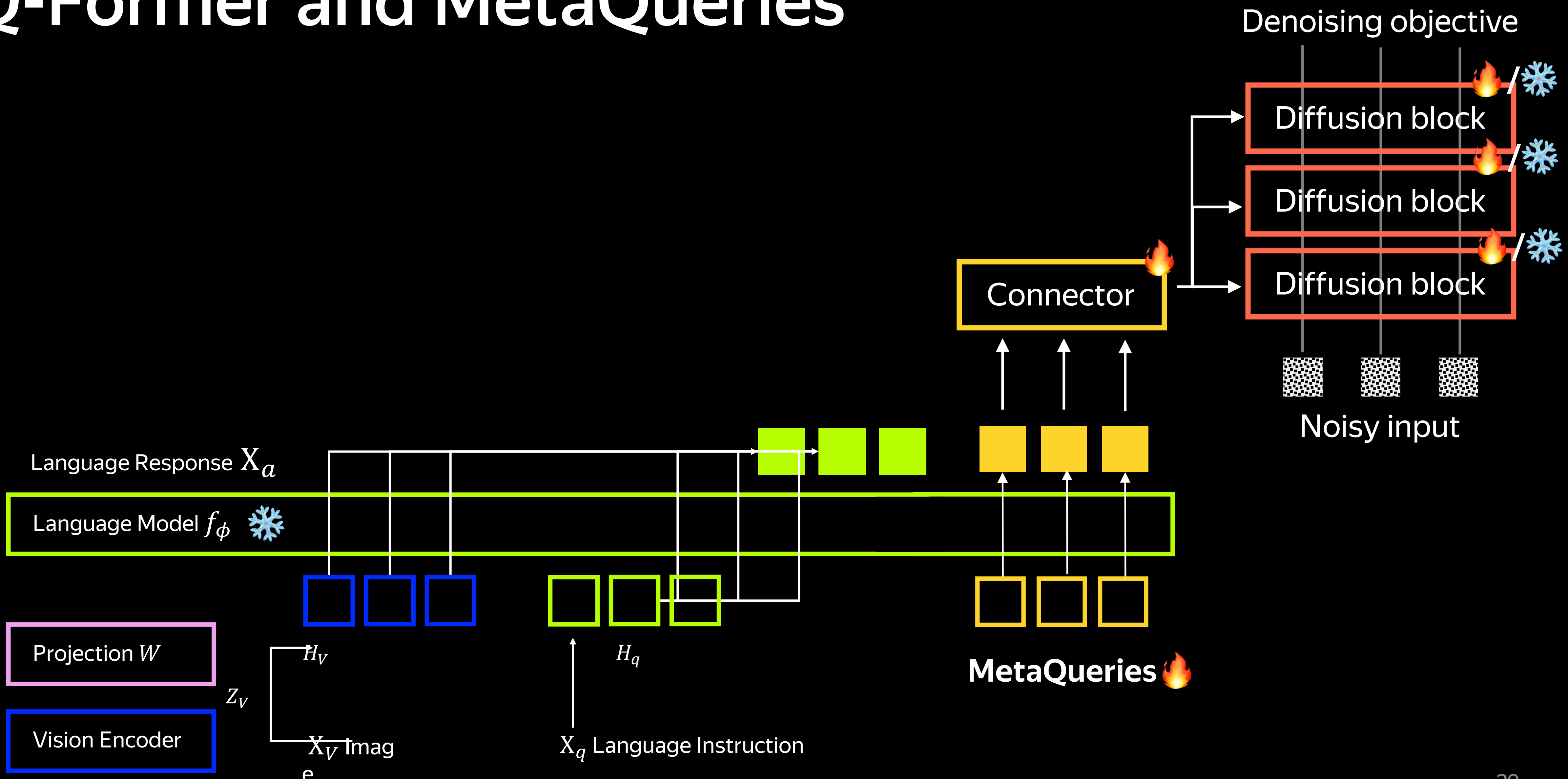
Baseline connection scheme



Baseline connection scheme

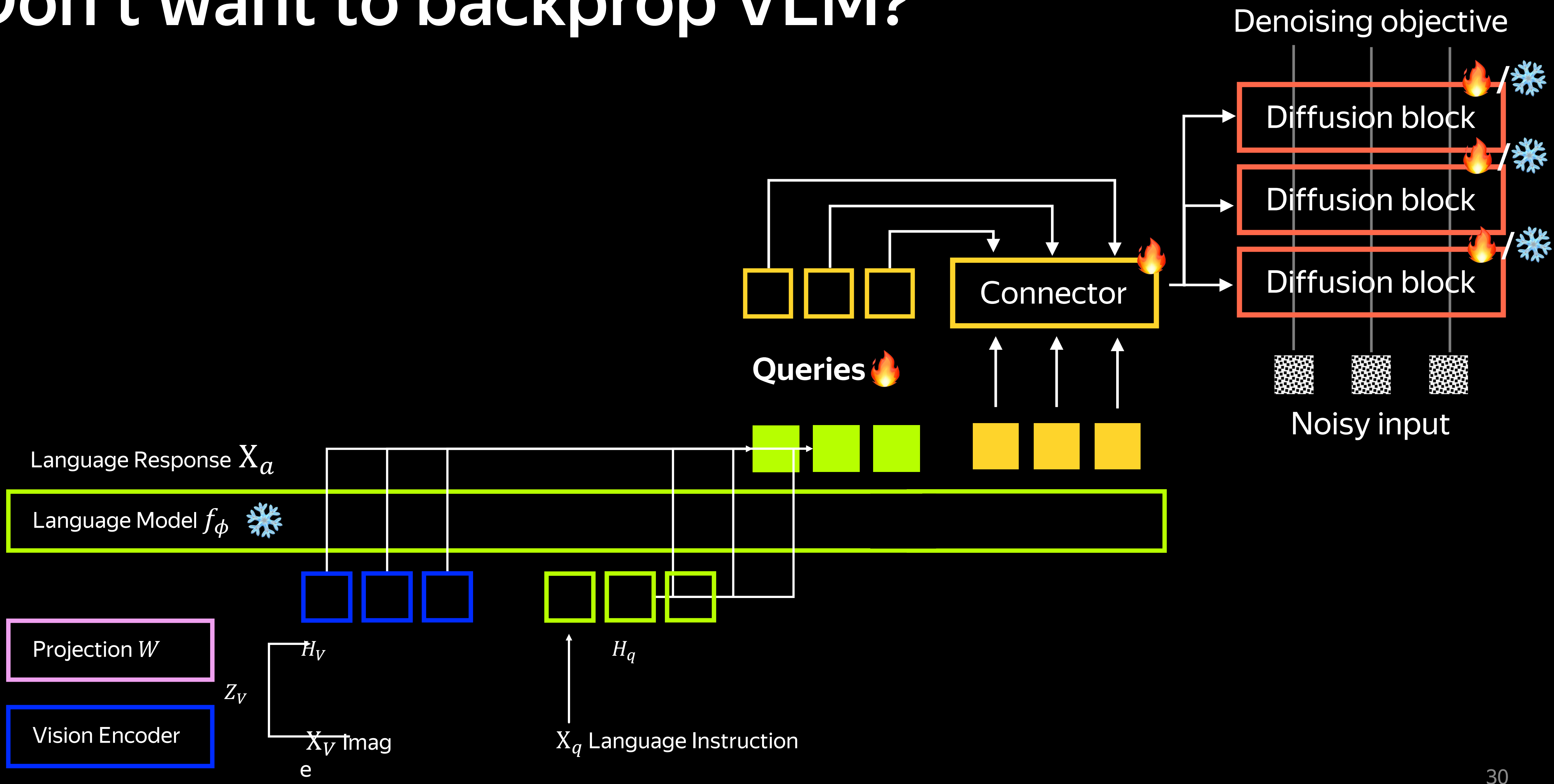


Q-Former and MetaQueries

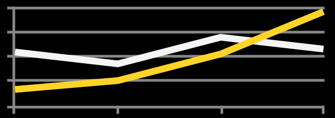


X. Pan et. al., "Transfer between Modalities with MetaQueries", 2025

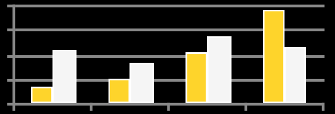
Don’t want to backprop VLM?



Recap

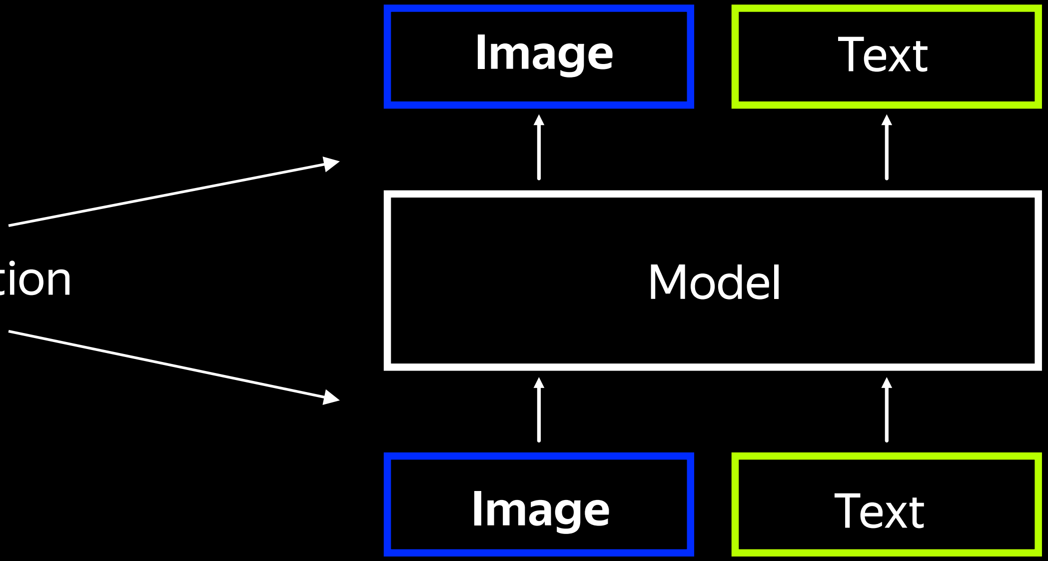


Continuous



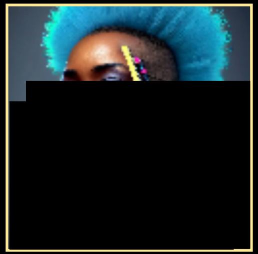
Discrete

Images representation



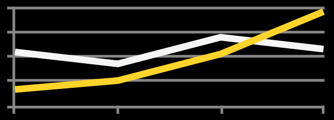
Recap

Diffusion

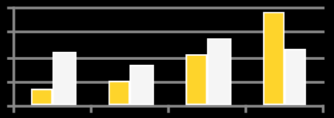


Autoregression

Loss

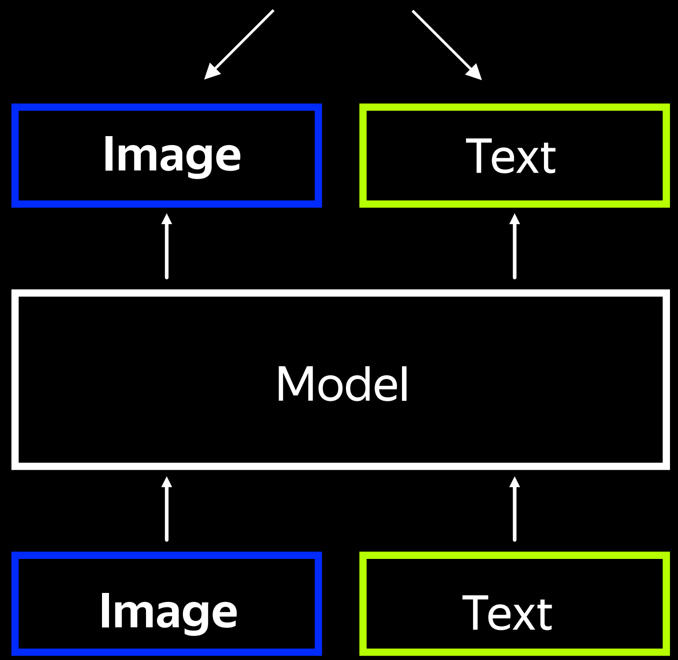


Continuous



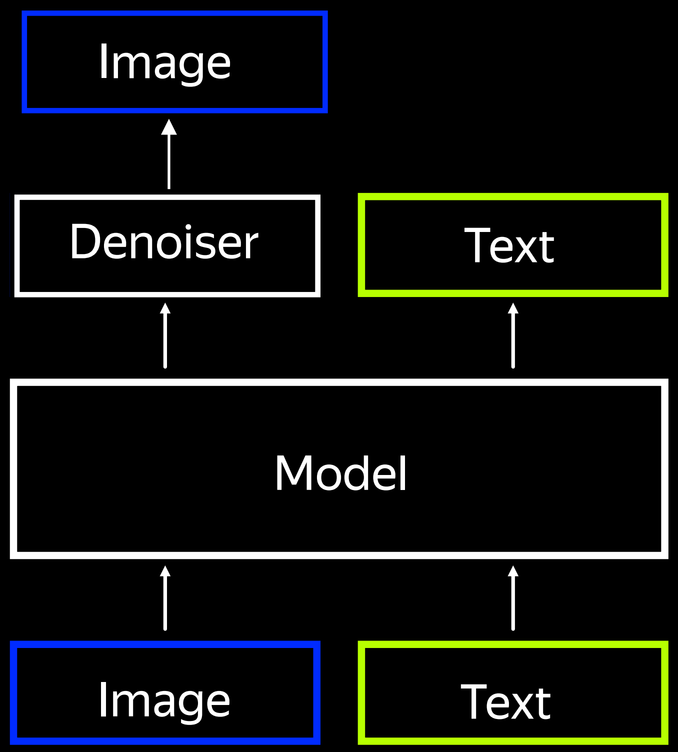
Discrete

Images representation

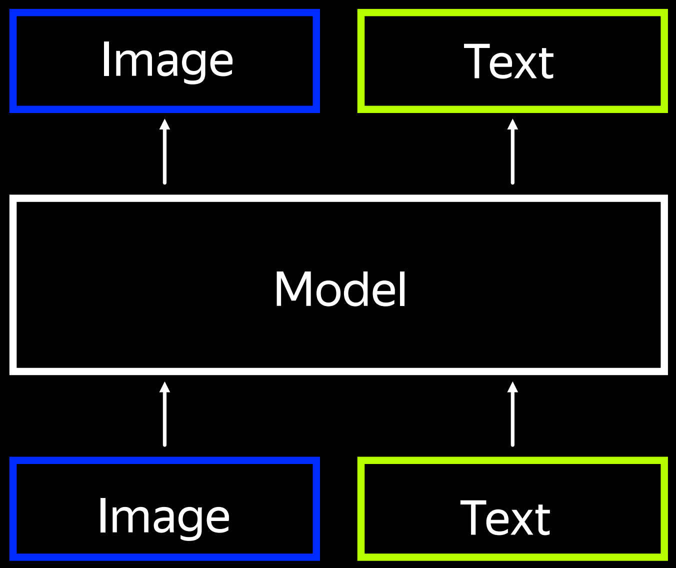


Recap

Image gen module



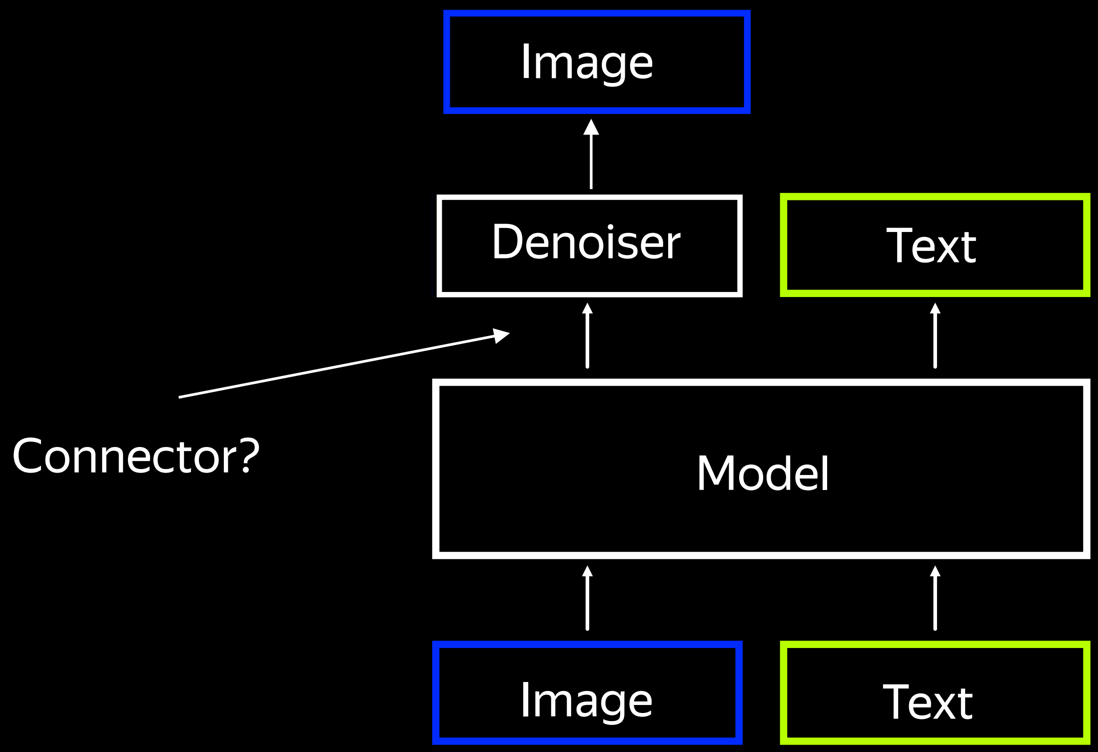
External



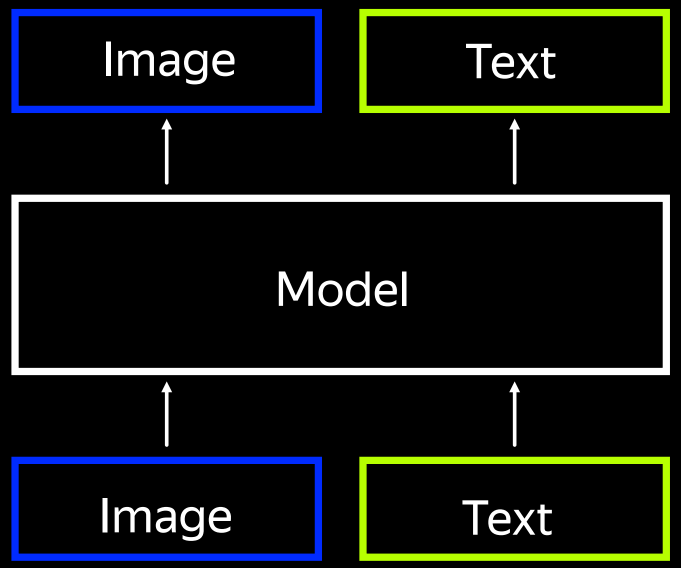
Integrated

Recap

Image gen module



External

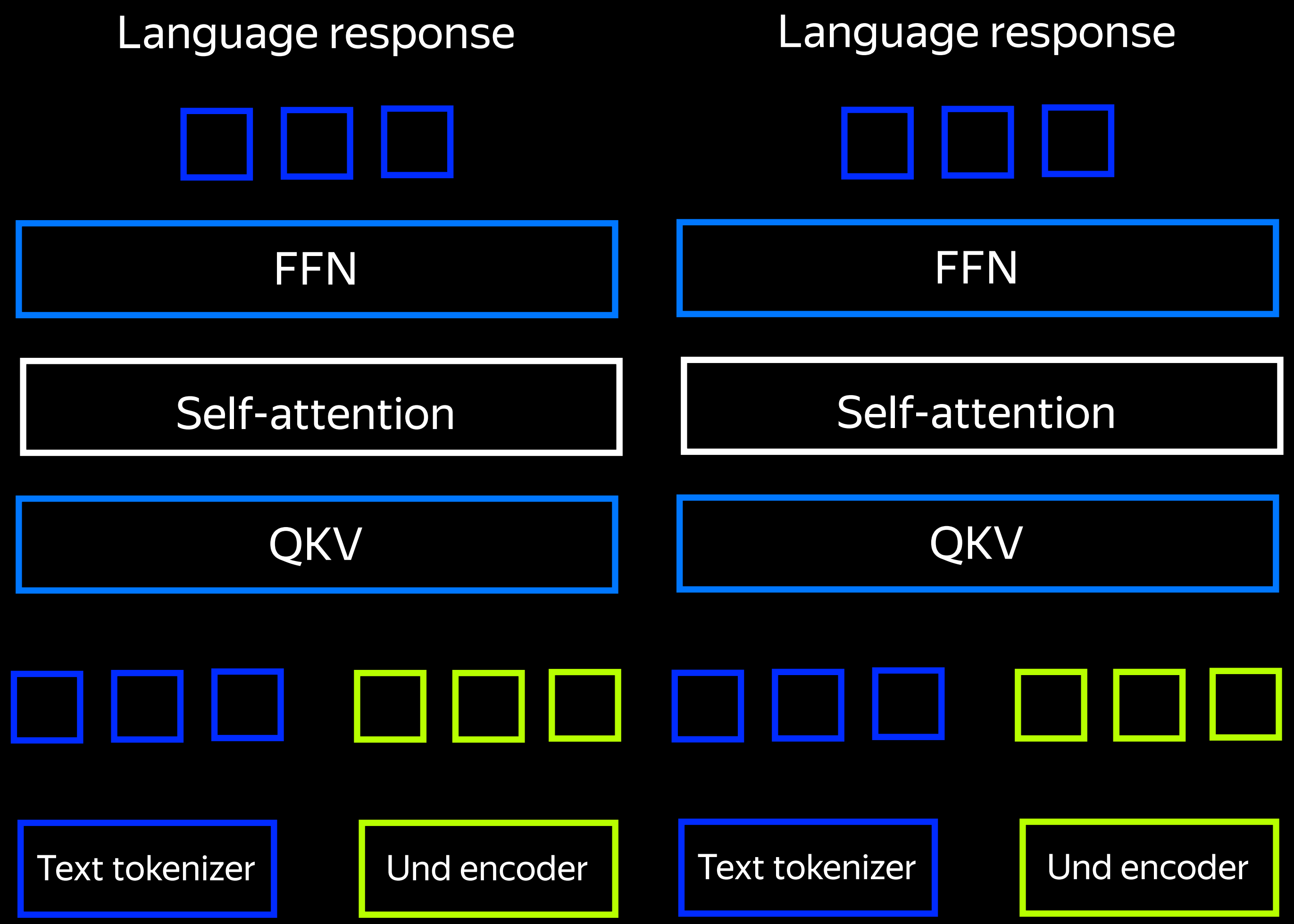


Integrated

BAGEL

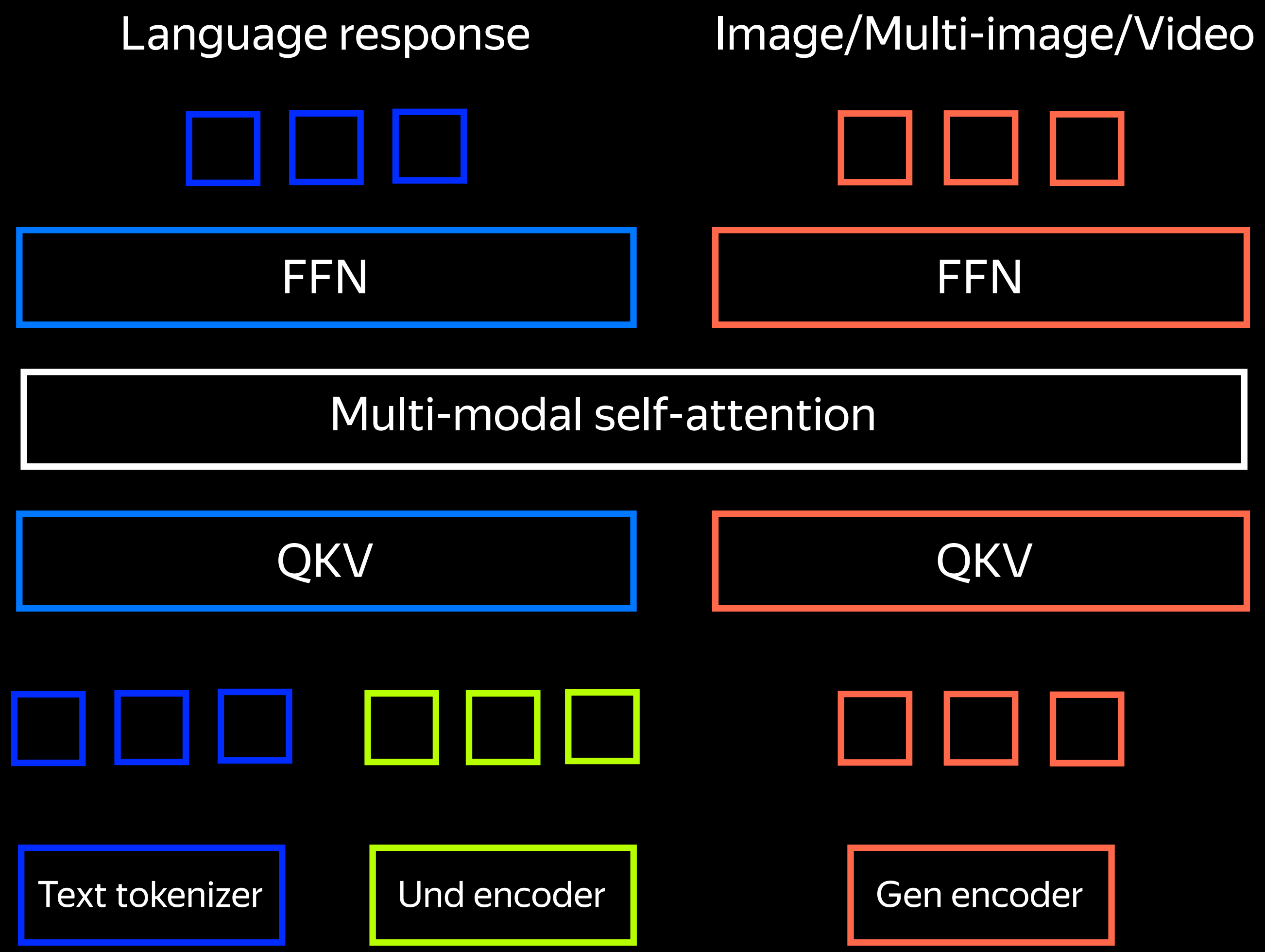
C. Deng et. al., “Emerging Properties in Unified Multimodal Pretraining”, 2025

BAGEL



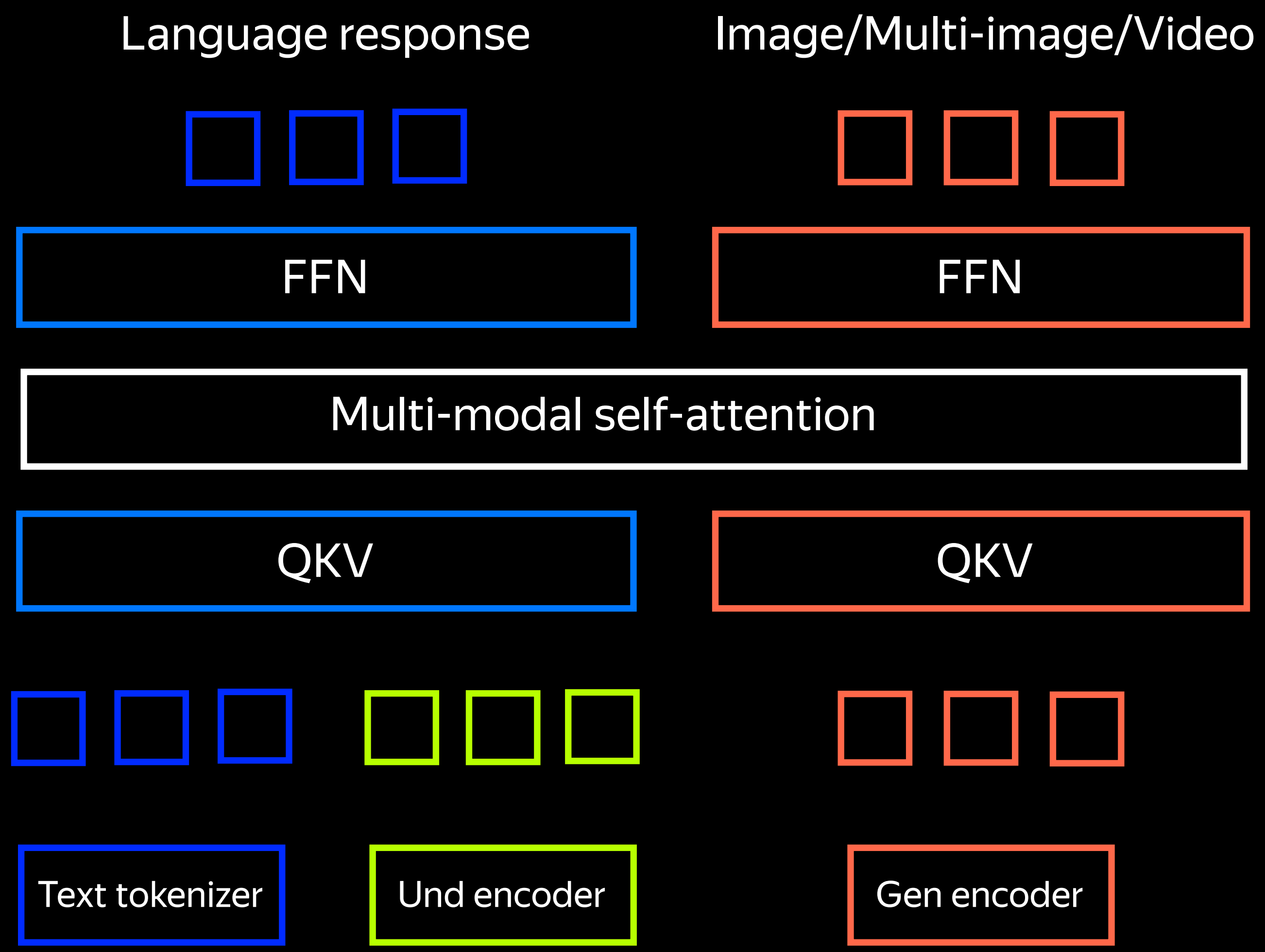
C. Deng et. al., “Emerging Properties in Unified Multimodal Pretraining”, 2025

BAGEL



C. Deng et. al., “Emerging Properties in Unified Multimodal Pretraining”, 2025

BAGEL



🤔 Integrated?

C. Deng et. al., “Emerging Properties in Unified Multimodal Pretraining”, 2025

Final recap

- Plenty of options to choose
- Options have ups and downs
- There is no clear winner yet

- Many open questions

Yandex Research

(Open) Questions

How to Train

Currently, we can model

- text-to-text — LLM
- text-to-image — text-cond diffusion
- image-to-text — VLM
- text+image-to-image — text+img-cond diffusion

How to Train

Currently we can model

- text-to-text — LLM
- text-to-image — text-cond diffusion
- image-to-text — VLM
- text+image-to-image — text+img-cond diffusion

The moment you start you be like: <u>how the hell do I combine all that?</u>

Is there cross-modal knowledge transfer?

- Editing-only training collapses T2I
- Training on T2I + editing > just T2I or just editing
- Training on all 4 tasks is not worse than training on them individually

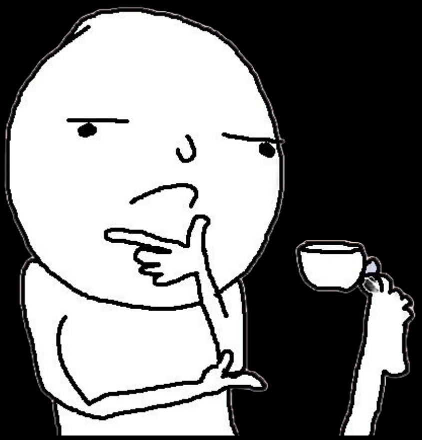
* all conclusions are from our internal experiments

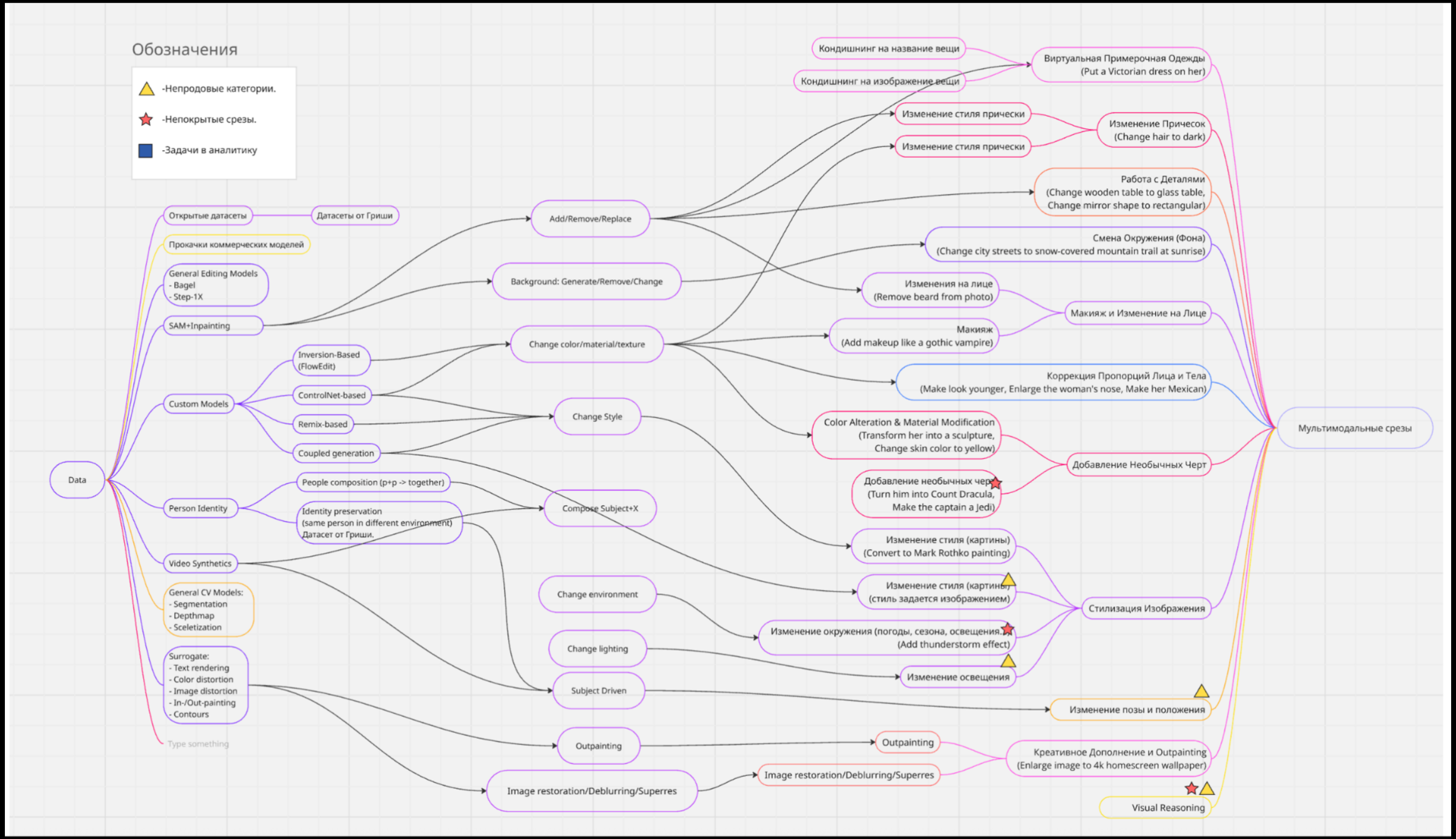
Adapted setup does magic




* all conclusions are from our internal experiments

DATA





Обозначения
-Непродовые категории.
-Непокрытые срезы.
-Задачи в аналитику
Data
Открытые датасеты
Датасеты от Гриши
Прокачки коммерческих моделей
General Editing Models
- Bagel
- Step-1X
SAM+Inpainting
Custom Models
Inversion-Based (FlowEdit)
ControlNet-based
Remix-based
Coupled generation
Person Identity
People composition (p+p -> together)
Identity preservation (same person in different environment) Датасет от Гриши.
Video Synthetics
General CV Models:
- Segmentation
- Depthmap
- Sceletization
Surrogate:
- Text rendering
- Color distortion
- Image distortion
- In-/Out-painting
- Contours
Type something
Add/Remove/Replace
Background: Generate/Remove/Change
Change color/material/texture
Change Style
Compose Subject+X
Change environment
Change lighting
Subject Driven
Outpainting
Image restoration/Deblurring/Superres
Кондишнинг на название вещи
Кондишнинг на изображение вещи
Виртуальная Примерочная Одежды (Put a Victorian dress on her)
Изменение стиля прически
Изменение стиля прически
Изменение Причесок (Change hair to dark)
Работа с Деталями (Change wooden table to glass table, Change mirror shape to rectangular)
Смена Окружения (Фона) (Change city streets to snow-covered mountain trail at sunrise)
Изменения на лице (Remove beard from photo)
Макияж (Add makeup like a gothic vampire)
Макияж и Изменение на Лице
Коррекция Пропорций Лица и Тела (Make look younger, Enlarge the woman's nose, Make her Mexican)
Color Alteration & Material Modification (Transform her into a sculpture, Change skin color to yellow)
Добавление необычных черт (Turn him into Count Dracula, Make the captain a Jedi)
Добавление Необычных Черт
Мультимодальные срезы
Изменение стиля (картины) (Convert to Mark Rothko painting)
Изменение стиля (картины) (стиль задается изображением)
Изменение окружения (погоды, сезона, освещения...) (Add thunderstorm effect)
Изменение освещения
Стилизация Изображения
Изменение позы и положения
Outpainting
Image restoration/Deblurring/Superres
Креативное Дополнение и Outpainting (Enlarge image to 4k homescreen wallpaper)
Visual Reasoning

How to Distill

- Copy-paste from T2I does not (optimally) work
- Fun cross-task interactions:
  - LCM on editing > LCM on T2I for T2I quality 🥴

* all conclusions are from our internal experiments

How to Distill

- For unmerged models, we can combine acceleration techniques
    - LoRA distill on the image part
    - Spec dec, KV-caching etc on the text part
- Can we use text-gen acceleration techniques on image-gen? 🤔

Yandex Research

Inference-time Compute Scaling

Very much known for LLMs

Question

↓

“Reasoning”
Large Language Model

↓

Thinking w/text...

Answer!

Very much known for LLMs

Question

↓

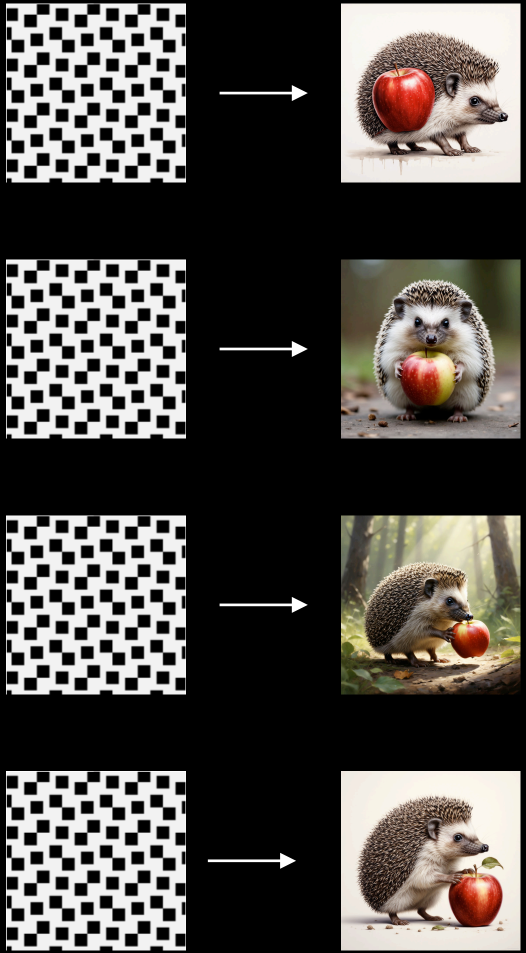
"Reasoning"
Diffusion/Multi-modal model

↓

Thinking w/text+images...

Answer!

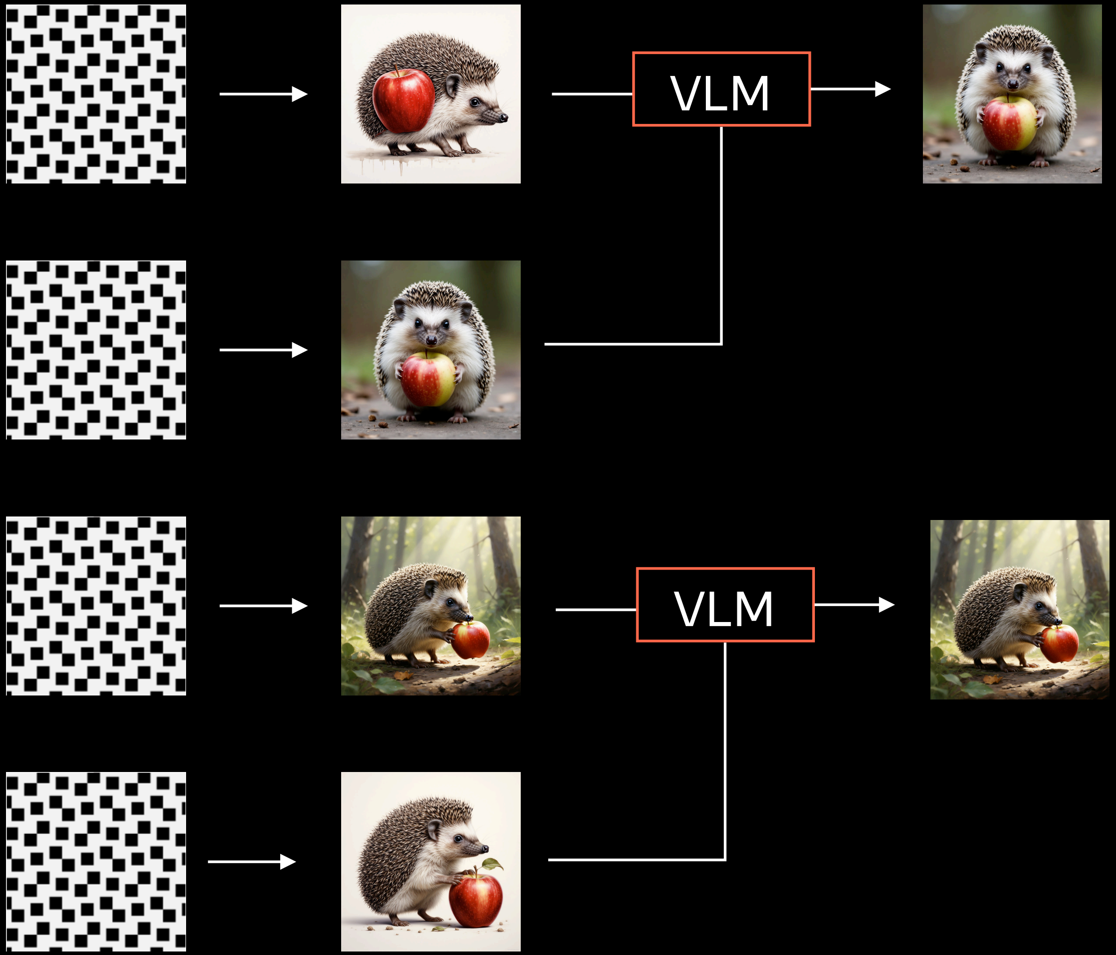
Best-on-N baseline



“hedgehog with an apple”

E. Xie et. al., “SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer“, 2025

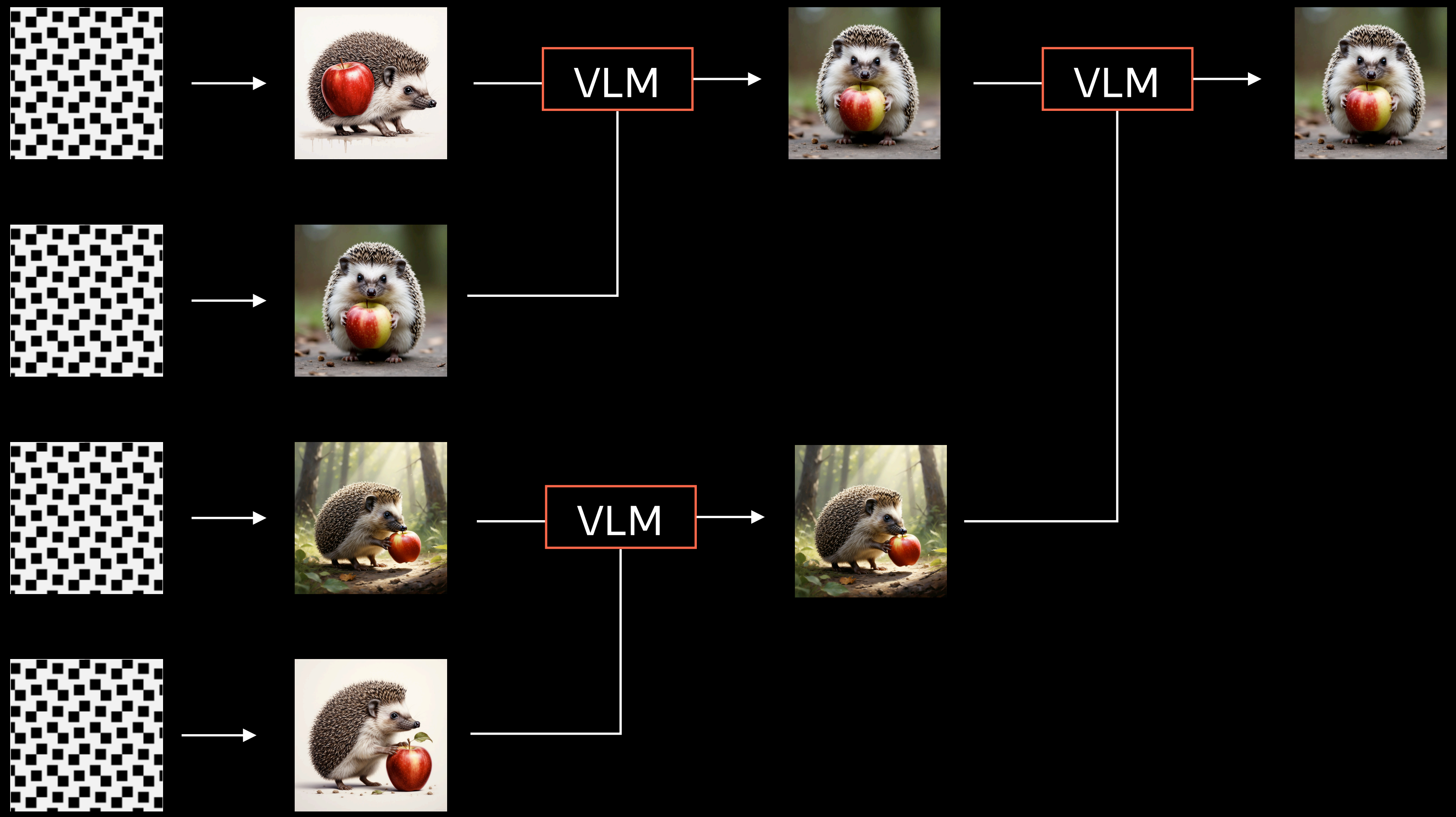
Best-on-N baseline



“hedgehog with an apple”

E. Xie et. al., “SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer“, 2025

Best-on-N baseline



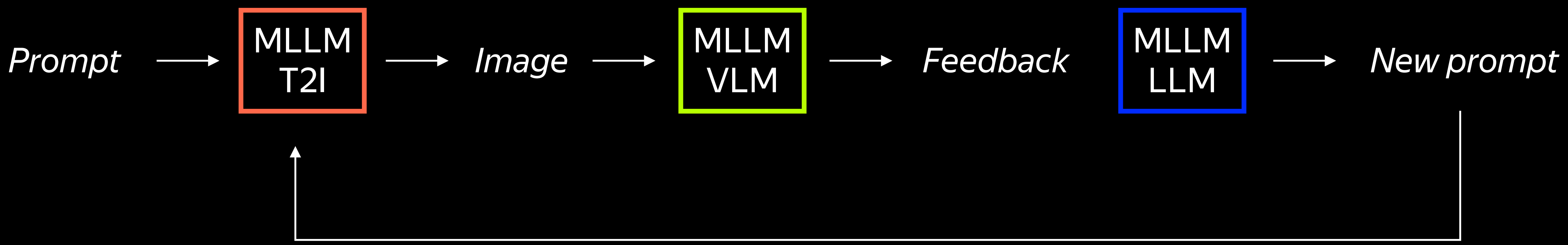
“hedgehog with an apple”

E. Xie et. al., “SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer“, 2025

External VLM-as-judge

M. Abdul et. al., “Test-time Prompt Refinement for Text-to-Image Models“, 2025

Can multi-modal model do all that itself?



Boosts text gen and `many objects`

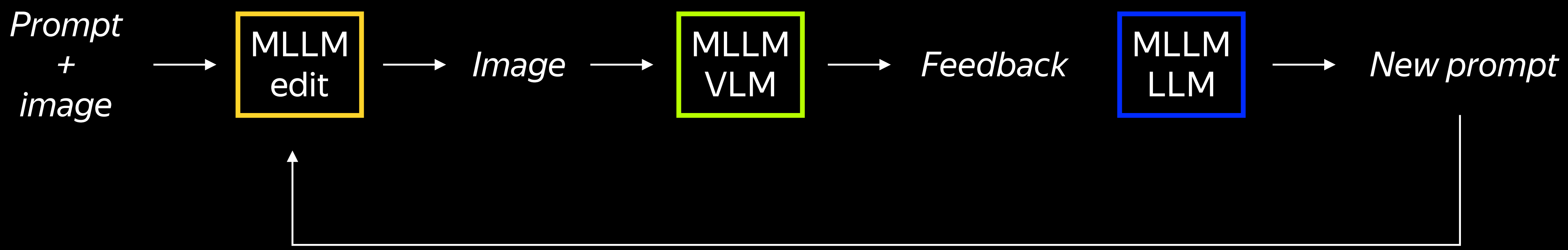
No TTS | TTS

“do an inscription `rash 435`”

“draw 5 dogs of different colors”

* preliminary results

Can we do the same for editing?



Also works pretty well

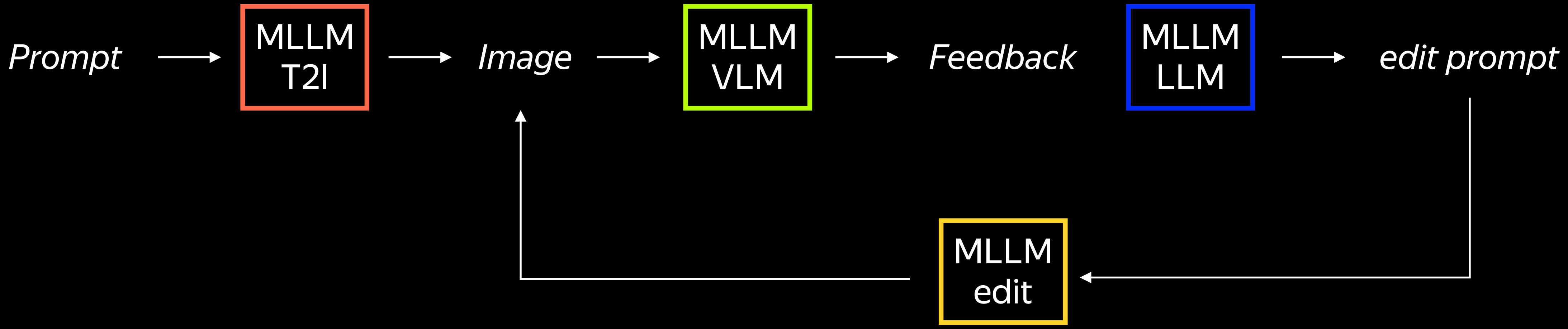
No TTS | TTS

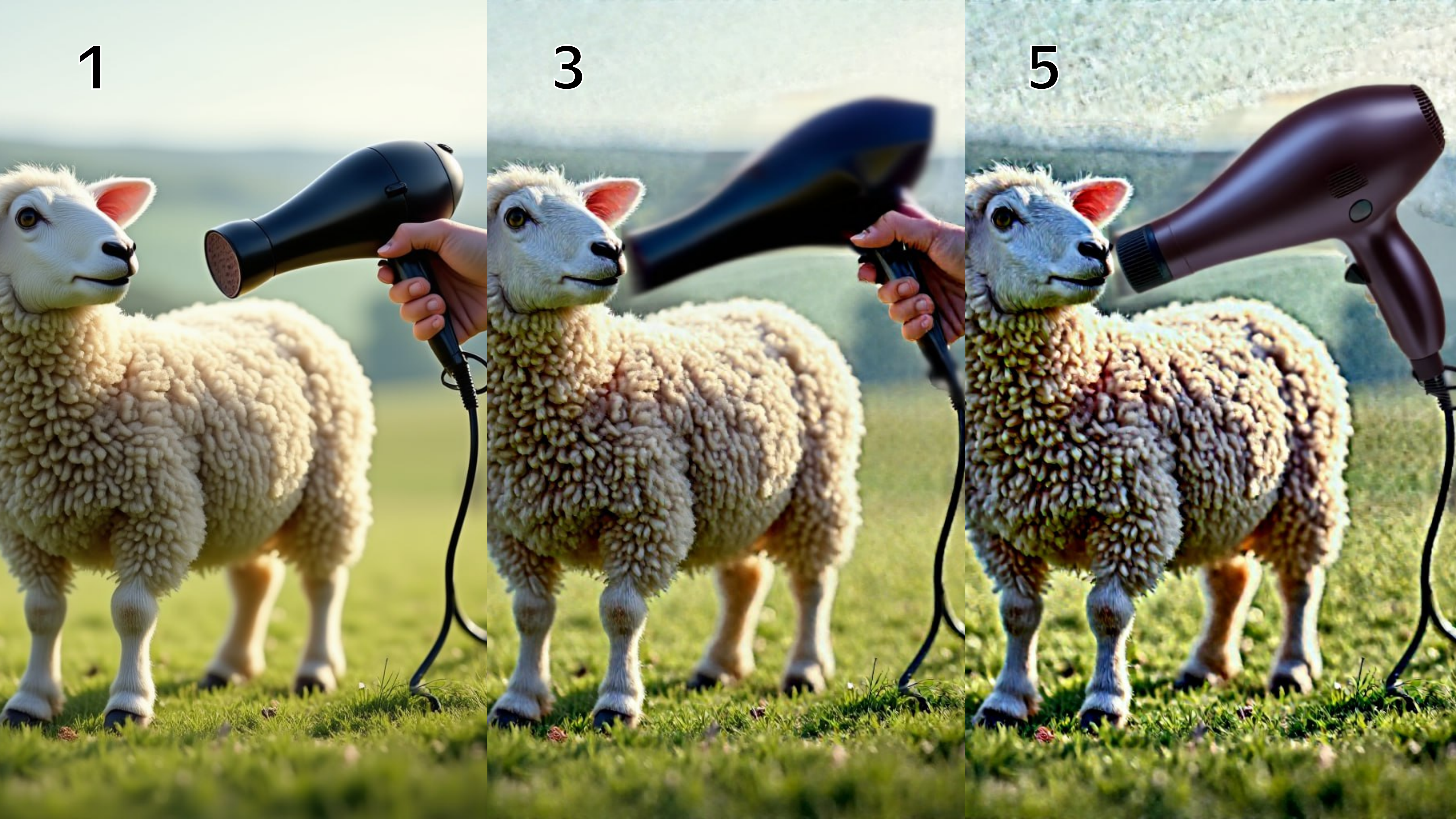
“change the action of the horses to galloping”

“Extract the person from the photo and dress them in a police uniform”

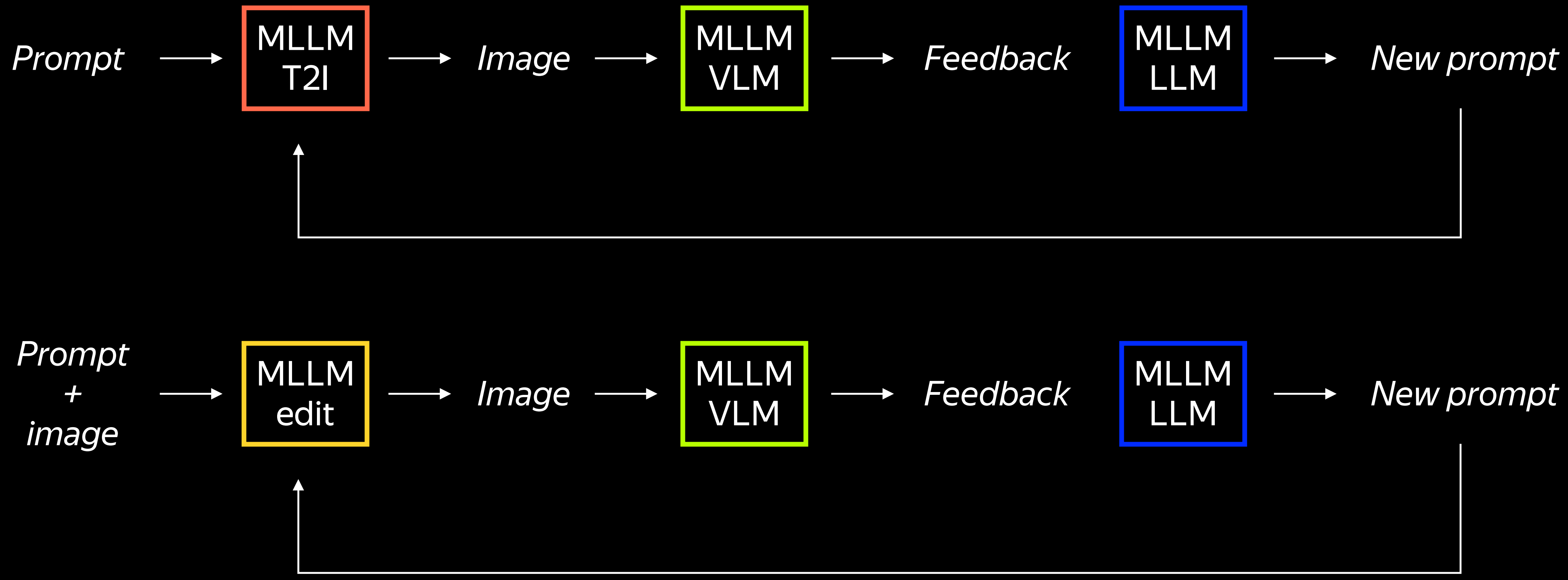
* preliminary results

Use all four tasks?




1
3
5

Distill again?



What stops us from training a reasoner on this synthetic chains?

Instead of Conclusion

Yandex Research also works on and presents:

Gen Models

Poster 39
Alchemist: Turning Public Text-to-Image Data into Generative Gold

Poster 38
Inverse Bridge Matching Distillation

Tabular DL

Poster 25
TabM: Advancing tabular deep learning with parameter-efficient ensembling

Poster 62
Leveraging Coordinate Momentum in SignSGD and Muon: Memory-Optimized Zero-

Effective Inference

Poster 42
AutoJudge: Judge Decoding Without Manual Annotation

Poster 48
Hogwild! Inference: Parallel LLM Generation via Concurrent Attention

Graph ML

Poster 27
GraphLand: Evaluating Graph Machine Learning Models on Diverse Industrial Data

Optimization

Poster 83
On Linear Convergence in Smooth Convex-Concave Bilinearly-Coupled Saddle-Point Optimization: Lower Bounds and Optimal Algorithms

Yandex Research

Read research papers

Yandex Research

Research with us

Develop Tech with us