Multimodality for story-level understanding and generation of visual data

Vicky Kalogeiton

Paris Generative AI Autumn School

25/10/2024

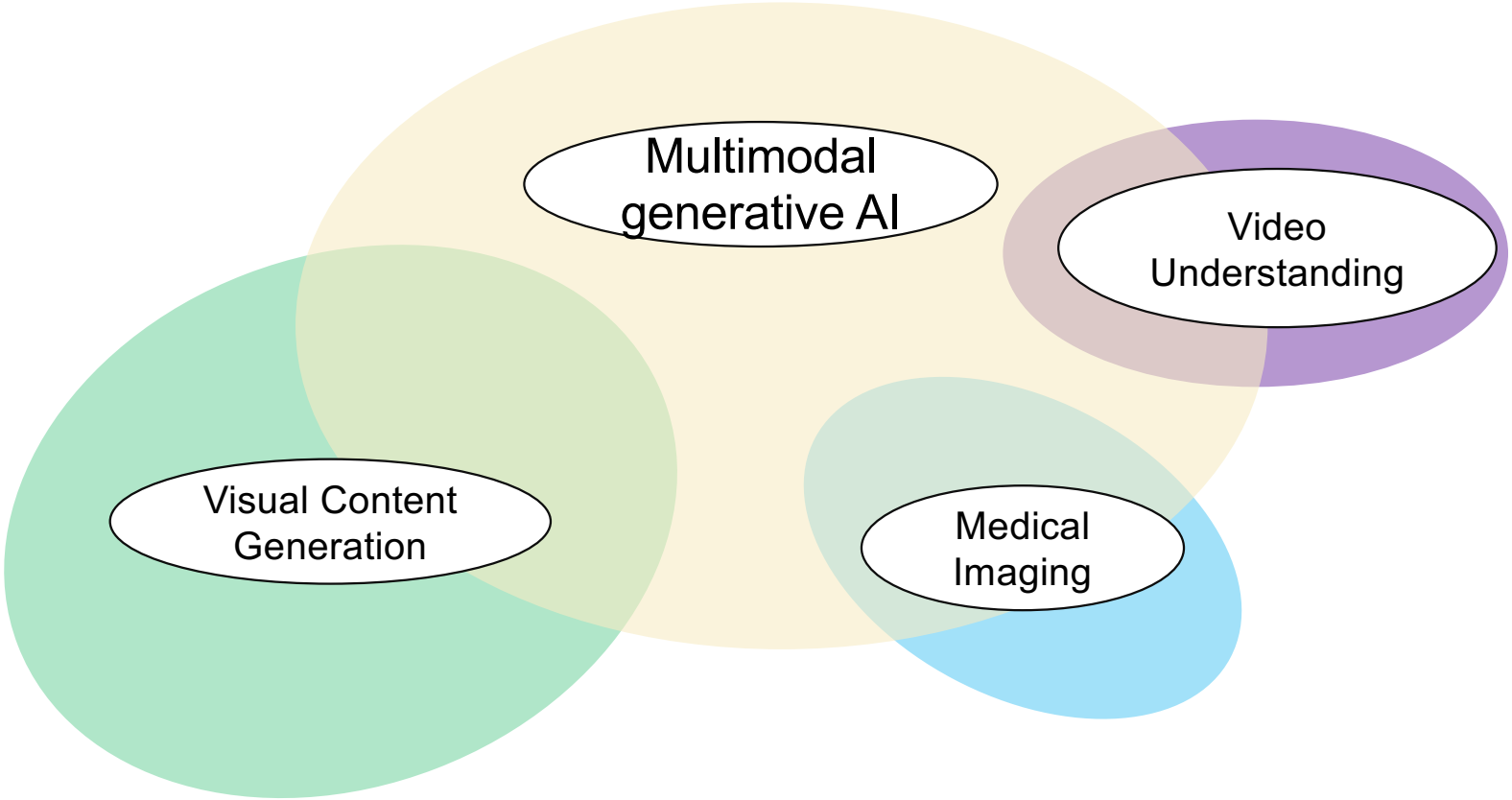
About me

- *Assistant Professor*, 2020 –
  - VISTA Group, Ecole Polytechnique, France

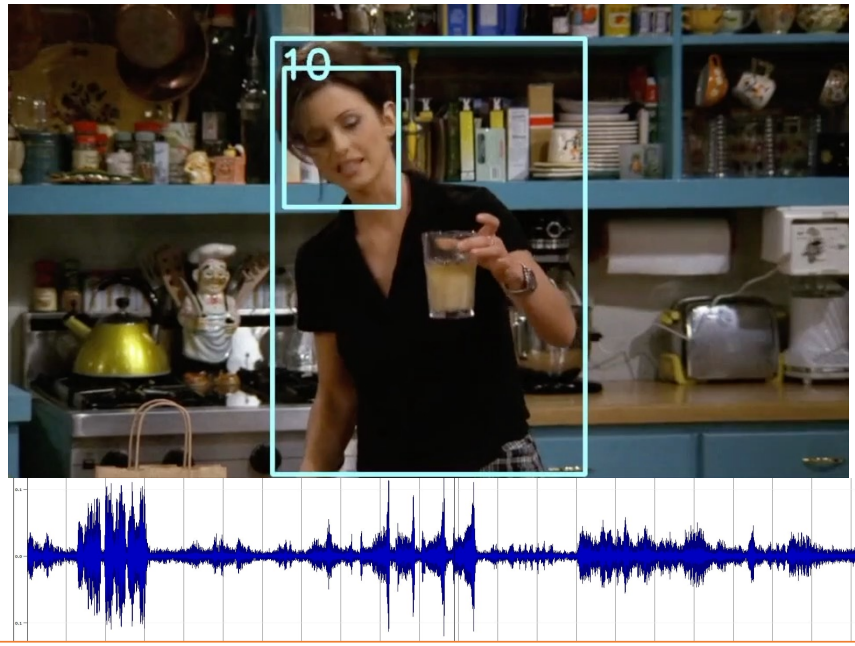
- *Research Fellow,* 2019 – 2021
- *Post-doc,* 2018 – 2019
  - Visual Geometry Group, University of Oxford, UK
  - Andrew Zisserman

- *PhD*, 2013 – 2017
  - University of Edinburgh, UK, INRIA, Grenoble, France
  - Vittorio Ferrari, Cordelia Schmid

Research agenda



Multimodality: video, audio, text



Subtitles

- **Rachel:** You guys, do this look like something the girlfriend of a paleontologist would wear?
- **Phoebe:** I don't know. You might be the first one

Low-level understanding

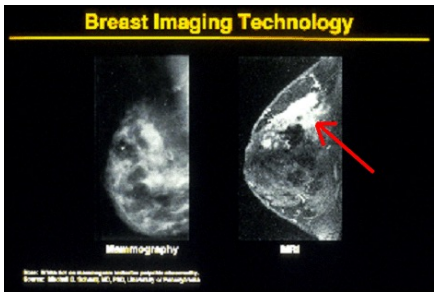
- Characters (Rachel, Phoebe, …)
- Located in an apartment
- Winter (clothing)

High-level reasoning

- Anxious (what to wear, …)
- Interactions
- Joke to diffuse the situation

Why does it matter?

Safety



Health

Security

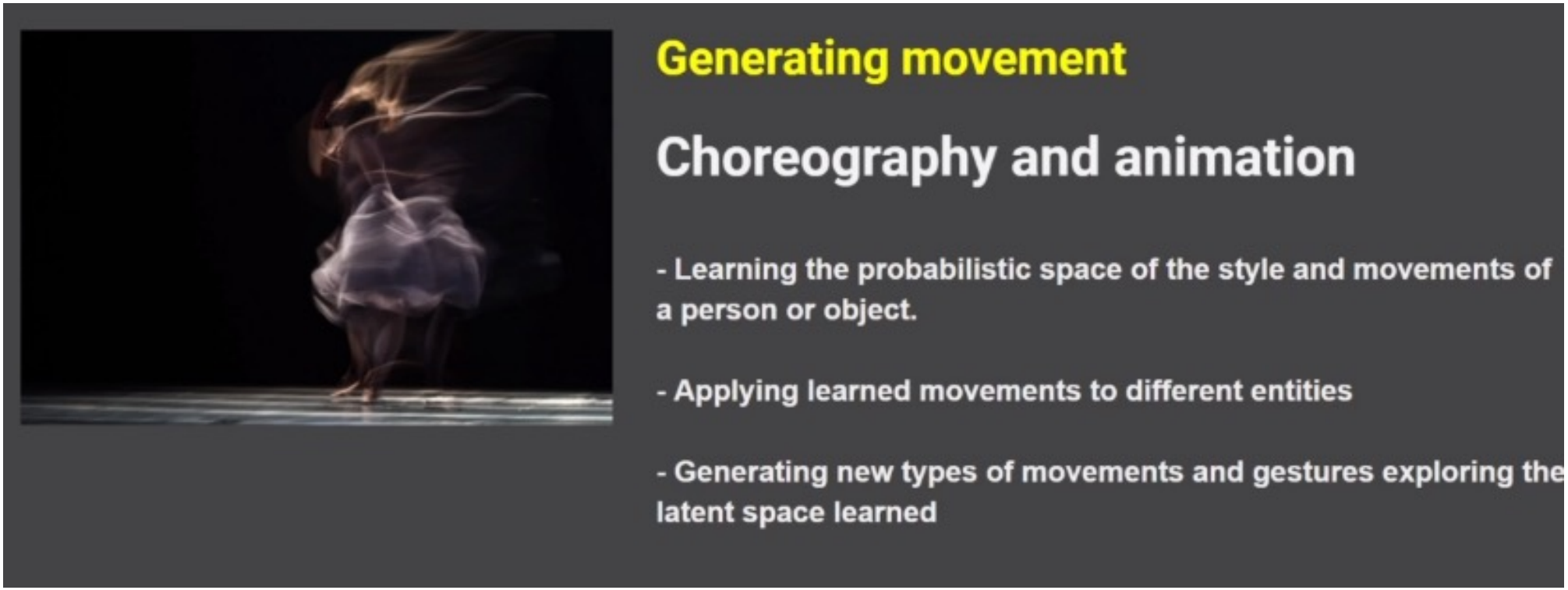
Comfort

Fun

Access

Generative AI

Choreography and animation



Automatic story telling

Grand entry of the king's horses and men. ARYA, wearing a helm and cloak, pushes her way into a tall wagon for a better look….

Grand entry of the king's horses and men. ARYA, wearing a helm and cloak, pushes her way into a tall wagon for a better look…. In rides JOFFREY, followed by the HOUND

Game of Thrones

Transforming Business with Generative AI

- Innovating product design
  - Generative Games, word synthesis
  - Shoes, jewelry, clothing, …
  - Rapid Prototyping
  - Material Innovation
- Enhancing customer interaction
  - Personalized recommendations
  - Automated customer support
  - Interactive virtual assistant
  - Sentiment Analysis
- Streamlining content creation
  - 3D house models
  - AI-assisted Video Production, Choreography
  - Dynamic Web Content
  - Marketing Material Creation

My personal long-term goal

- "`Once upon a time in a faraway land, there was a dragon called Zoe and she was different…`" started the mother's bedtime story
  - Video → visual illustration
  - Colors, sounds, characters → story
  - Characters and music → personalized

Story-level generation

Long-term story level understanding

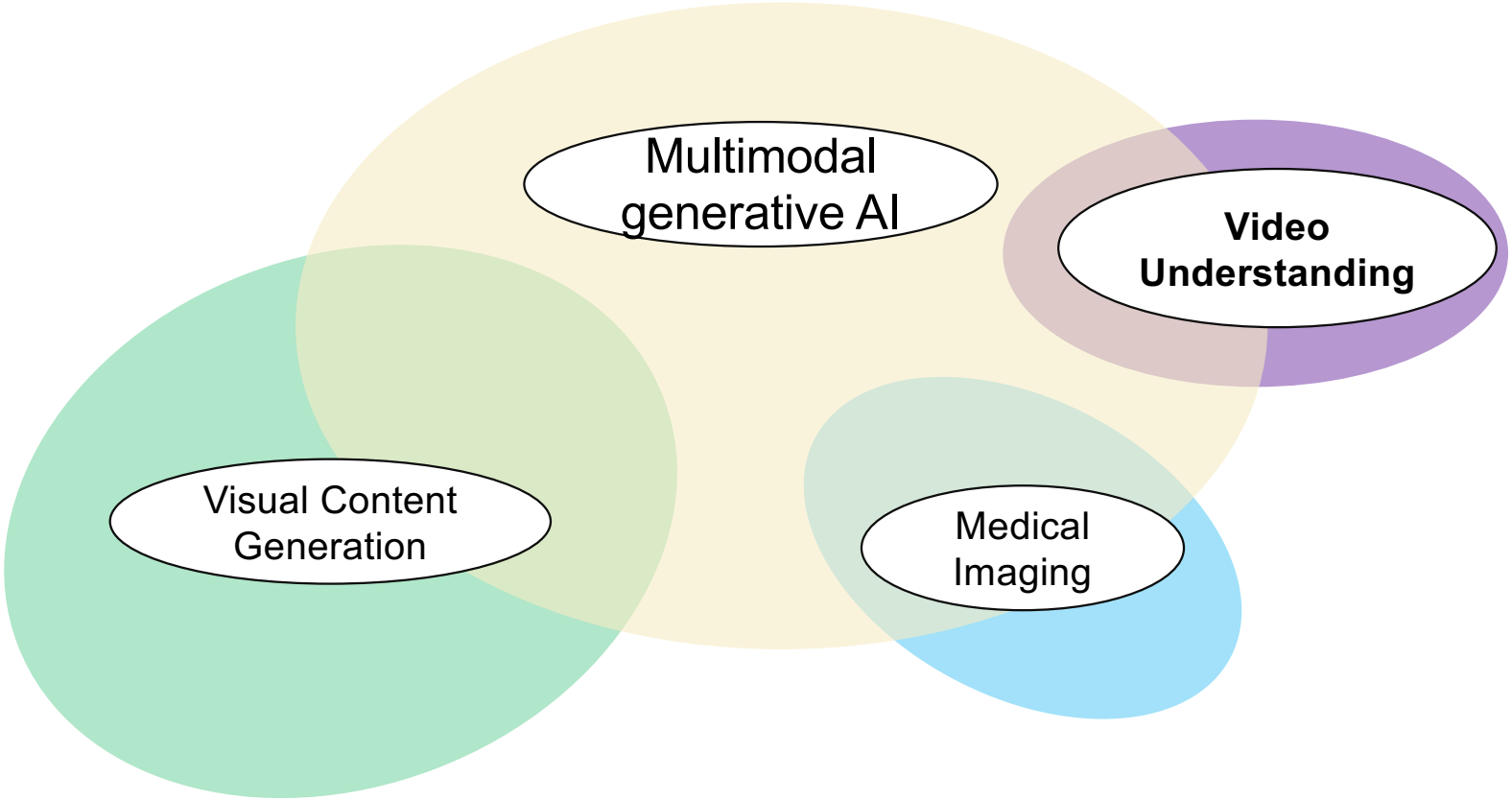
- Multimodality
- Long-term reasoning

→

Text-to-video generation

- Dynamic storytelling techniques
- Text to visual content alignment

Research agenda



Part I: Multimodal Video Understanding

- FunnyNet-W: Multimodal Learning of Funny Moments in Videos

- Movie Question Answering

Is this funny?

Multimodal Learning of Funny Moments in Videos

Mia: Three tomatoes are walking down the street -- a poppa tomato, a momma tomato, and a little baby tomato. Baby tomato starts lagging behind.
Poppa tomato gets angry, goes over to the baby tomato, and squishes him... and says, …. "Catch up."

[Video scene from Pulp fiction, 1994, source video: https://www.youtube.com/watch?v=4L5LjjYVsHQ]

FunnyNet-W: Multimodal Learning of Funny Moments in Videos in the Wild



[IJCV 2024, ACCV 2022, Oral, Honorable Mention Award
Z.S. Liu, R. Courant, V. Kalogeiton]

Code & demo:
https://www.lix.polytechnique.fr/vista/projects/2024_ijcv_liu/

CIHE Caritas Institute of Higher Education 明愛專上學院

VISTA
Visual Worlds: Temporal Analysis, Animation and Authoring

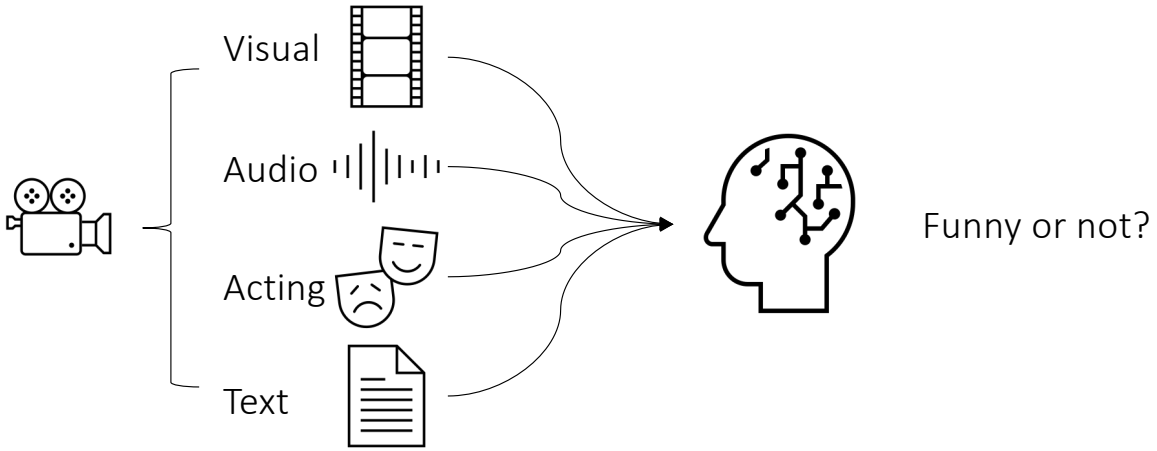
Why does it matter ?

In-the-wild funny moment detection: detecting funny moments in any content.

Human-machine interactions: making conversational AI more spontaneous.

Make computers funny! comprehending what is funny.

Background and introduction



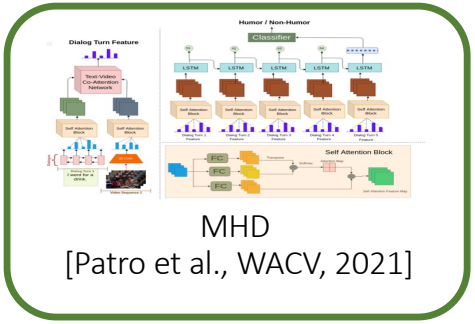
Understanding funnyness: complex → Purely visual / auditory / mix both

Multimodality

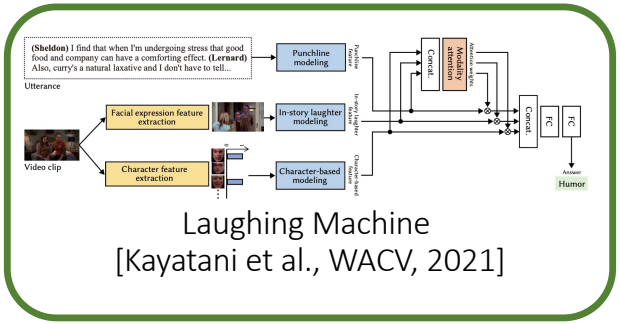
No recipe for the perfect joke!

Related work

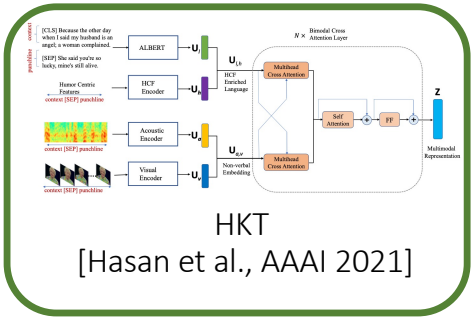
MUStARD
[Castro et al., ACL, 2019]



MHD
[Patro et al., WACV, 2021]



Laughing Machine
[Kayatani et al., WACV, 2021]



HKT
[Hasan et al., AAAI 2021]

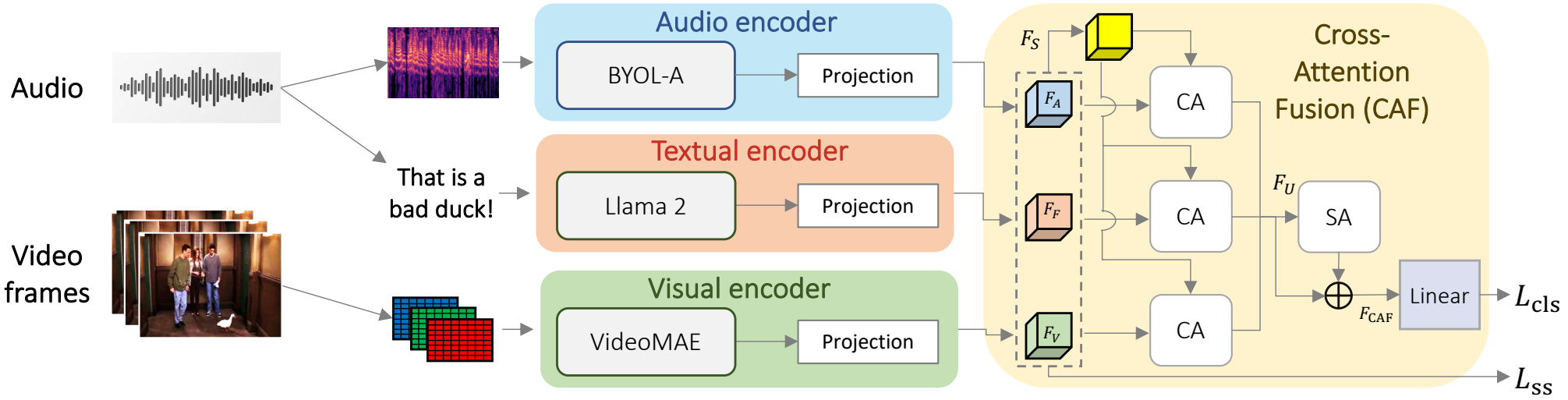
Focus mainly on textual modality

- Limited, imperfect
- Not flexible

Ours

+ Exploit only raw video modalities: audio, visual, text
+ Self-supervised: use canned laughter for supervision

Method: FunnyNet-W



Audio encoder: $F_A \in \mathbb{R}^{512}$

Visual encoder: $F_V \in \mathbb{R}^{512}$

Text encoder: $F_T \in \mathbb{R}^{512}$

$\oplus \rightarrow$ Fused vector: $F_U \in \mathbb{R}^{3\times512}$

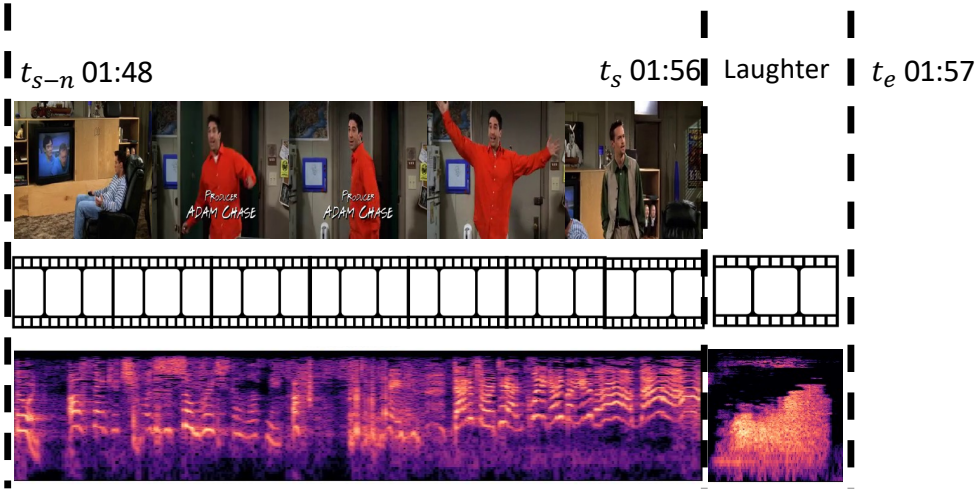
Cross Attention (CA) for across-cue correlations

$$F_S = \sum_{i\in\{A,V,T\}} \sigma\left(\frac{Q_U K_i^T}{\sqrt{d}}\right) V_i$$

Self Attention (SA) for inter correlations

$$F_{CAF} = F_S + \sigma\left(\frac{Q_S K_S^T}{\sqrt{d}}\right) V_S$$

Training and data pre-processing



Laughter at timestep (*ts*, *te*) →

- Positive: N-sec clip followed by laughter
  i.e., *N*-sec clip (*ts*−*n*, *ts*) split into audio and video
- Negative: N-sec clip not followed by laughter

Comparison to the state of the art

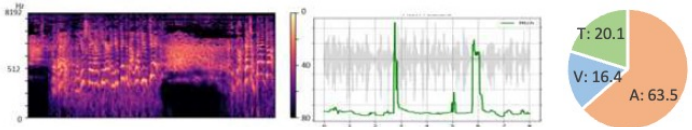
Method / Metrics	Wild	TBBT		MHD		MUStARD		UR-Funny		Friends	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Random	–	46.3	50.0	56.1	50.9	48.3	48.7	50.2	50.2	51.0	51.0
All positive	–	60.3	43.2	75.6	60.8	66.7	50.0	75.4	50.7	66.7	50.0
All negative	–	0.0	56.8	0.0	39.2	0.0	50.0	0.0	49.3	0.0	50.0
MUStARD 2019 (V+A+T^{gt}) [9]	–	-	-	-	-	71.7	71.8	-	-	-	-
MSAM 2021 (V+T^{gt}) [69]	–	-	-	81.3	72.4	-	-	-	-	-	-
MISA 2020 (V+A+T^{gt}) [31]	–	-	-	-	-	-	66.2	-	69.8	-	-
HKT 2021 (V+A+T^{gt}) [29]	–	-	-	-	-	-	79.4	-	77.4	-	-
LaughM† 2021 (T^{gt}) [39]	–	64.2	70.5	**86.5**	76.3	68.6	68.7	71.9	67.6	74.7	59.8
FunnyNet (V+A+T^{gt}) [51]	–	73.8	75.8	83.4	78.6	79.5	79.9	84.1	79.9	88.2	85.8
FunnyNet (V+F+A+T^{gt}) [51]	–	75.9	78.3	85.2	79.6	83.2	82.0	84.4	80.2	88.8	86.4
FunnyNet-W (V+A+T^{gt})	–	**78.5**	**80.0**	84.6	**80.1**	**85.9**	**84.1**	**84.5**	**80.2**	**89.3**	**86.7**

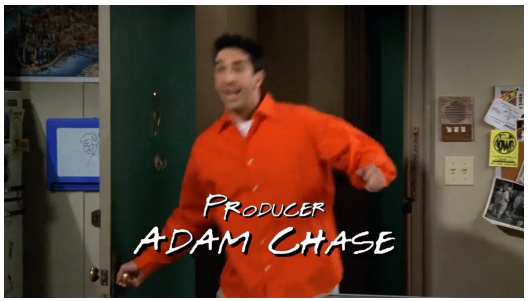
Qualitative results: Modality impact

Funny predictions

Chandler: Okay! Now you stay out here and think about what you did!!

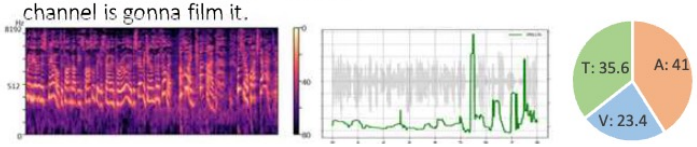
Ross: That is a duck.
Chandler (high pitch): That is a bad duck!!





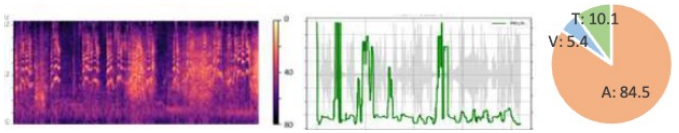
Ross: they are putting together this panel to talk about fossils they just found in Peru and the Discovery channel is gonna film it.

Chandler: Oh my God! (pause) Who is gonna watch that?



Phoebe: I am setting the phone down. Don't go anywhere, I am still here.

Phoebe (speech rate change): One sec! One second! Wait! One second! Just!



- Positive: high pitch, pause, speech rate change indicate punchline for laughter

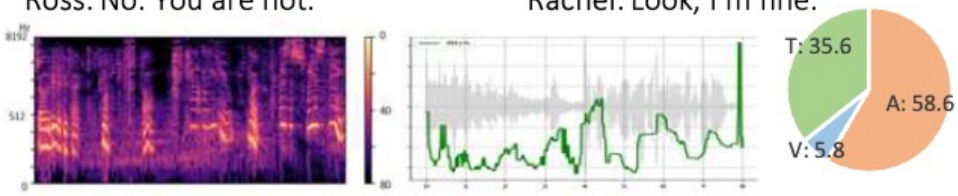
FunnyNet-W: Text FunnyNet-W: Visual FunnyNet-W: Audio

Qualitative results: Modality impact

Not-funny predictions

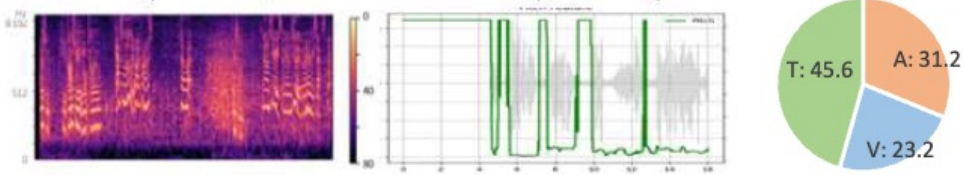
Ross: Is that still…
Rachel: I'm fine. I'm fine.
Ross: No. You are not.

Rachel: Yes, I am.
Ross: Rach
Rachel: Look, I'm fine.



Pete: Can you promise you won't tell he though?
Phoebe: I promise, Tell her what?

Pete: Thanks a lot.
Phoebe: No. I'm intuitive but my memory sucks.



- Positive: high pitch, pause, speech rate change indicate punchline for laughter
- Negative: neutral voice and facial expressions

FunnyNet-W: Text | FunnyNet-W: Visual | FunnyNet-W: Audio

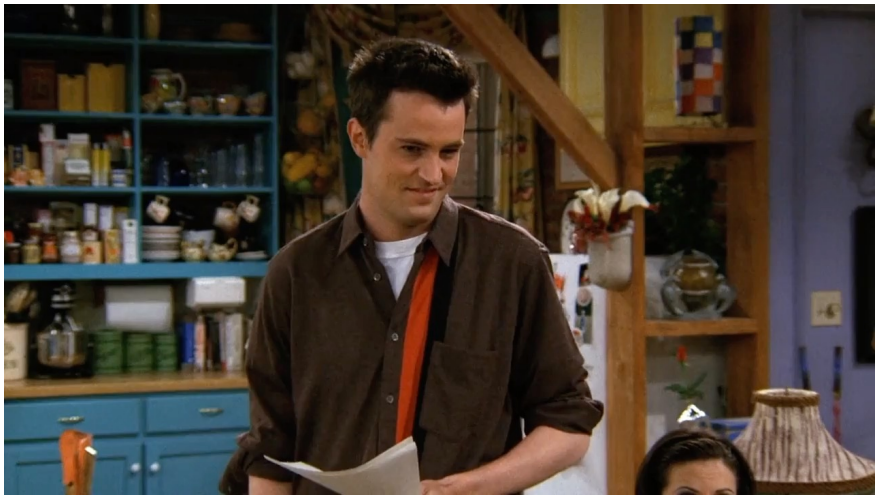
Quantitative results: Comparison to chatbot

Prompt engineering	Prompt training	F1	Accuracy
Generic	-	14.5	41.8
	✓	44.3	46.5
Specific	-	64.1	53.2
	✓	71.1	55.9
FunnyNet-W (T)		77.8	68.1
FunnyNet-W (A+V+T)		88.2	85.6

Models		LlaMa-2		FunnyNet-W
		w/o PT	w PT	
Funny (positive)	They are putting together this panel to talk about fossils they just found in Peru and the Discovery channel is gonna film it. Oh my god, who's gonna watch that?	No	Yes	Yes
	I didn't wear this suit for a year because you hated it. You're not my girlfriend anymore, Now that you're on your own, you're free to look as stupid as you'd like.	No	Yes	Yes
Not funny (Negative)	I hope it won't be too weird. will it? Rache? No, not at all. I'm actually gonna bring someone myself	No	Yes	No
	Let me walk you home and stop by every newsstand and burn every copy of The Times and The Post.	Yes	Yes	No

Failure cases: False Positives

Strong emotional responses expressed by single wording



Chandler: Something else I just said?
Rachel: I don't know. Weren't you the guy who told me to quit my job when I had absolutely nothing else to do?
Ha! Ha! Ha! Ha!

Gunther: Rachel, I just made you cocoa.
Rachel: OMG, you are so nice.
Monica: (screaming) Ah!!
Phoebe: Are you guys OK?

Failure cases: False Negatives

Subtle sarcastic comments with straight face and no follow-up indications or inside jokes that require long-term understanding

Ross: I made a mistake.
Rachel: A mistake? Where were you trying to put it in? Her purse?
Phoebe: Where? Where did he put in?

Joey: You know, they call it “The Ross”.
Joey: People like, huh, he’s got a Ross.
Ross: Yeah, that would be cool.

Sitcom w/ canned laughter

Sitcoms w/o canned laughter

Manny: I wish I could stay home with you and fly toy airplanes.

Jay: These aren't toy airplanes, Manny. These are *models* and they're very complicated. You wanna fly one of these, you gotta be familiar with *air foyle, drag, lift and thrust, and these are all principles of aerodynamics.*

Manny: The box says twelve and up.

Jay: What?!

Conclusion

- FunnyNet-W: self-supervised audio-visual model for funny moment detection
- Exploit only raw video modalities: *audio, visual, textual*
- Audio is the *dominant* cue
- Outperforms the state of the art
- Future work: *other languages, other types of humor*

Movie Question-Answering


Ridouane Ghermin, Vicky Kalogeiton, Ivan Laptev
submission 2024

Time scale

Motion planning

Task planning

Event planning

e.g. vacation planning

e.g. planning education; building a house…

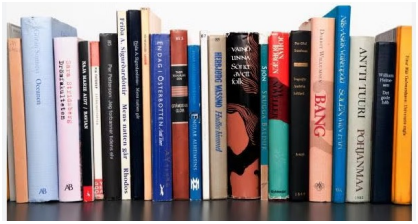
seconds | minutes | hours | days | years | time

Time scale

Text resources

Cook boks, travel guides, blogs,

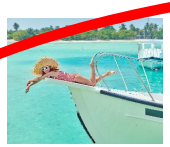
documentaries, history books



seconds

minutes

hours



days

years

time

Image/Video resources

Instructional videos,

SFD: Short Film Dataset Ghermi et al., ArXiv 2024

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Comedy

The Beast, the Phantom and the Hunchback use a dating app to find love.

Romance

A couple on a first date clash over astrology.

Horror

*A young woman trapped in a bathroom stall during an active sho*ting.*

Action

The biggest boxing fight of 1960 takes an unexpected turn.

Drama

A mother struggles with bullies who torment her disabled daughter.

Animation

A young boy gets lost in a strange forest. Then his father tries to rescue him.

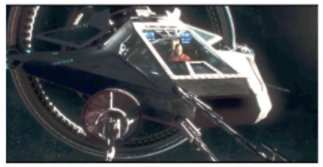
Documentary

How Movie Sounds are Made. An Inside Look at the World of Foley Artist.

Experimental

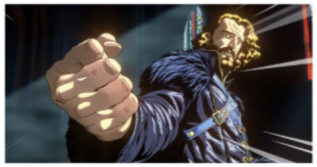
A stranded soul searches an endless desert for his purpose.

Sci-Fi



A lone astronaut testing the first faster-than-light spacecraft travels farther than he imagined

AI movie



After their father dies, two brothers turn to a traditional battle method to decide who will rule the kingdom.

Dance

Testament to timidity and enthusiasm; the dancers shed their hardened pandemic-built exteriors

Western

A young woman in the Old West sets off on a journey of revenge after her sister's murder

SFD is a VideoQA dataset, containing 1,078 movies and 4,885 questions. Videos last 13 minutes on average.

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

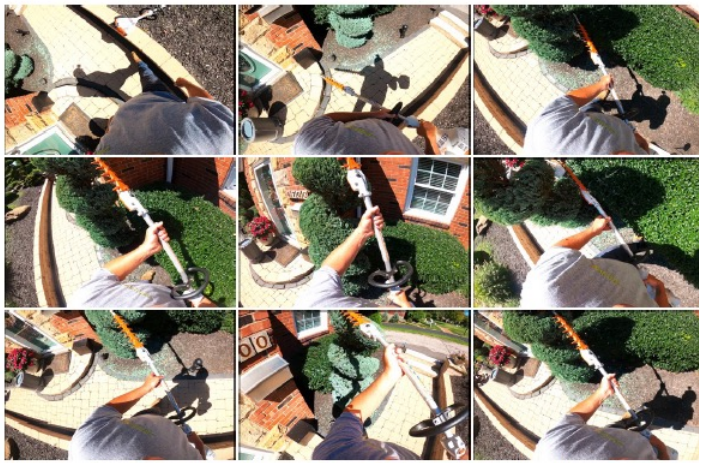
Movies

A young girl lives in a bunker, protected from the monsters. Today she goes outside...

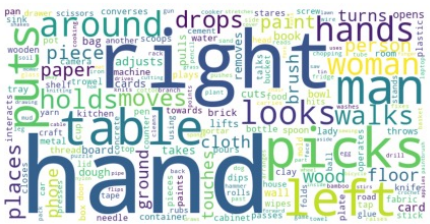
Instructional videos

How to tie a necktie?

Egocentric videos



C was in the front yard, pulled the starter string and trimmed the tree with a hedge trimmer



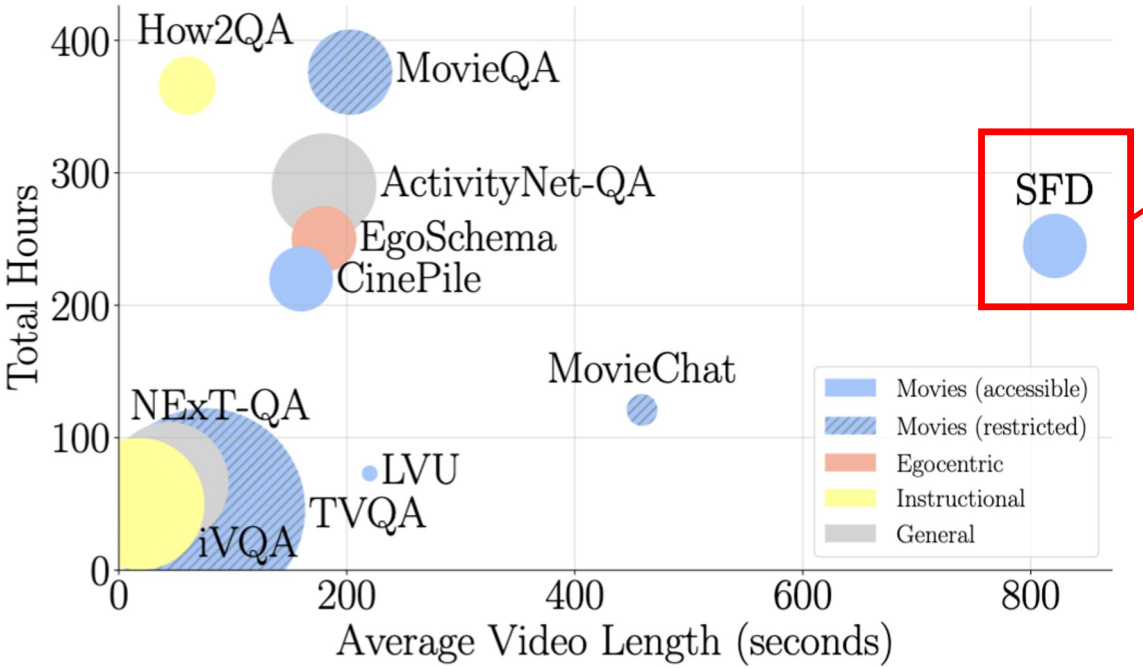
Vicky Kalogeiton

25/10/2024 Multimodal story-level genAI

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/



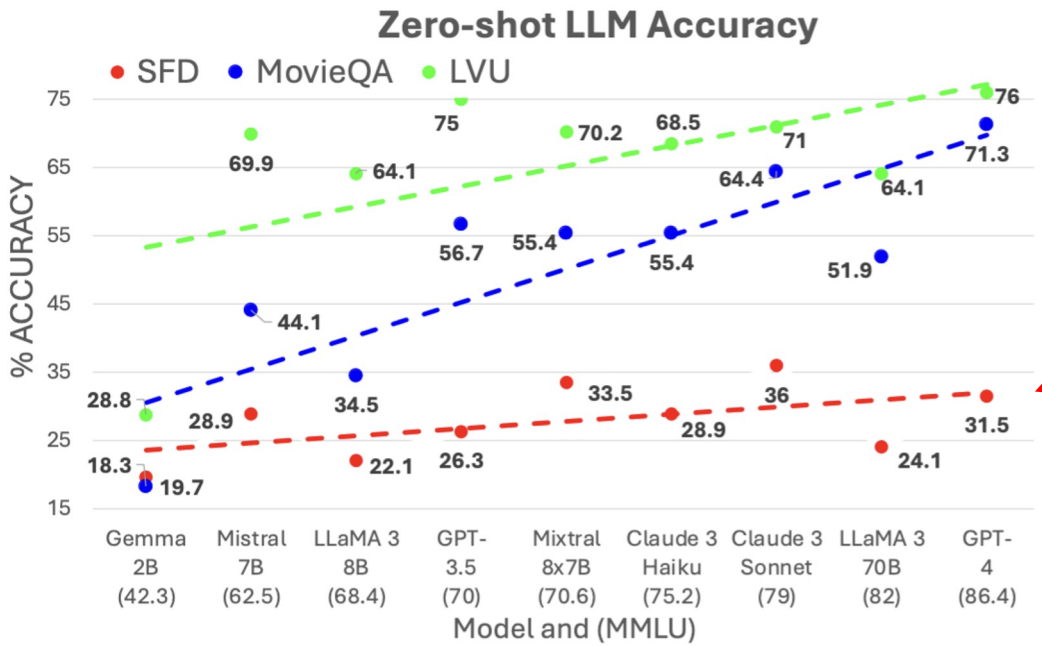
Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/



Modern LLMs memorize common movies and can answer Questions in in LVU and MovieQA given movie names

Why another VideoQA dataset?

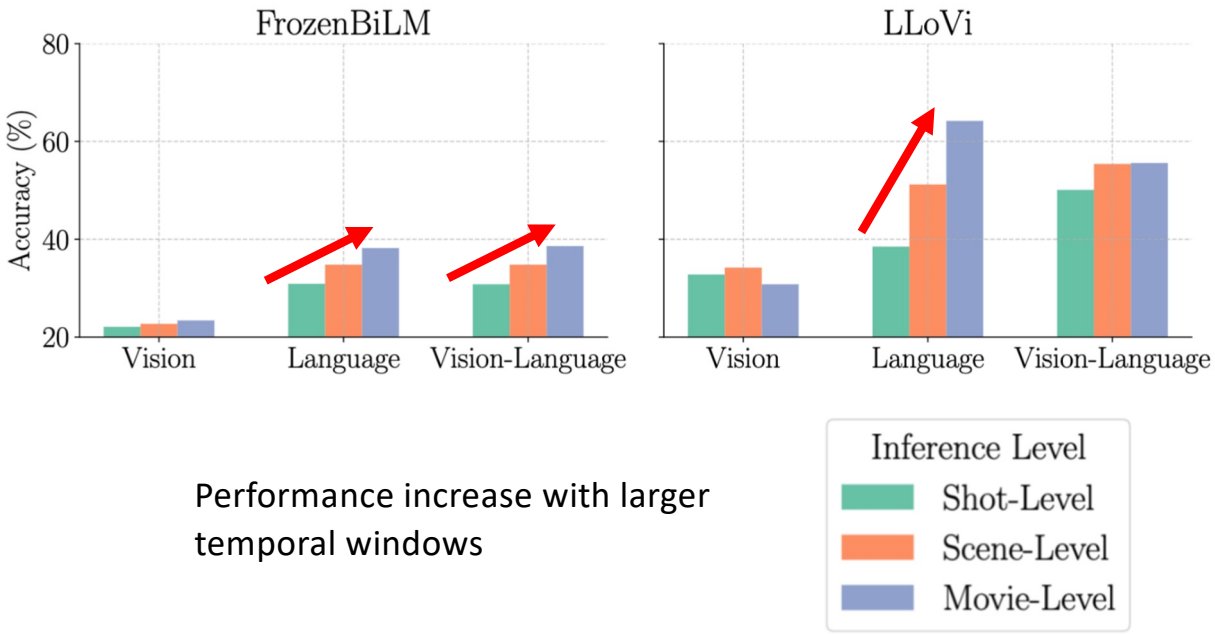
- Story-level QAs
- Publicly available videos
- Limited/No data leakage

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Why another VideoQA dataset?



Performance increase with larger temporal windows

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	30.8	64.2	55.6	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		59.0	70.9	89.8	-	-	-

V: Only visual input

L: Only text input (speech transcripts)

VL: Visual+text input

Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	30.8	64.2	55.6	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		59.0	70.9	89.8	-	-	-

V: Only visual input

L: Only text input (speech transcripts)

VL: Visual+text input

Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

Method	Venue	% Accuracy Multiple-Choice QA V	L	VL	Open-Ended QA V	L	VL
Random		20.0	20.0	20.0	-	-	-
FrozenBiLM [5]	NeurIPS 2021	23.4	38.2	38.6	-	-	-
mPLUG-Owl2 [74]	CVPR 2024	38.3	20.7	21.3	22.1	1.8	1.6
Video-LLaVA [33]	arXiv 2023	34.2	21.3	24.7	19.2	6.4	8.0
LLoVi [79]	arXiv 2023	30.8	64.2	55.6	16.2	40.3	24.7
LangRepo [26]	arXiv 2024	29.0	32.1	31.0	3.5	10.4	9.5
MovieChat [54]	CVPR 2024	8.4	6.4	8.0	14.0	15.7	11.8
TimeChat [47]	CVPR 2024	25.5	6.4	31.8	26.4	9.4	5.9
Human		59.0	70.9	89.8	-	-	-

V: Only visual input

L: Only text input (speech transcripts)

VL: Visual+text input

Why another VideoQA dataset?

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

SHORT FILM DATASET (SFD)

A Benchmark for Story-Level Video Understanding

https://shortfilmdataset.github.io/

SONGKRAN

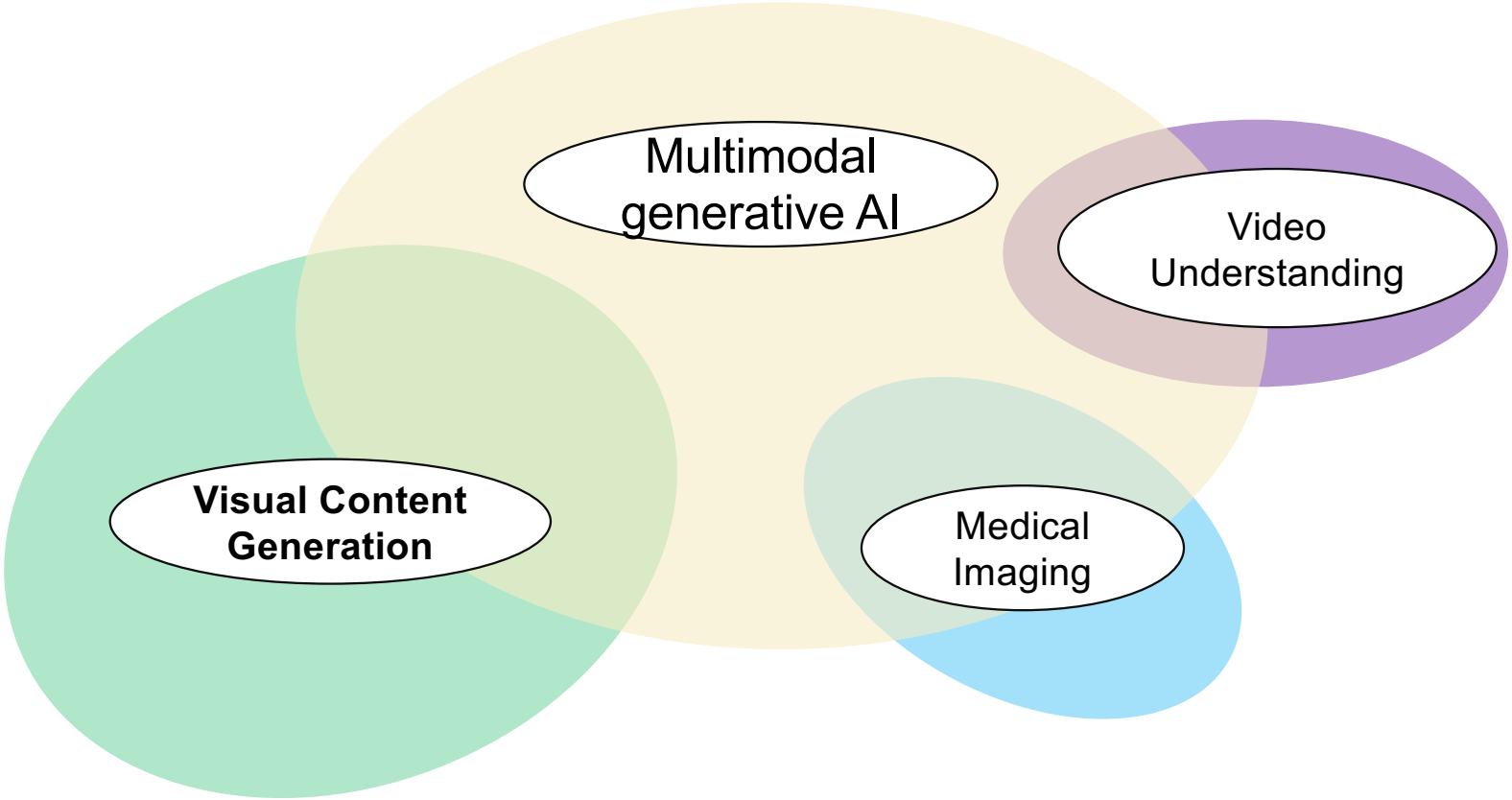
A coffee machine salesman falls for a boutique cafe owner on a business trip to Thailand.

What problem does Pete encounter on his way to the hotel?

A) He loses his passport and must navigate Bangkok's bureaucracy to get a temporary one.
B) He is pickpocketed in a crowded market and loses his money and phone.
C) He gets stuck in Bangkok's traffic and decides to walk, getting lost in the process. ✔
D) He mistakenly takes the wrong bus and ends up in a distant part of the city.
E) He finds that his hotel reservation has been mistakenly cancelled.

- Story-level QAs
- Publicly available videos
- Limited/No data leakage
- Questions with long temporal context
- Finding #1: Transcript-only performance of best LLMs is approaching human
- Finding #2: Vision-only performance of best VLMs is **18% below human**

Research agenda

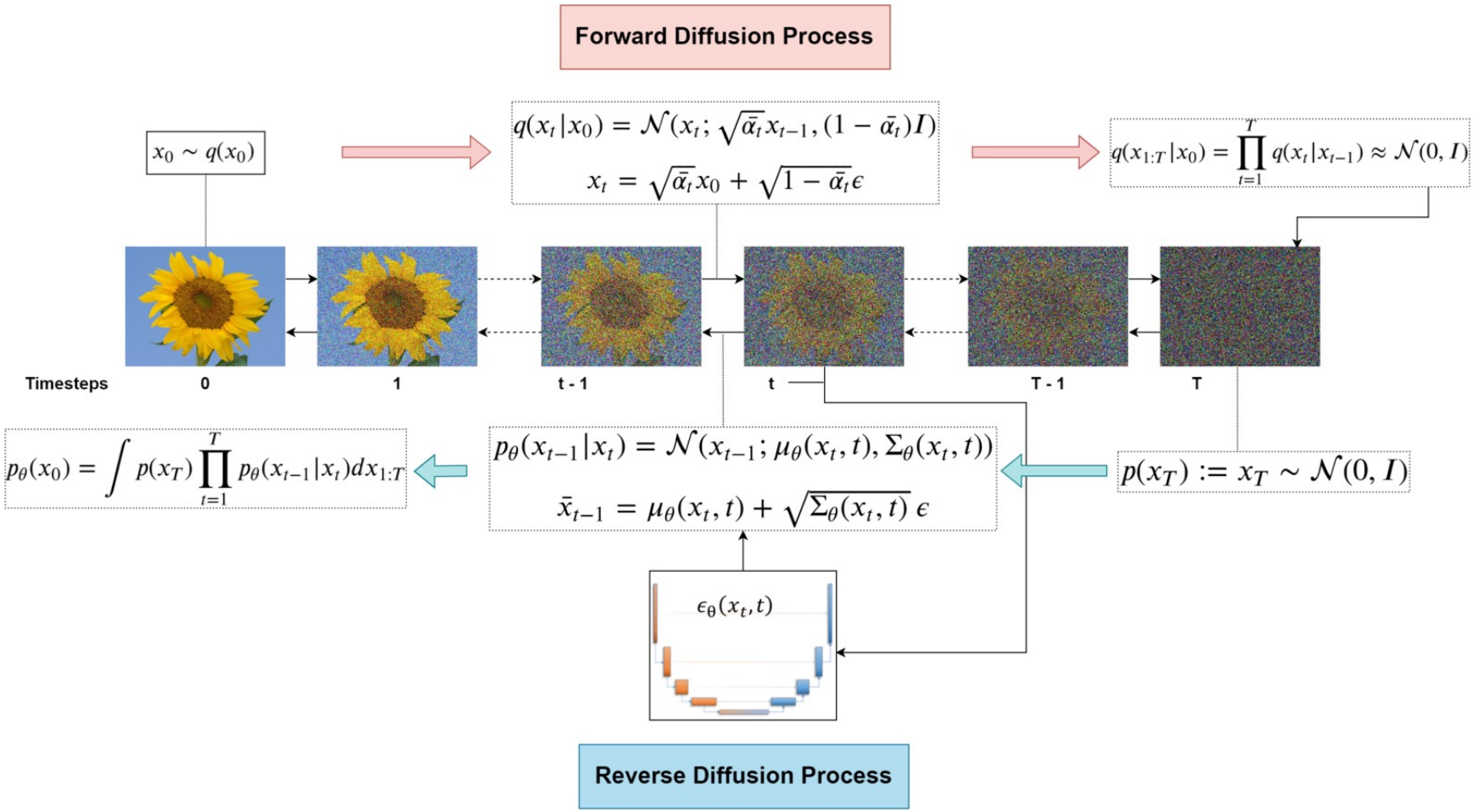


Part II: Visual Content Generation

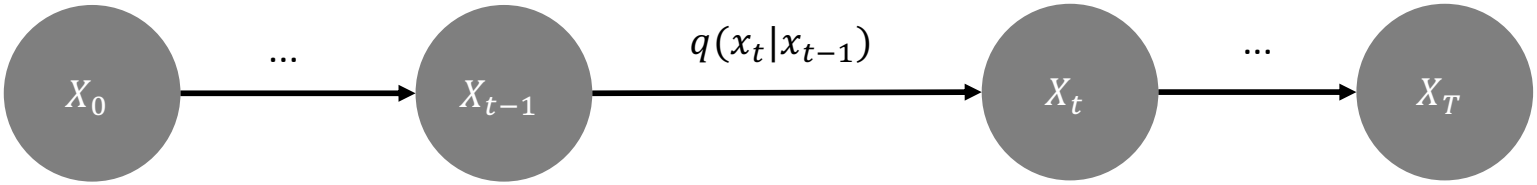
- Diffusion recap
- Guidance
  - Guided diffusion
  - Control the diffusion
  - Explicit condition
  - Guided diffusion
  - Why not guided diffusion?
  - Classifier-free guidance
- Analysis of Classifier-Free Guidance Weight Schedulers
- ControlNet

Denoising Diffusion Probabilistic Models

Denoising Diffusion Probabilistic Model (DDPM)



DDPM: Forward Process

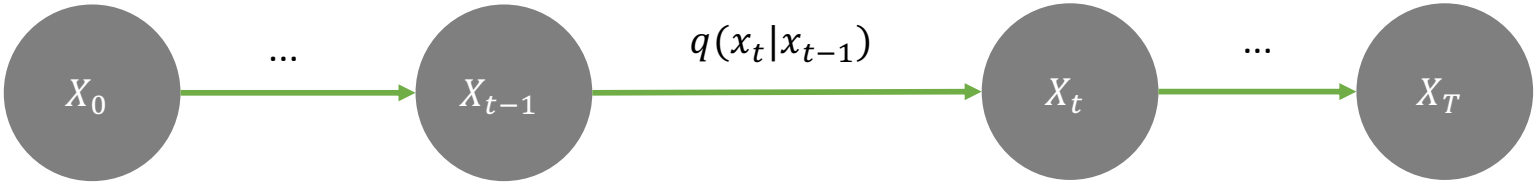


- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule

Forward:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

DDPM: Forward Process



- Original image at X_0 and pure noise at X_T
- We repeat the noising T times
- $\beta_t \in (0,1)$ is a noise schedule

Forward:

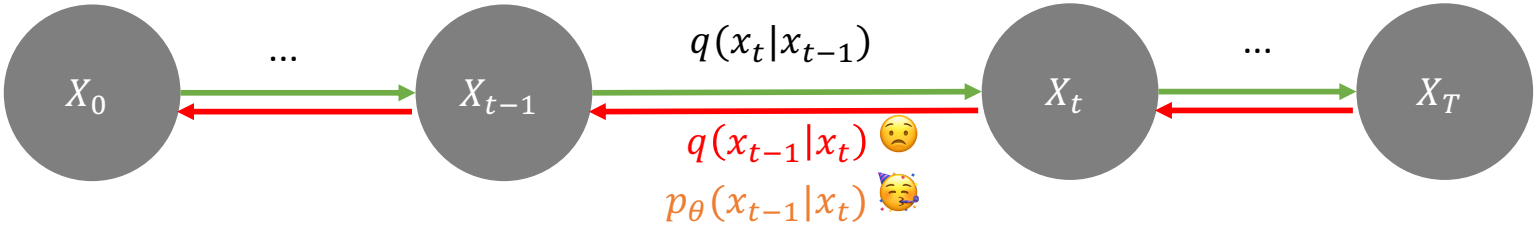
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

("Shortcut")
Sample any step using x_0 :

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$$

$$\alpha_t = 1 - \beta_t \qquad \bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$$

DDPM: Reverse (=Generative) Process



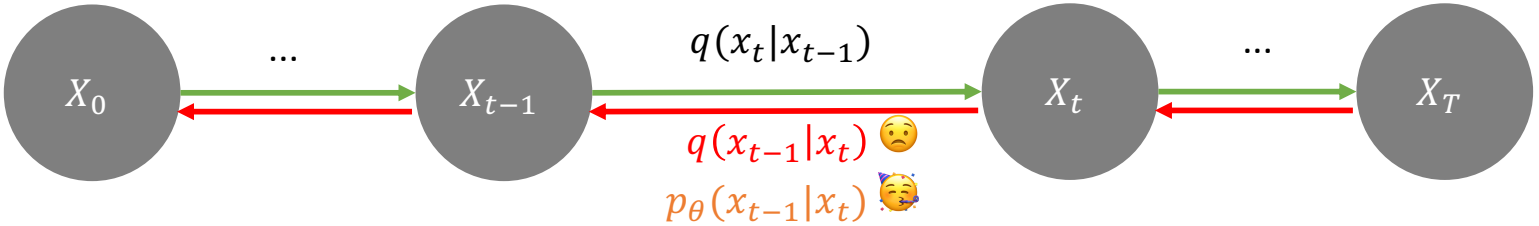
A very nice property of Gaussian:
if $q(x_t|x_{t-1})$ is a Gaussian with small β (another reason we need many steps!)
→ then, $q(x_{t-1}|x_t)$ is also a Gaussian.
Therefore, we learn this Gaussian's mean and variance
by a network approximated $p_\theta(x_{t-1}|x_t)$

Generation:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Learnable parameters

DDPM: Generative Process



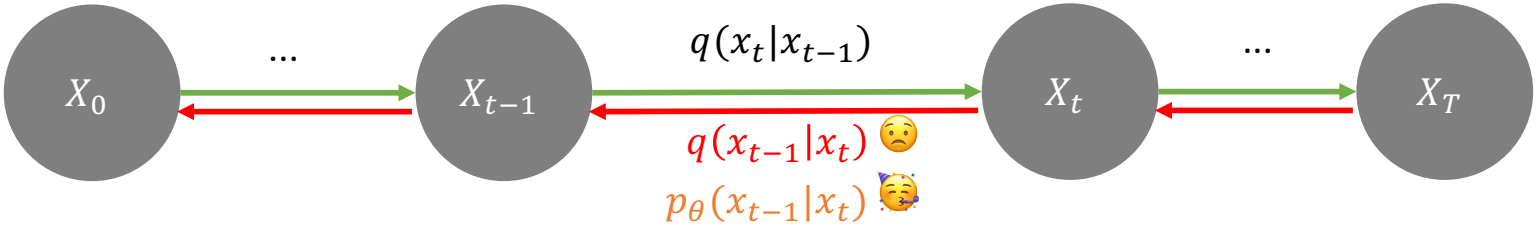
Forward

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Generation

DDPM: Reverse Process

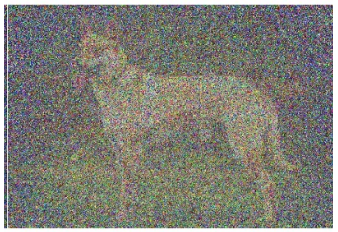


Forward

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

Bayes

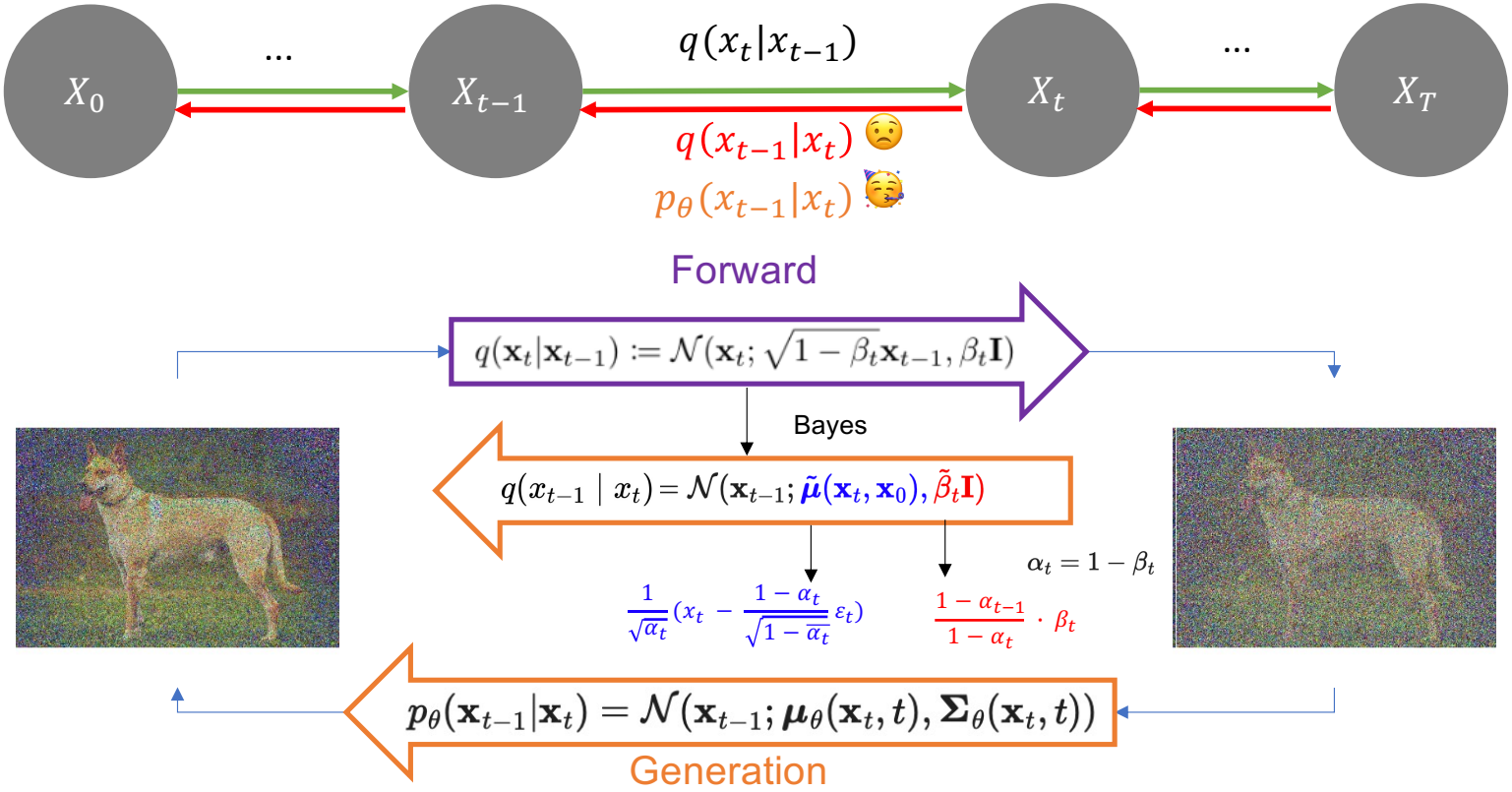
$$q(x_{t-1} \mid x_t) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I})$$



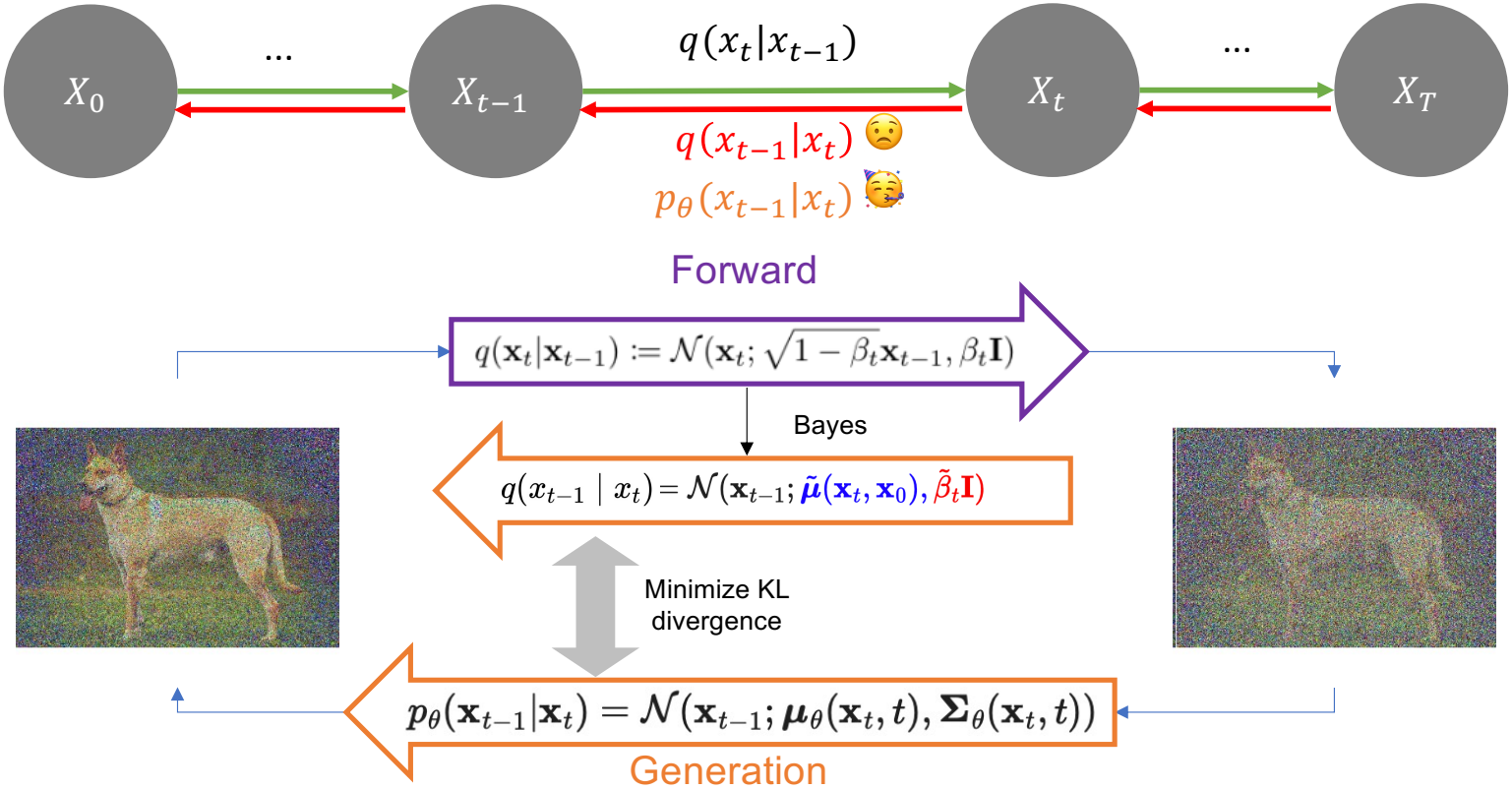
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Generation

DDPM: Reverse/Generative Process



DDPM: Reverse/Generative Process



Loss

DDPM: Loss

- KL divergence between two Gaussians:
  - Variance is fixed: equivalent to minimizing the distance between their means

- Training objective at step t becomes:

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right]$$

DDPM: Loss

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \| \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t) \|^2 \right]$$

- Recall that we used our "shortcut" property, and the mean is expressed by x_t and ε_t:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}} \varepsilon_t \right) \qquad \text{, where } \varepsilon_t \sim N(0,1)$$

DDPM: Loss

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2 \right]$$

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1 - \alpha_t}{\sqrt{1 - \overline{\alpha_t}}} \varepsilon_\theta)$$

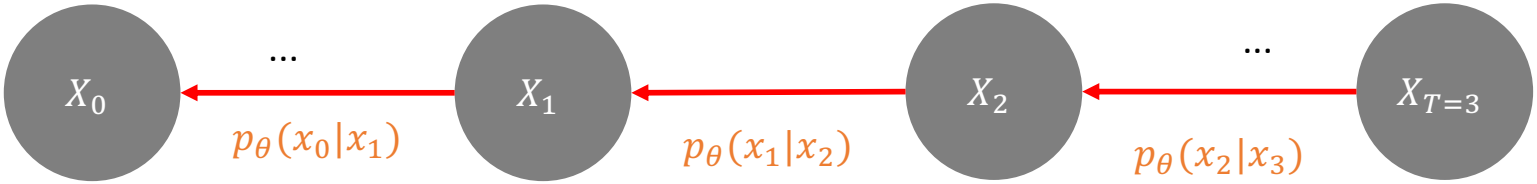
- New learning objective:

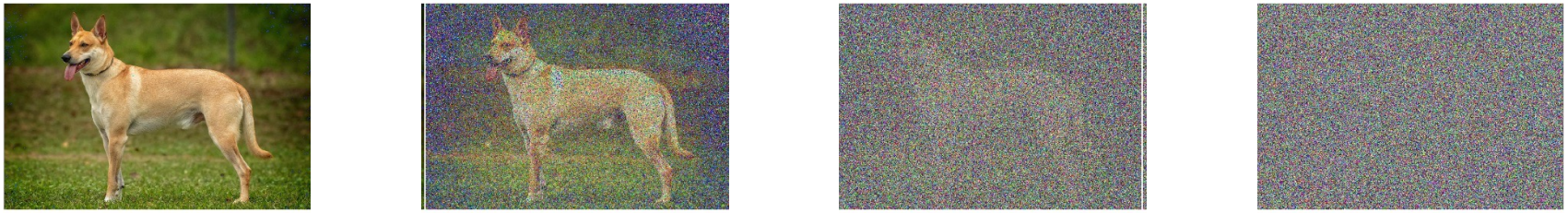
$$\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2$$

Training data

Network to predict the noise at each step

DDPM: Sampling

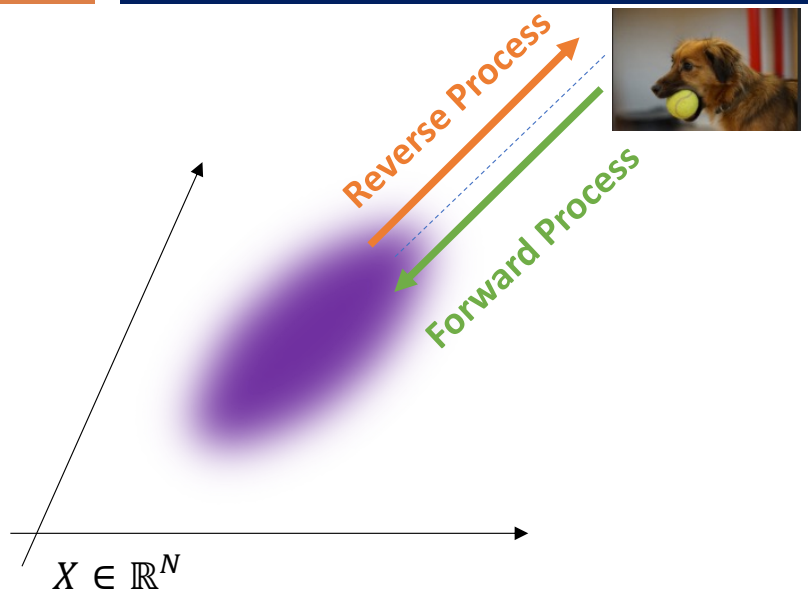




$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1};\ \mu_\theta(x_t,t), \boxed{\mathbf{\Sigma}_\theta(\mathbf{x}_t,t)})$$

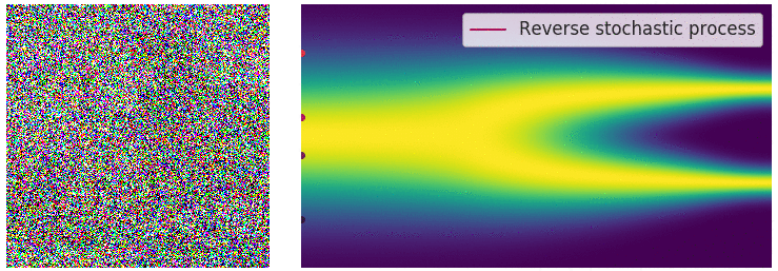
$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t,t)\right) + \boxed{\sigma_t \mathrm{z}}$$

Control the Diffusion Model



Distribution of Learnt Data $P_\theta(X)$
with parameter $\theta \in \mathbb{R}^M$

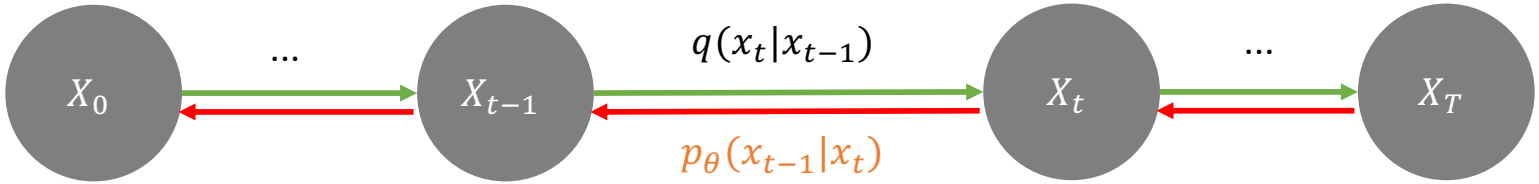
Good, it means one noise gives me an image!



But how can I achieve **control** on this? For example, I want a cat image, rather than others.

Or even more complicated: "*A stained glass window of a panda eating bamboo*." – text-to-image generation

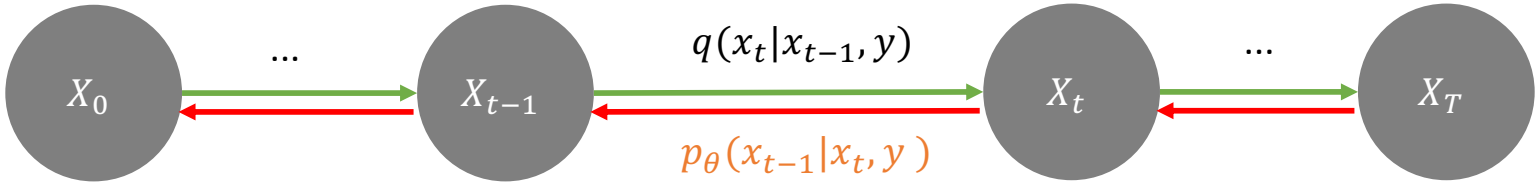
Control the Diffusion Model



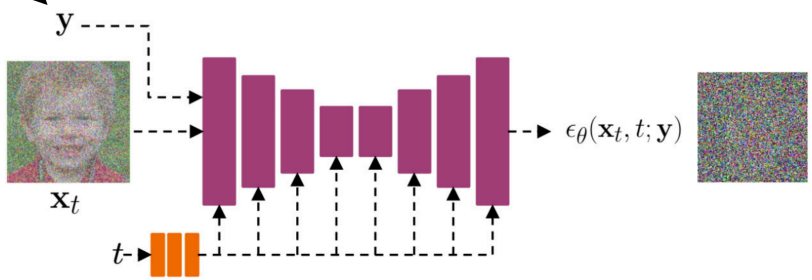
Where is the control?
How did we do with VAE? This sounds a familiar question.

Explicit condition

Control the Diffusion Model: Explicit Condition

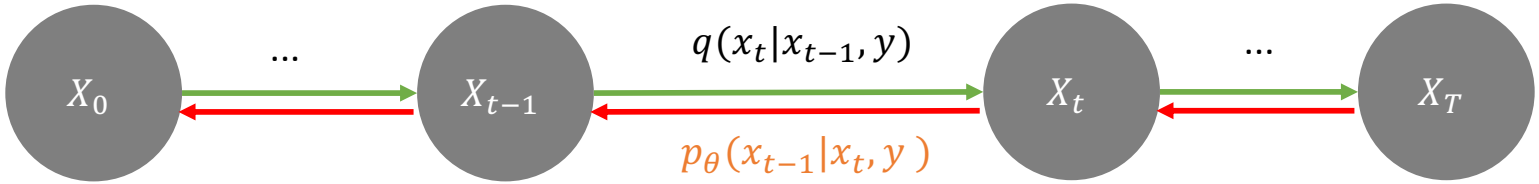


"a young schoolboy in a red shirt"

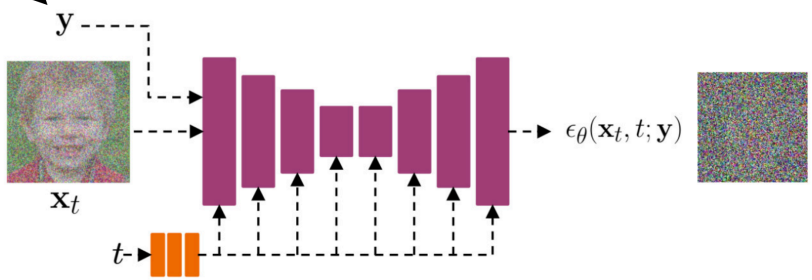


We can add it directly.

Control the Diffusion Model: Explicit Condition



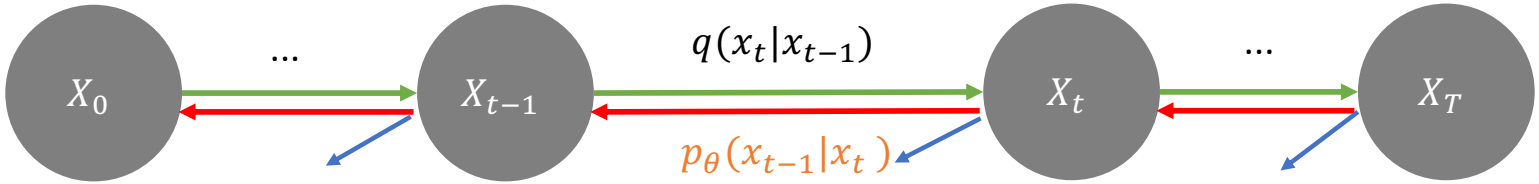
"a young schoolboy in a red shirt"



We can add it directly, but is this an effective way? Why?

Guided diffusion

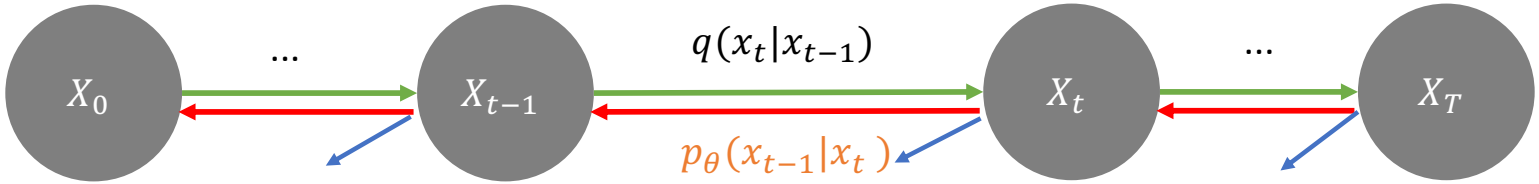
Control the Diffusion Model: Guided Diffusion



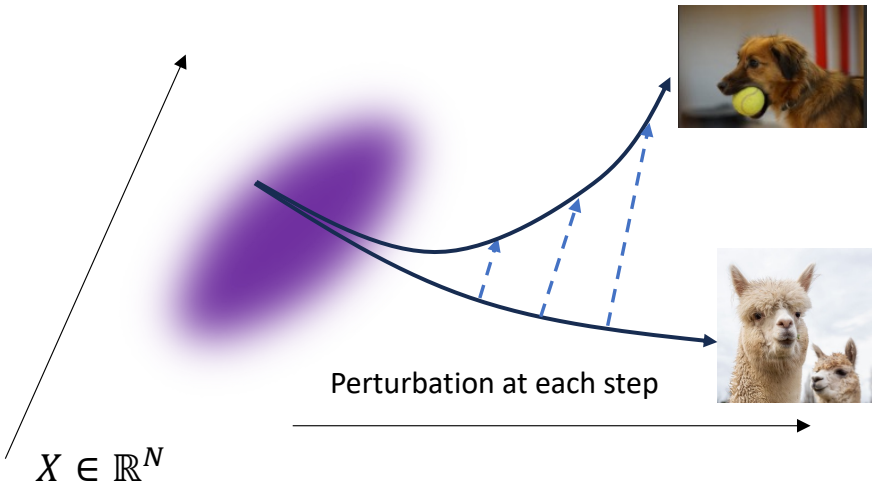
Shift to a dog label!

Let's perturb it step-by-step during the generation!

Control the Diffusion Model: Guided Diffusion

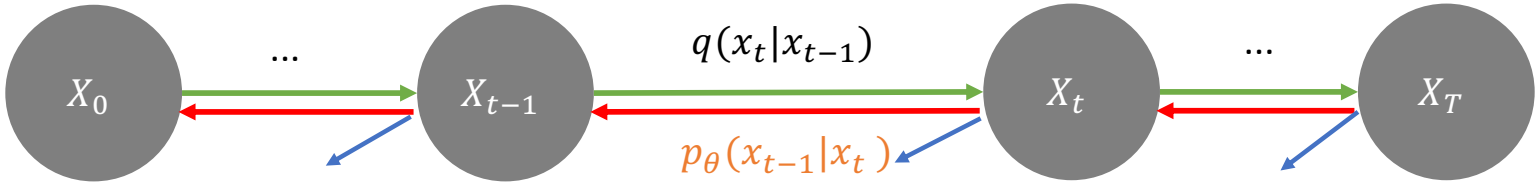


Shift to a dog label!

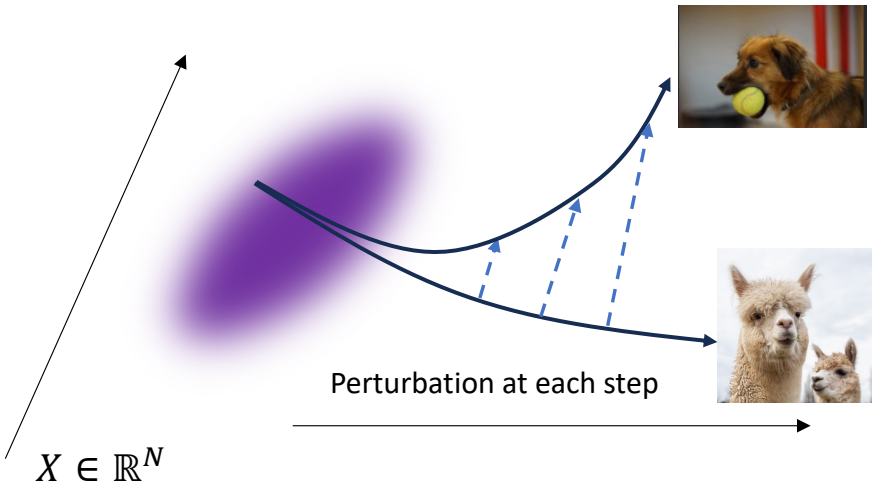


How to do it?

Control the Diffusion Model: Guided Diffusion



Shift to a dog label!



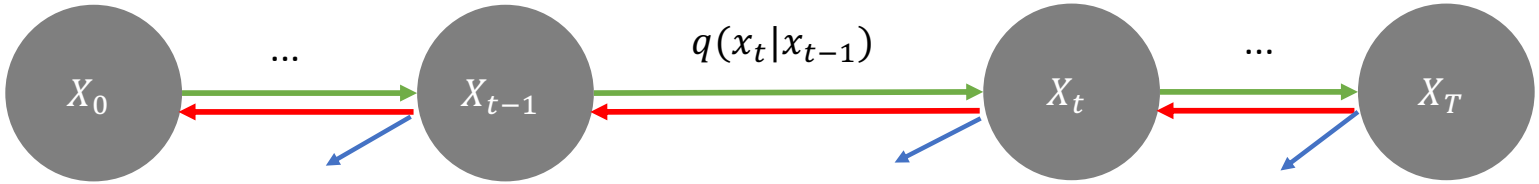
It's like a classifier: $p_\phi(y|x)$

We just look at its gradient:

$\nabla_x \log p_\phi(y|x)$

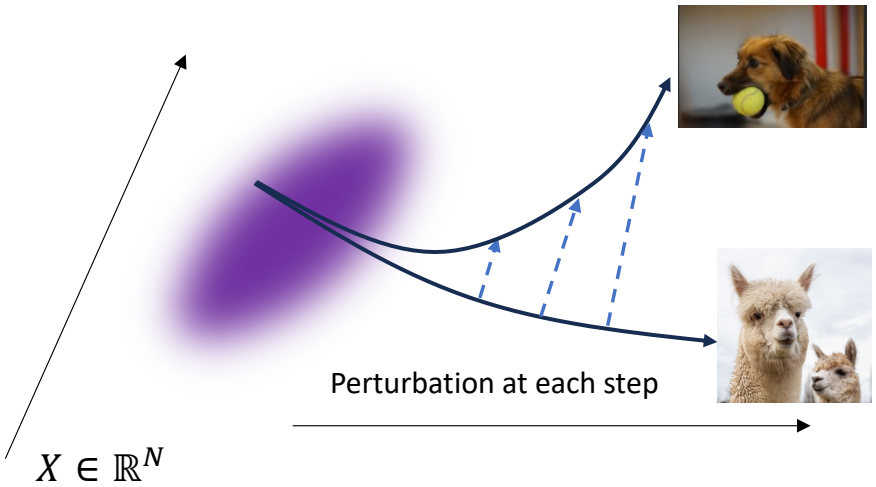
Such that the generated x is similar to the condition label.

Control the Diffusion Model: Guided Diffusion



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!



It's like a classifier: $p_\phi(y|x)$

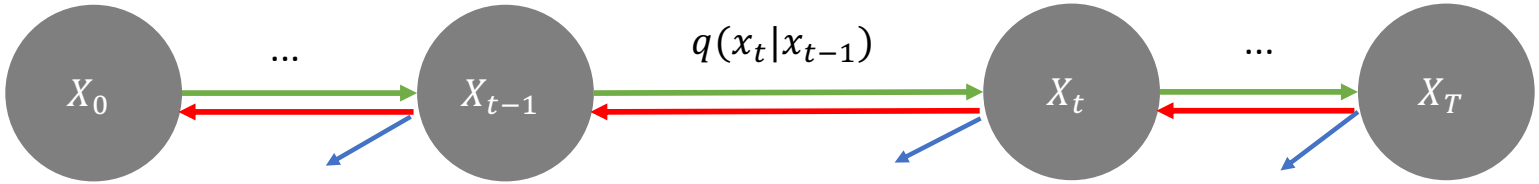
We just look at its gradient:

$$\nabla_x \log p_\phi(y|x)$$

Such that the generated x is similar to the condition label.

In sampling: $\epsilon_\theta(x_t, t) + \nabla_x \log p(y|x)$

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

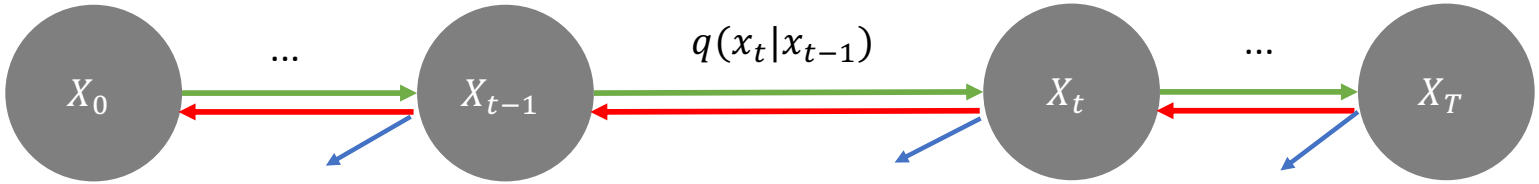
Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \nabla_x \log p_\phi(y|x)$

$p_\phi(y|x)$

A Dog

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

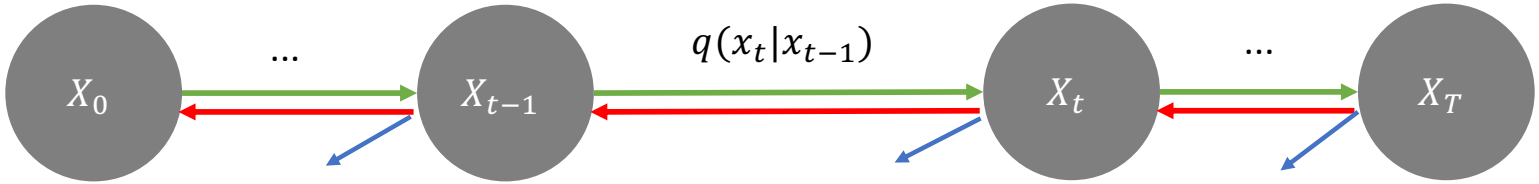
In sampling: $\epsilon_\theta(x_t, t) + \nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**

We need to train a classifier: $p_\phi(y|x)$, with the awareness of noise

$p_\phi(y|x)$

A Dog

Control the Diffusion Model: Guided Diffusion

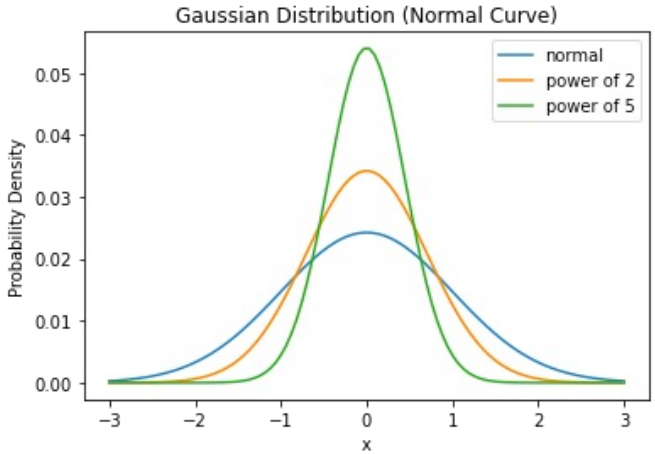


$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

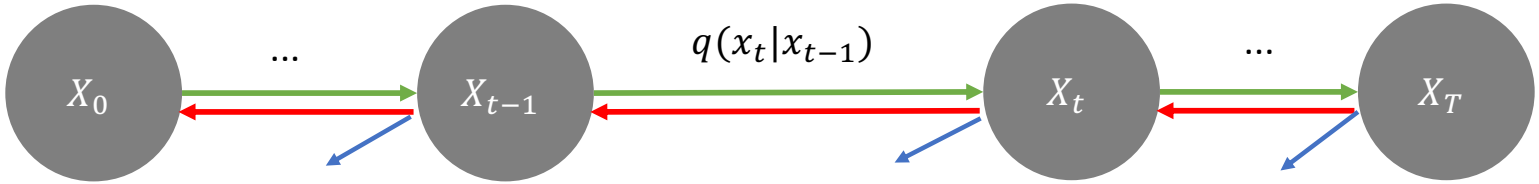
In sampling: $\epsilon_\theta(x_t, t) + \gamma\nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**

$\gamma\nabla_x \log p_\phi(y|x) \sim \nabla_x \log p_\phi(y|x)^\gamma$



Dhariwal and Nichol, 2021

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma\nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**

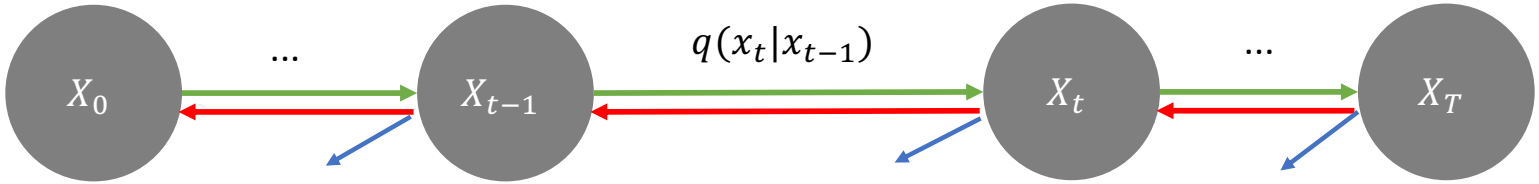
Label: Corgi

γ = 1

γ = 3

Dhariwal and Nichol, 2021

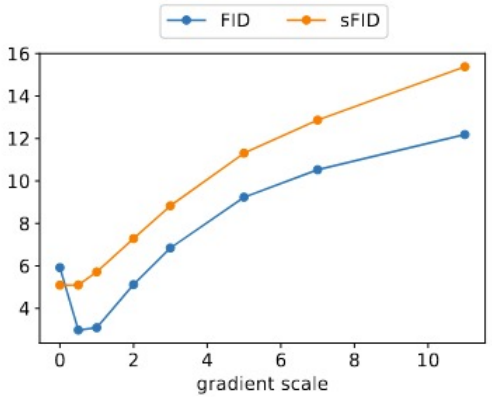
Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma \nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**



Dhariwal and Nichol, 2021

Guided Diffusion: Nearest Neighbors for Samples

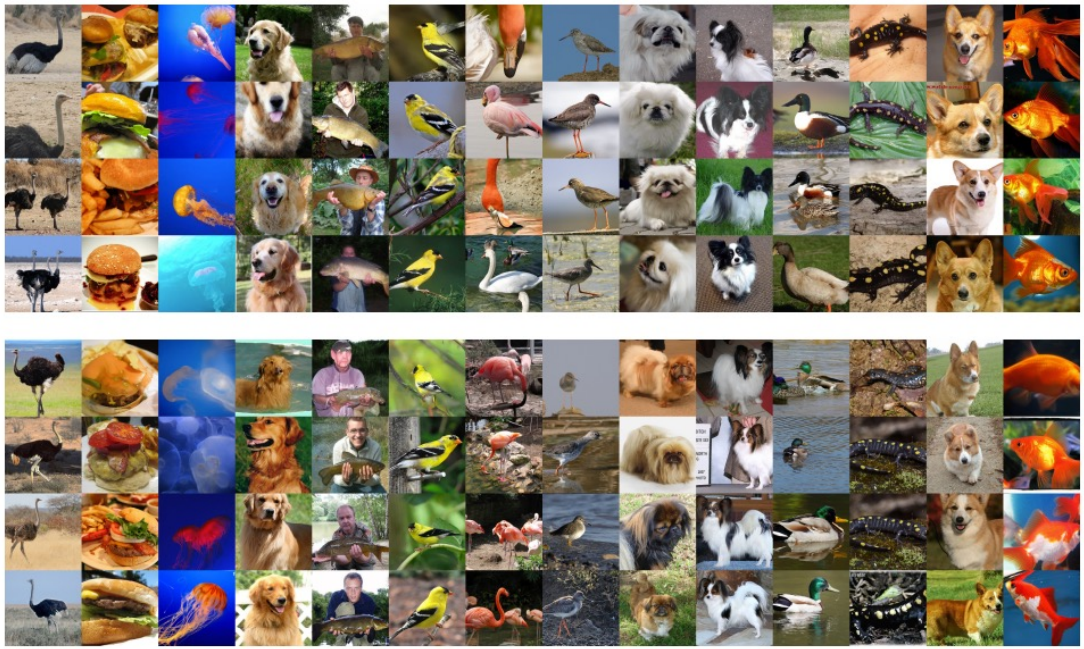


Figure 7: Nearest neighbors for samples from a classifier guided model on ImageNet 256×256. For each image, the top row is a sample, and the remaining rows are the top 3 nearest neighbors from the dataset. The top samples were generated with classifier scale 1 and 250 diffusion sampling steps (FID 4.59). The bottom samples were generated with classifier scale 2.5 and 25 DDIM steps (FID 5.44).

Guided Diffusion: Effect of Varying the Classifier Scale

Figure 8: Samples when increasing the classifier scale from 0.0 (left) to 5.5 (right). Each row corresponds to a fixed noise seed. We observe that the classifier drastically changes some images, while leaving others relatively unaffected.

Guided Diffusion: Examples

Figure 13: Samples from our best 512×512 model (FID: 3.85). Classes are 1: goldfish, 279: arctic fox, 323: monarch butterfly, 386: african elephant, 130: flamingo, 852: tennis ball.

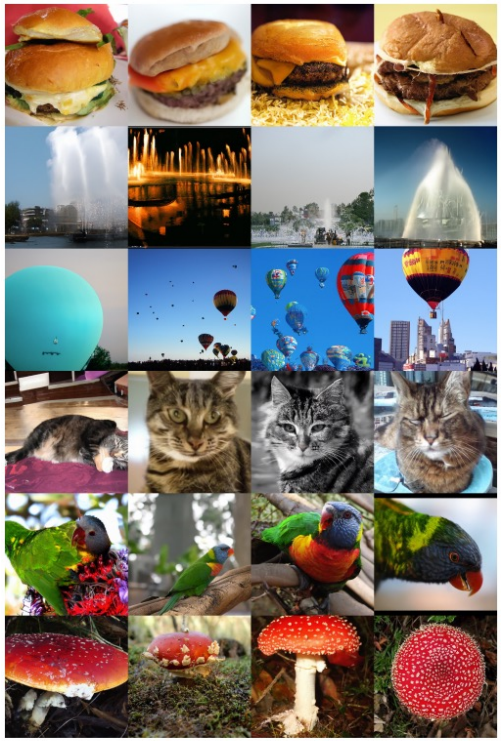
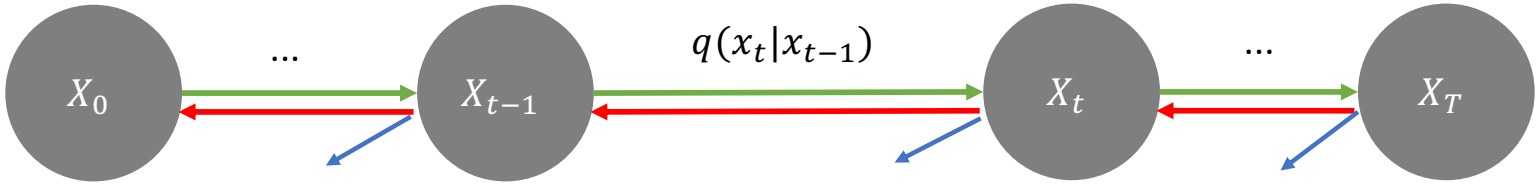


Figure 14: Samples from our best 512×512 model (FID: 3.85). Classes are 933: cheeseburger, 562: fountain, 417: balloon, 281: tabby cat, 90: lorikeet, 992: agaric.

Why not guided diffusion?

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

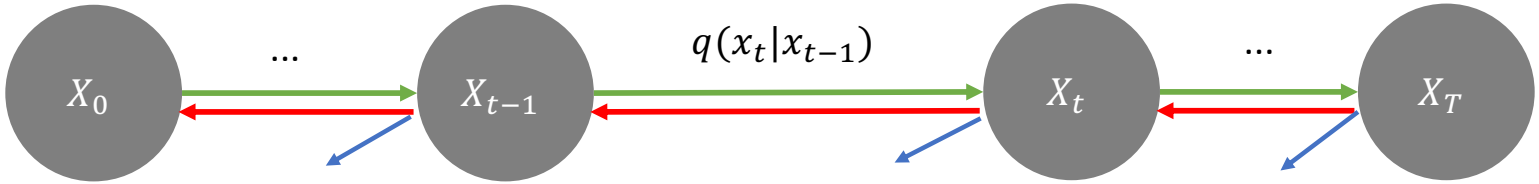
Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma\nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**

What do we NOT like in guided diffusion?

Dhariwal and Nichol, 2021

Control the Diffusion Model: Guided Diffusion



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

In sampling: $\epsilon_\theta(x_t, t) + \gamma\nabla_x \log p_\phi(y|x)$ ← **Guided Diffusion**

What do we NOT like in guided diffusion?

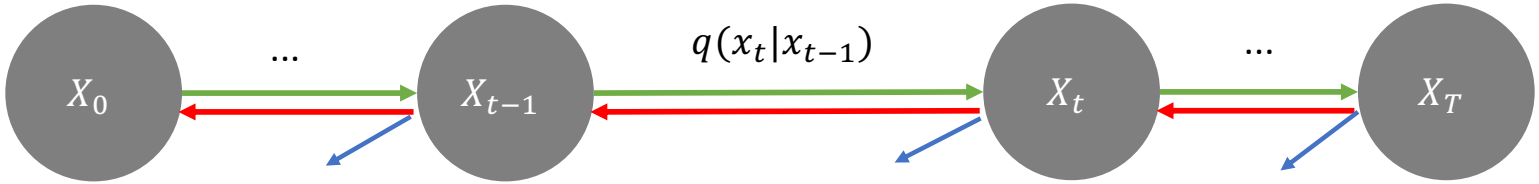
- Need to fine-tune and train a classifier
- Condition can only be label-based, hard to support other conditions like “text input”

Because for text, the classifier $p_\phi(y|x)$ ***does not*** exist.

Dhariwal and Nichol, 2021

Classifier-free guidance

Control the Diffusion Model: Classifier-Free Guidance



$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

Shift to a dog label!

At training: $p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

In sampling: $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma \nabla_x \log p(y|x)$

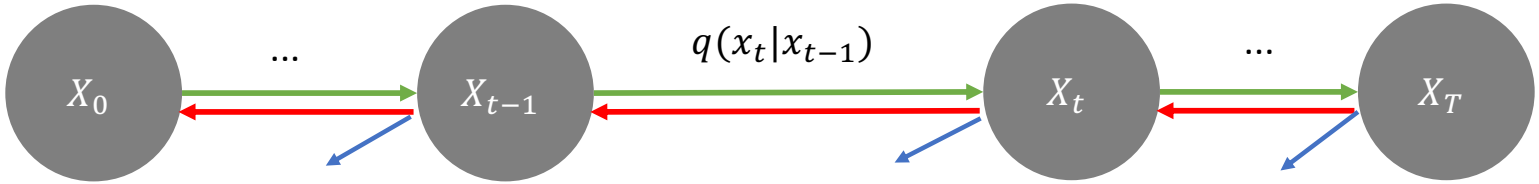
$$p(y|x) \propto \frac{p(x|y)}{p(x)}$$

$$\nabla_x \log p(y|x) \propto \nabla_x \log p(x|y) - \nabla_x \log p(x)$$

Thanks to Bayes

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

At training: $p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$

In sampling: $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma \nabla_x \log p(y|x)$

$$\nabla_x \log p(y|x) \propto \epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t)$$

$$p(y|x) \propto \frac{p(x|y)}{p(x)}$$

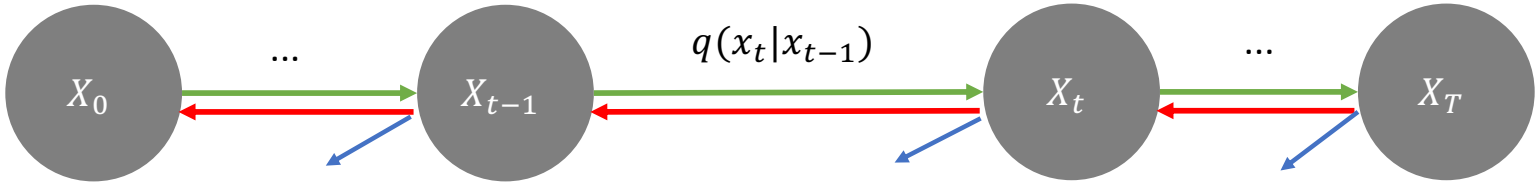
$$\nabla_x \log p(y|x) \propto \nabla_x \log p(x|y) - \nabla_x \log p(x)$$

Thanks to Bayes

Finally: $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\underbrace{\epsilon_\theta(x_t, t, y)}_{\text{conditional generation}} - \underbrace{\epsilon_\theta(x_t, t)}_{\text{unconditional generation}})$

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

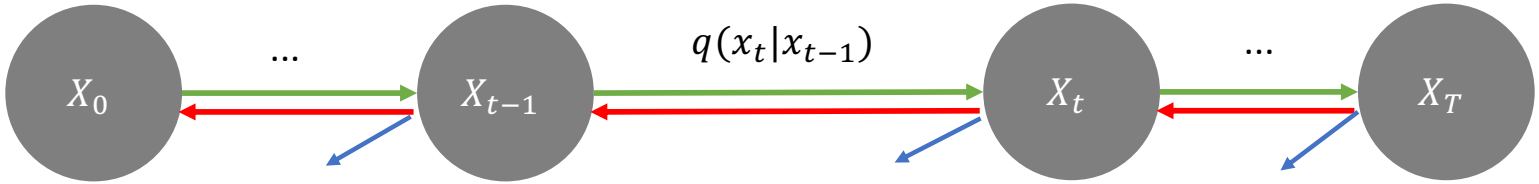
$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\underbrace{\epsilon_\theta(x_t, t, y)}_{\text{conditional generation}} - \underbrace{\epsilon_\theta(x_t, t)}_{\text{unconditional generation}})$$

How to compute: $\epsilon_\theta(x_t, t, y)$ →
- explicit condition

How to compute: $\epsilon_\theta(x_t, t)$ →

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)\, p_\phi(y|x)$$

Shift to a dog label!

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\underbrace{\epsilon_\theta(x_t, t, y)}_{\text{conditional generation}} - \underbrace{\epsilon_\theta(x_t, t)}_{\text{unconditional generation}})$$

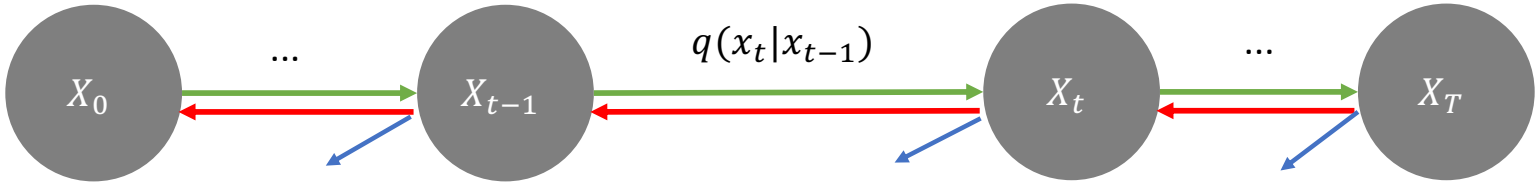
Implicit classifier

How to compute: $\epsilon_\theta(x_t, t, y)$ →
- explicit condition

How to compute: $\epsilon_\theta(x_t, t)$ →
- We randomly set the condition to null (drop-out condition)
- $\epsilon_\theta(x_t, t, y)$ → $\epsilon_\theta(x_t, t, \emptyset)$

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)p_\phi(y|x)$$

Shift to a dog label!

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

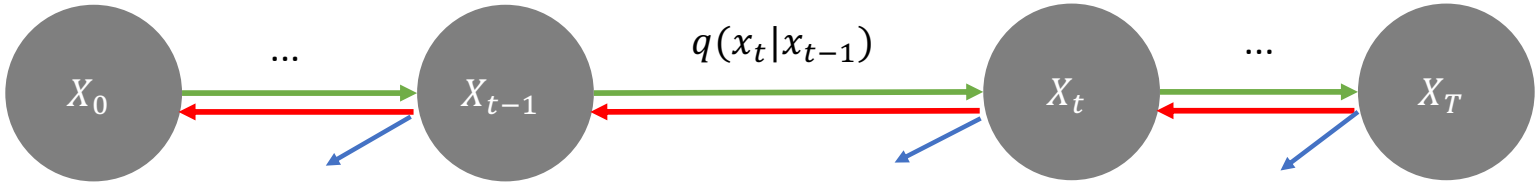
Label: Husky

γ=1

γ=3

Ho and Salimans, 2022

Control the Diffusion Model: Classifier-Free Guidance



$$p_\theta(x_{t-1}|x_t, y) = p_\theta(x_{t-1}|x_t)\, p_\phi(y|x)$$

Shift to a dog label!

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \gamma(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

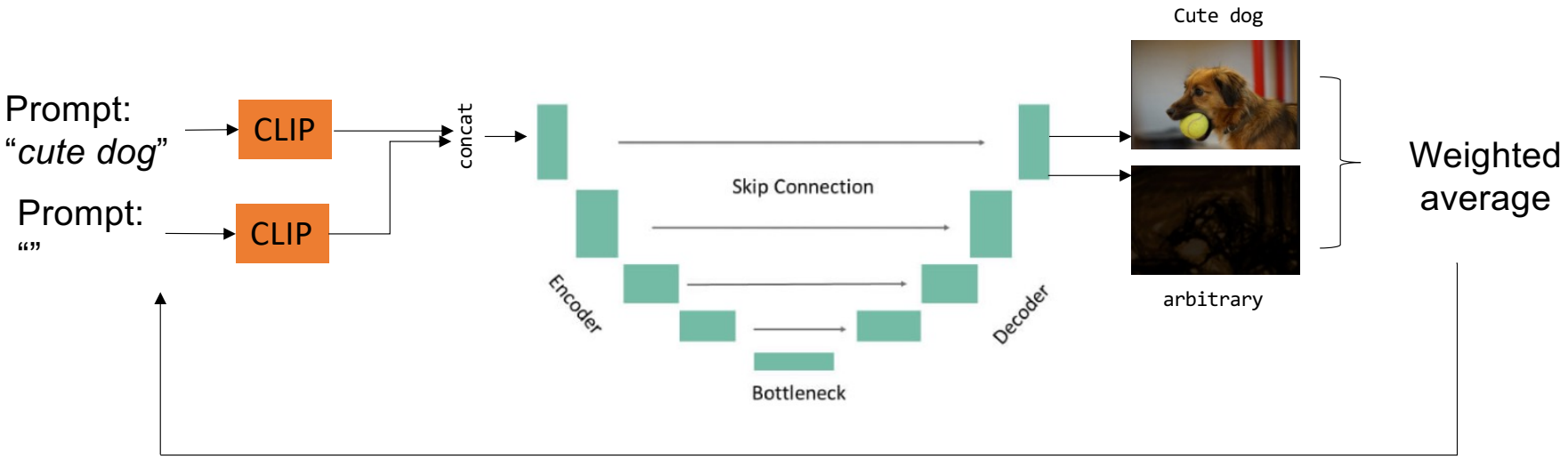
We do not need the explicit classifier: we can use text-encoder to condition on text.
Called : Classifier-Free Guidance (CFG)

"A panda is eating ice-cream" → CLIP/T5 → 768x1 dim

Condition Latent

Ho and Salimans, 2022

Classifier free guidance



Classifier free guidance

Control the Diffusion Model: Classifier-Free Guidance

γ = 1

γ = 3

Caption: "*A stained glass window of a panda eating bamboo.*"

Control the Diffusion Model: Classifier-Free Guidance

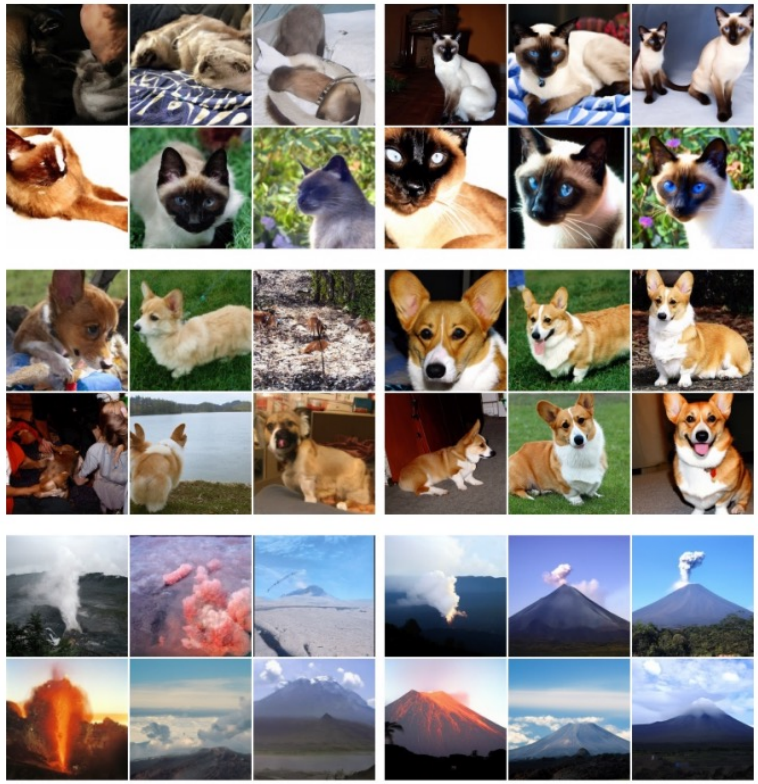


Figure 3: Classifier-free guidance on 128x128 ImageNet. Left: non-guided samples, right: classifier-free guided samples with $w = 3.0$. Interestingly, strongly guided samples such as these display saturated colors. See Fig. 8 for more.

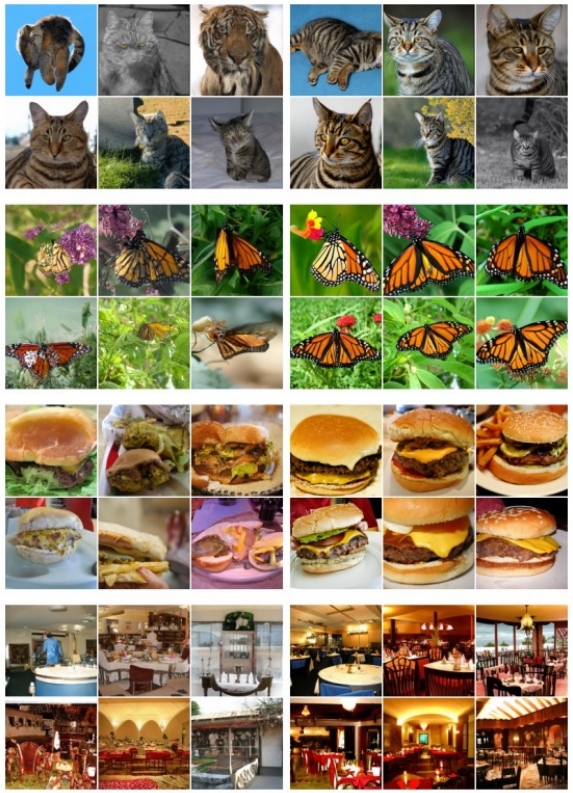
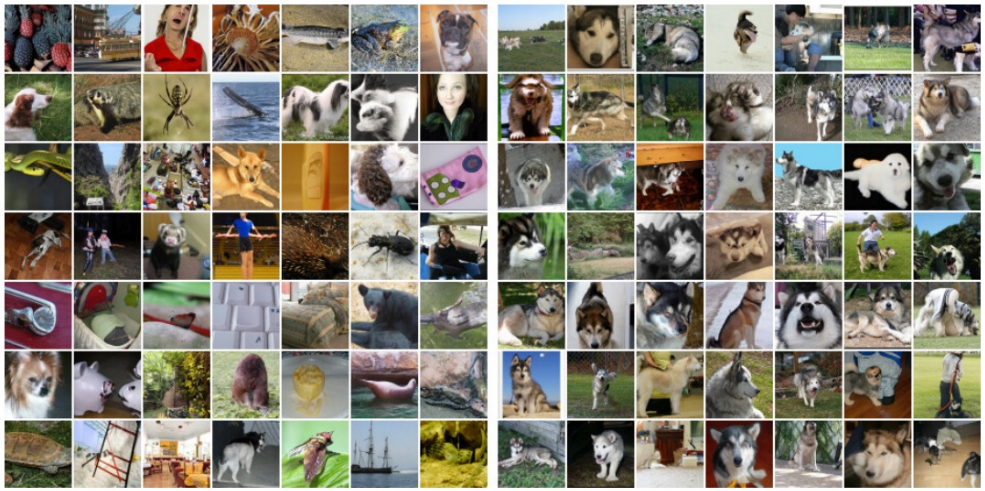
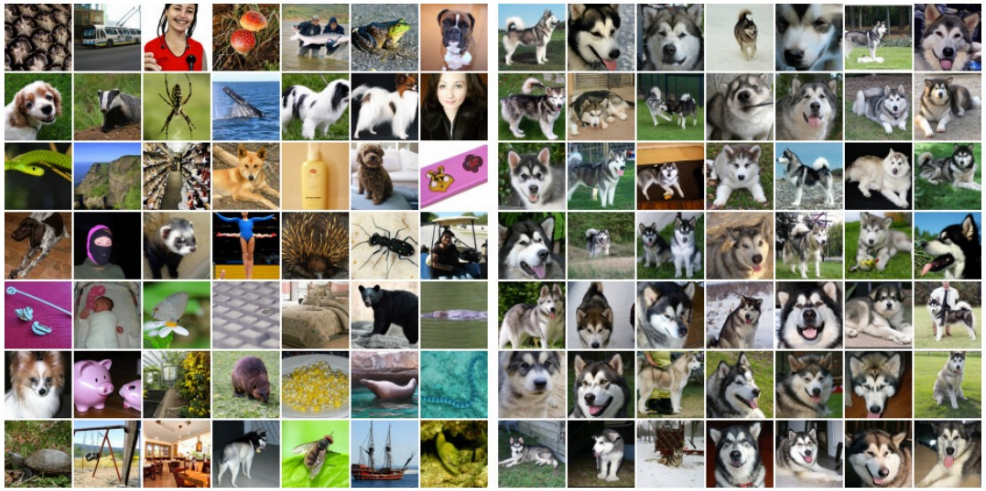


Figure 8: More examples of classifier-free guidance on 128x128 ImageNet. Left: non-guided samples, right: classifier-free guided samples with $w = 3.0$.

Control the Diffusion Model: Classifier-Free Guidance

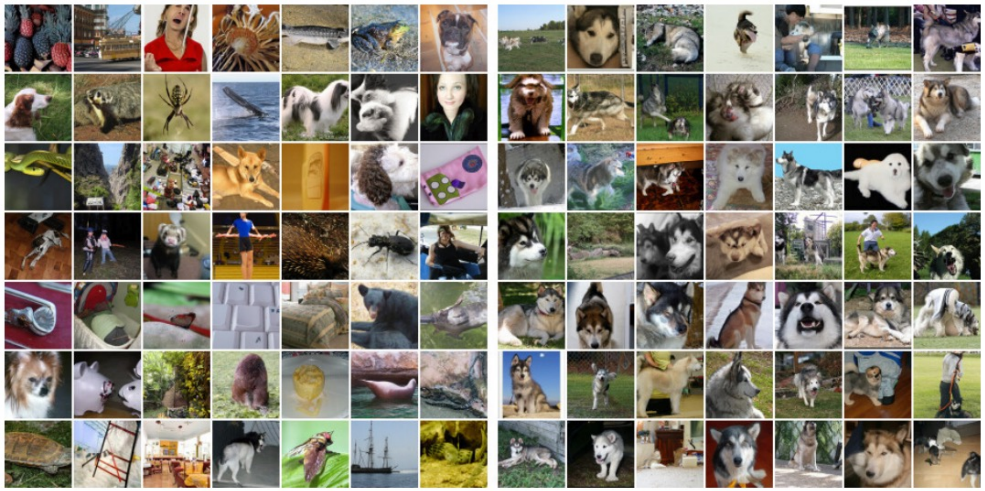


(a) Non-guided conditional sampling: FID=1.80, IS=53.71

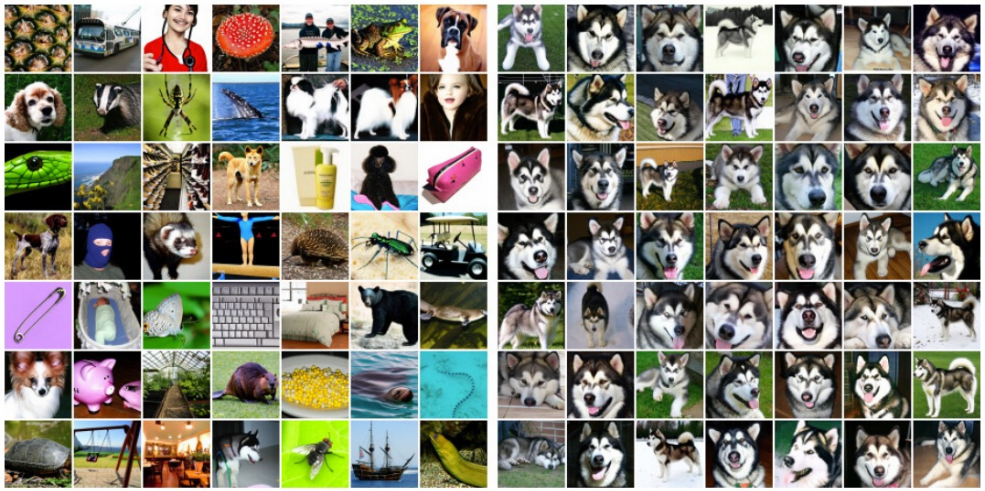


(b) Classifier-free guidance with $w = 1.0$: FID=12.6, IS=170.1

Control the Diffusion Model: Classifier-Free Guidance



(a) Non-guided conditional sampling: FID=1.80, IS=53.71



(c) Classifier-free guidance with $w = 3.0$: FID=24.83, IS=250.4

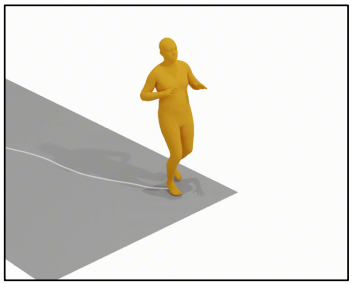
Analysis of Classifier-Free Guidance Weight Schedulers

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, Vicky Kalogeiton, submission 2024

Introduction

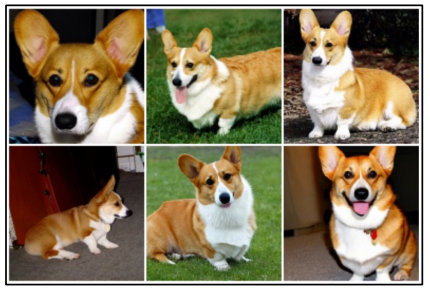
- Classifier-Free Guidance is the key method for conditioning diffusion models based on various input modalities (label, text, etc.)
- $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$
- CFG consists of generation term + guidance term and ω is used to control the conditioning magnitude

prompt condition: Darth Vader is surfing on the waves. [From SVD]



prompt condition: A person is running backwards quickly. [From MDM]

prompt condition: An astronaut is riding a green horse. [From SDXL]



Label condition: "Corgi"[From CFG]

[Ho et al., 2022]

Introduction

- Classifier-Free Guidance is the key method for conditioning diffusion models based on various input modalities (label, text, etc.)
- $\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \omega\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$
- CFG consists of generation term + guidance term and ω is used to control the conditioning magnitude
- As a hyperparameter, tuning guidance scale ω is important to balance the generation quality, textual adherence and generation diversity

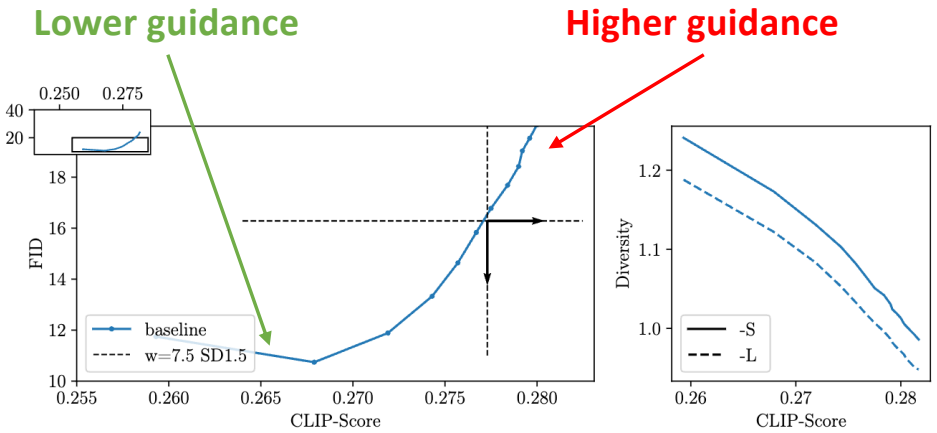
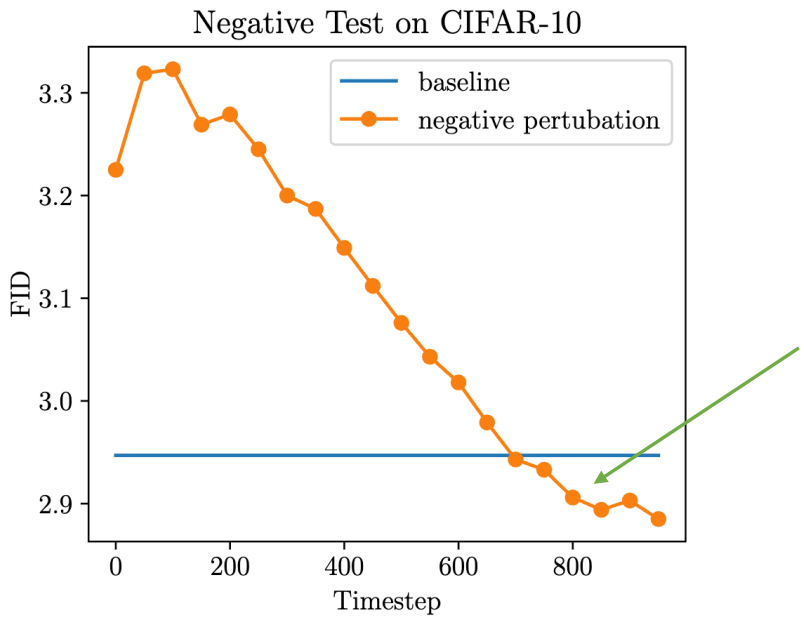


Figure. FID vs. CLIP-Score and Diversity vs. CLIP-Score on different guidance scale

[Ho et al., 2022]

Negative Perturbation Experiment

- Remove varying intervals of guidance scale with respect to the timestep of the generation



Observation:
Removing the initial stage of Classifier-Free Guidance
→
improves generation quality (FID)
→
constant guidance: not effective design

Solution

$$\epsilon_\theta(x_t, t, y) = \epsilon_\theta(x_t, t) + \boldsymbol{\omega(t)}\,(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t))$$

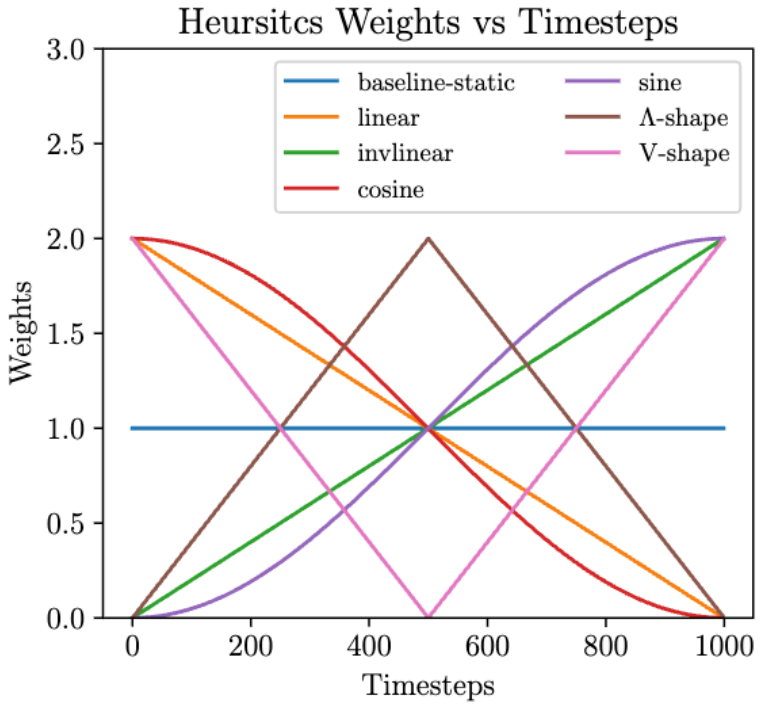
Replace constant guidance, we with guidance schedulers **ω(t)** that vary according to generation timesteps

- Two families:
  - Heuristic functions
  - Parametrized functions

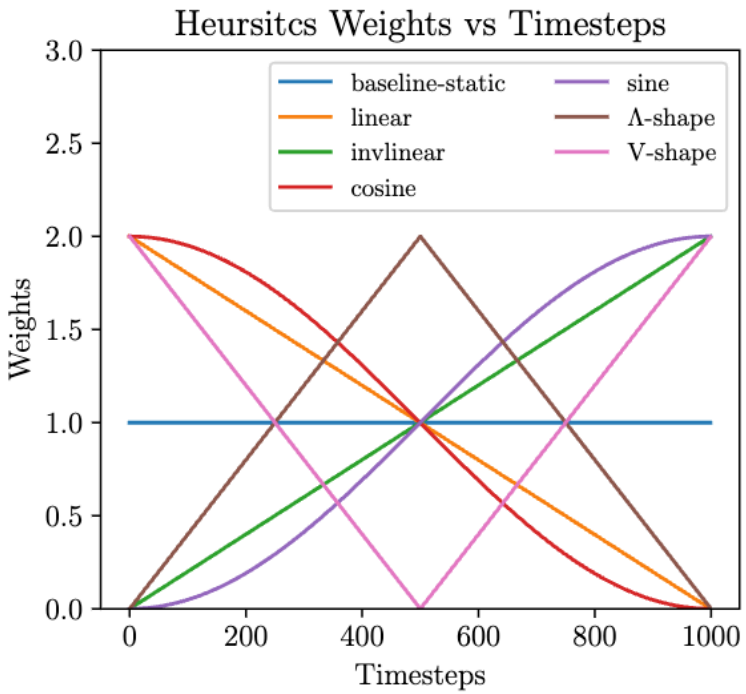
- Analyze results

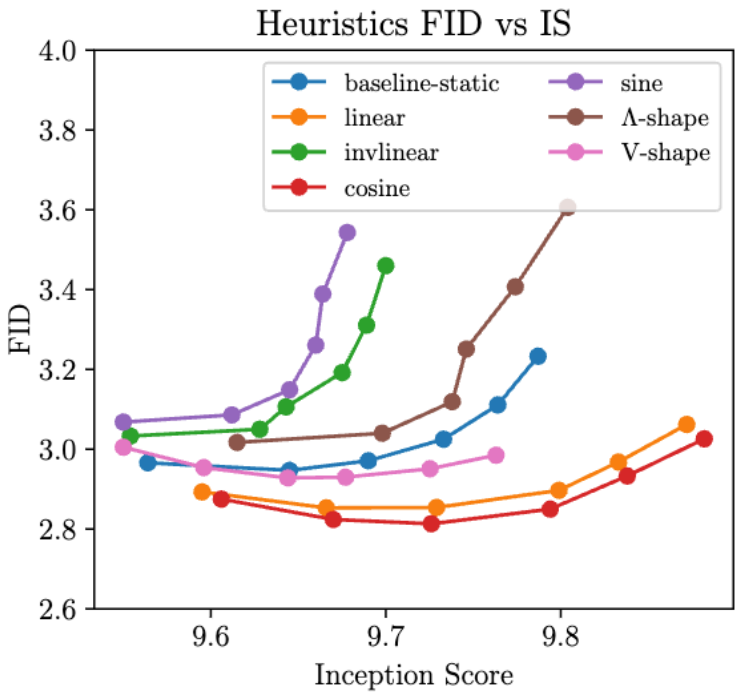
Replace static by Heuristic functions

$$\begin{aligned}
\text{linear: } \omega(t) &= 1 - t/T, \\
\text{invlinear : } \omega(t) &= t/T, \\
\text{cosine: } \omega(t) &= \cos(\pi t/T) + 1, \\
\text{sine: } \omega(t) &= \sin(\pi t/T - \pi/2) + 1, \\
\text{V-shape: } \omega(t) &= \text{invlinear}(t) \ \textit{if } t < T/2, \ \text{linear}(t) \ \textit{else}, \\
\Lambda\text{-shape: } \omega(t) &= \ \text{linear}(t) \ \textit{if } t < T/2, \ \text{invlinear}(t) \ \textit{else}.
\end{aligned}$$



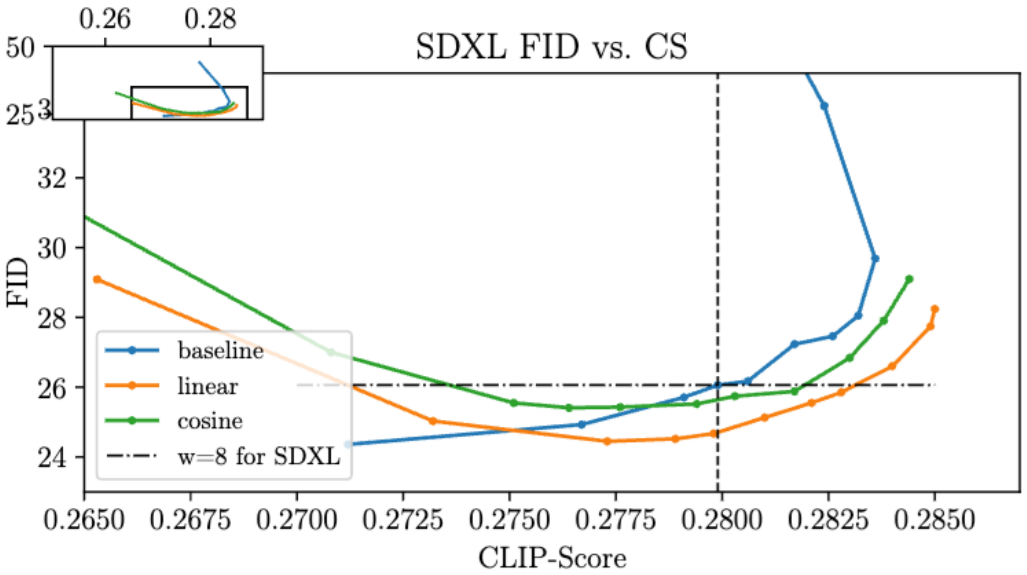
Quantitative Results: Heuristic functions Class-conditional generation

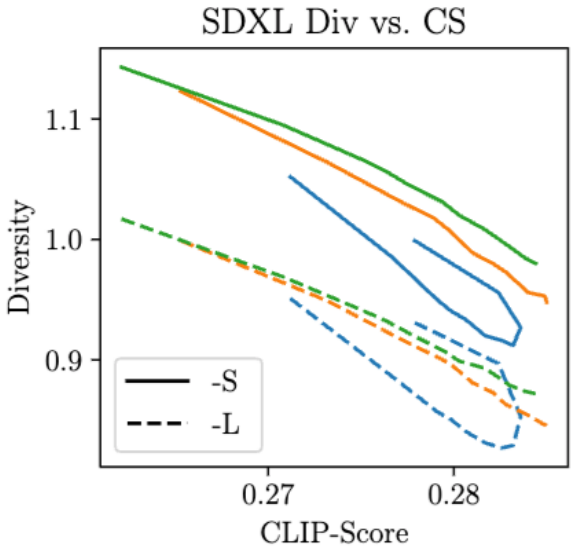




Monotonically increasing shape heuristic performs the best

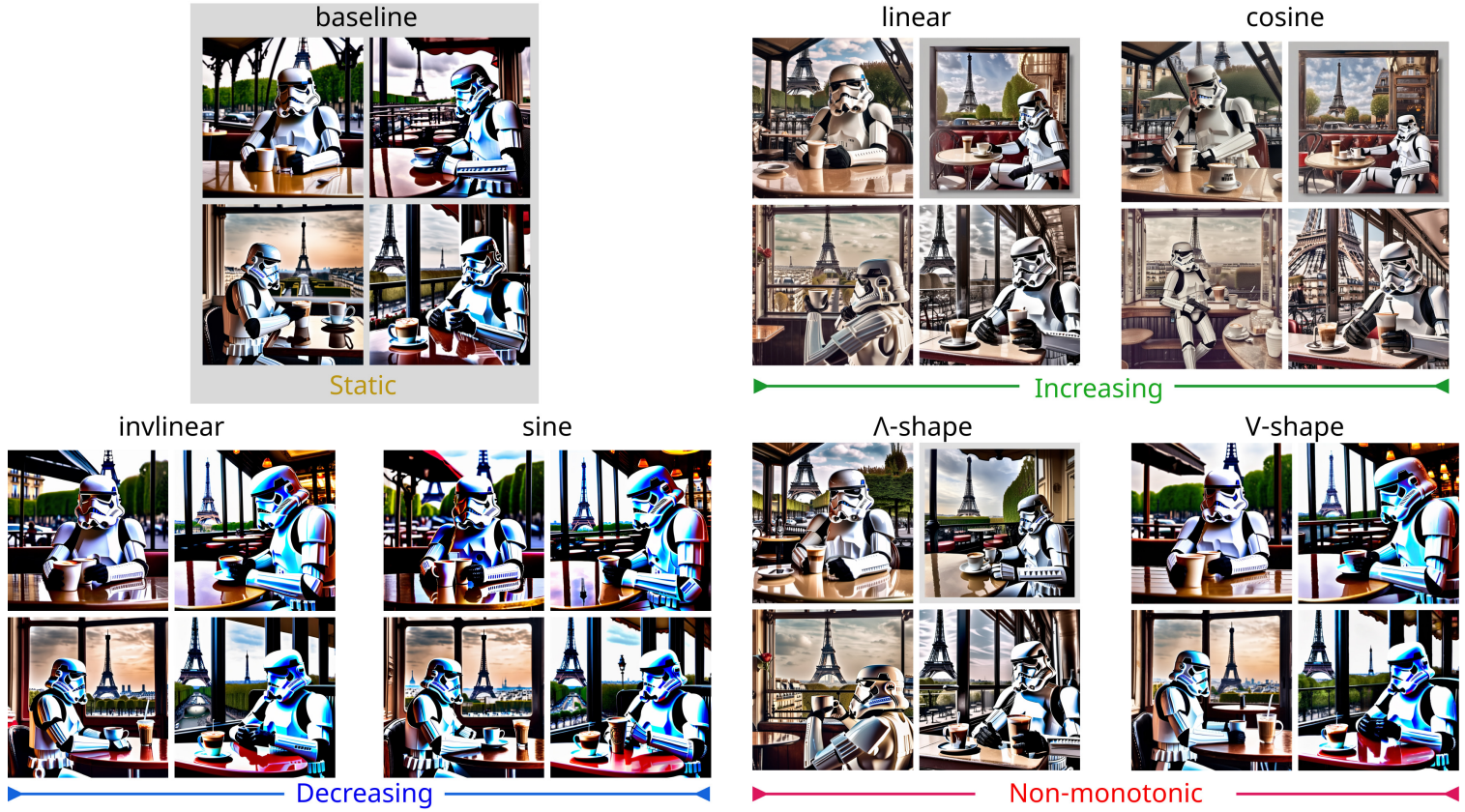
Quantitative Results: Heuristic functions Text-to-image generation



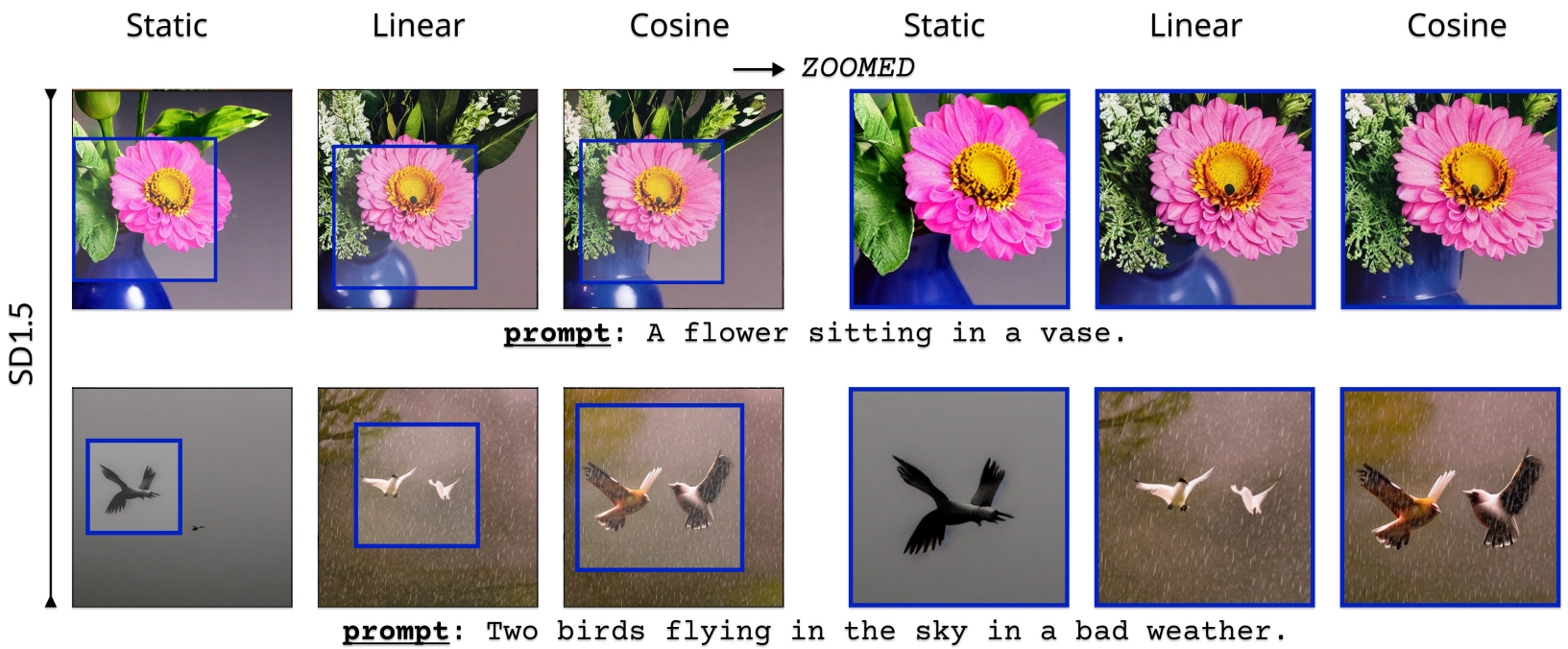


Monotonically increasing shape guidance schedulers achieve a better balance *of quality, conditional adherence, and diversity*

Qualitative Results: Heuristic functions
Text-to-image generation



Qualitative Results: Heuristic functions Text-to-image generation



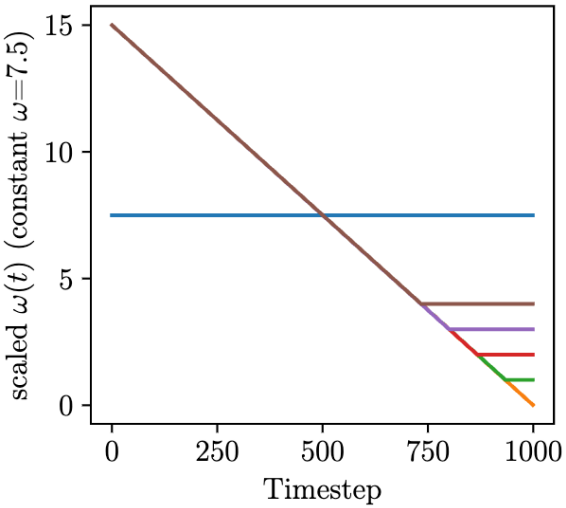
Better quality

Qualitative Results: Heuristic functions Text-to-image generation

Better diversity

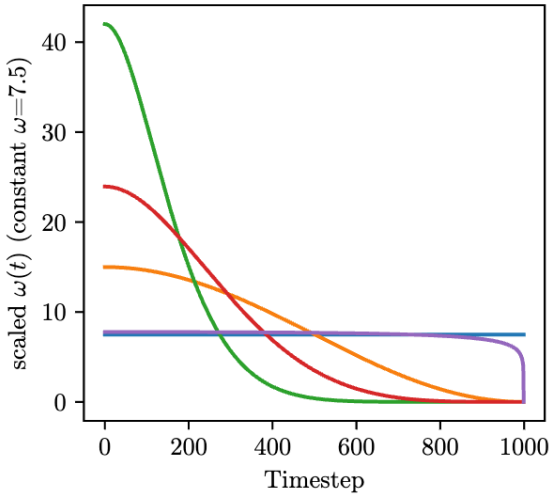
Replace static by Parametrized functions

Clamping



$\omega(t, c) = \max(\omega(t), c)$;
$\omega(t)$ = linear, cosine, etc.

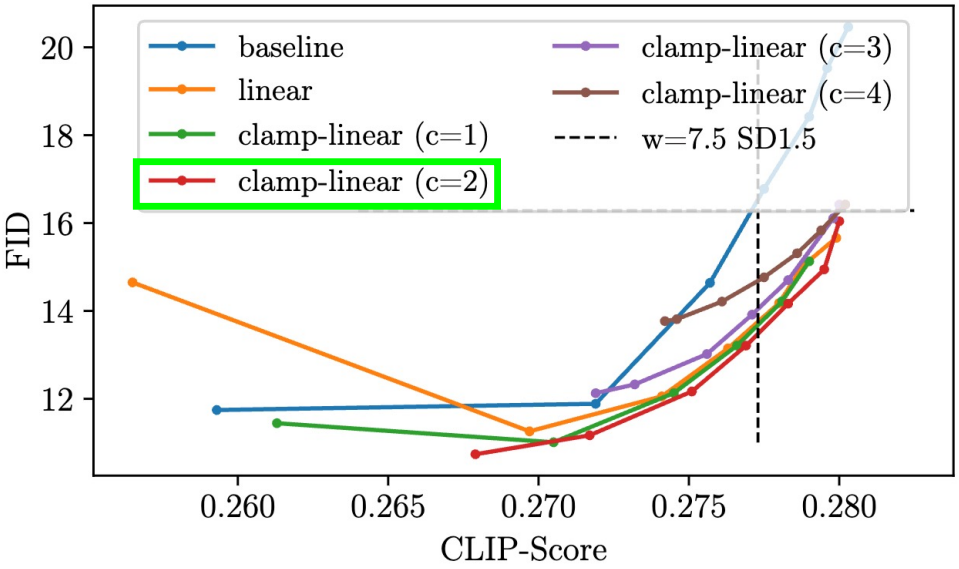
Parametrized cosine



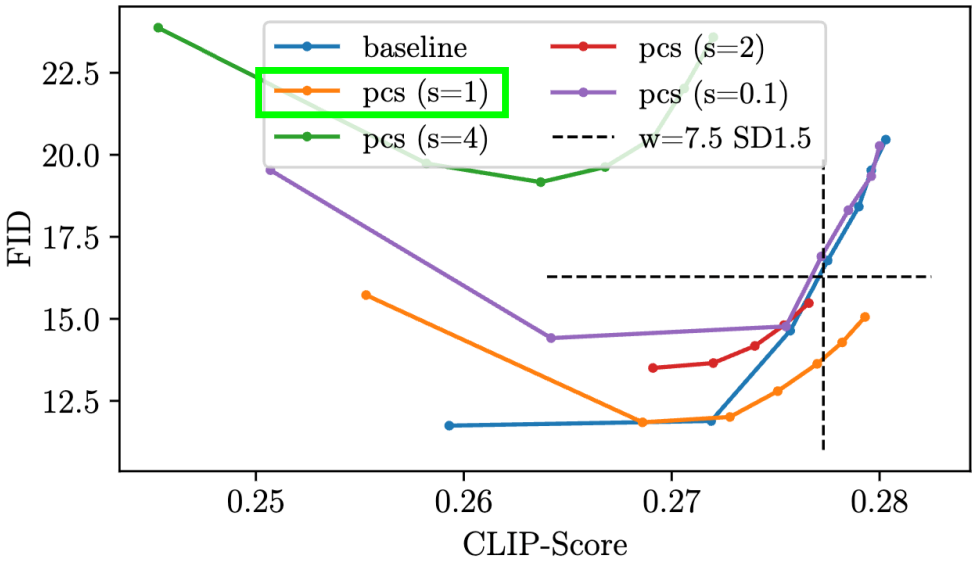
$$\omega(t, s) = \frac{1 - \cos \pi(\frac{T - t}{T})^s}{2} \omega$$

Quantitative Results: Parametrized functions

Clamping



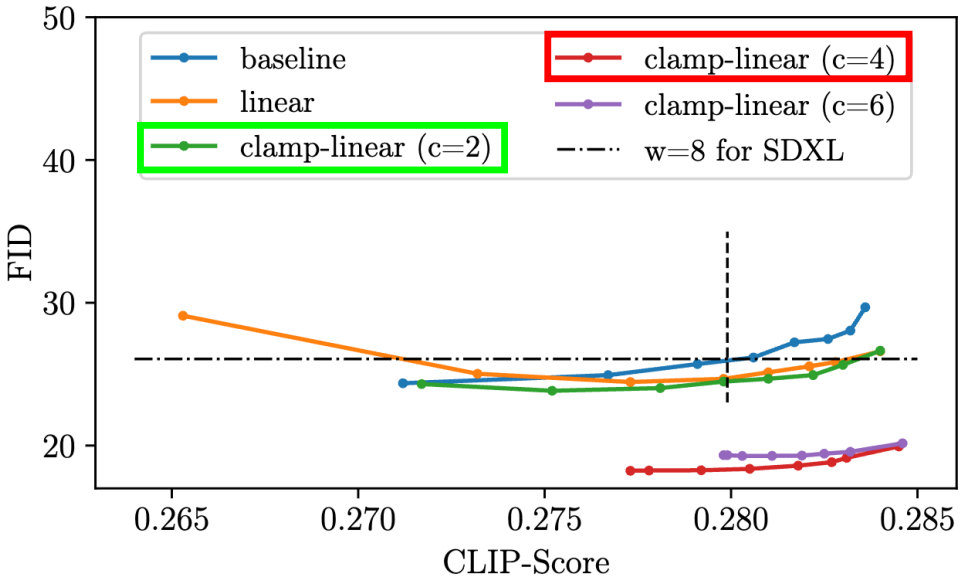
Parametrized cosine



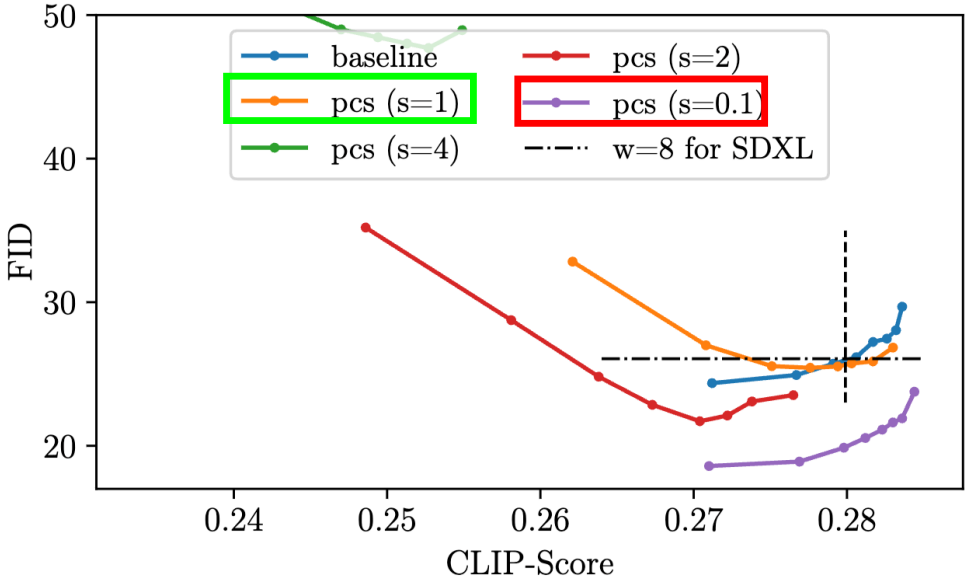
SD1.5

Quantitative Results: Parametrized functions

Clamping



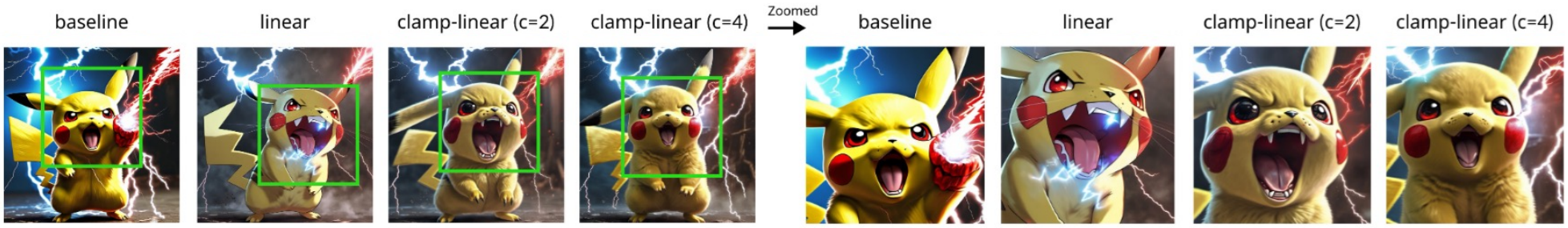
Parametrized cosine



Observation: tuning correctly can improve the performance, but the tuning is *not generalizable*

SDXL

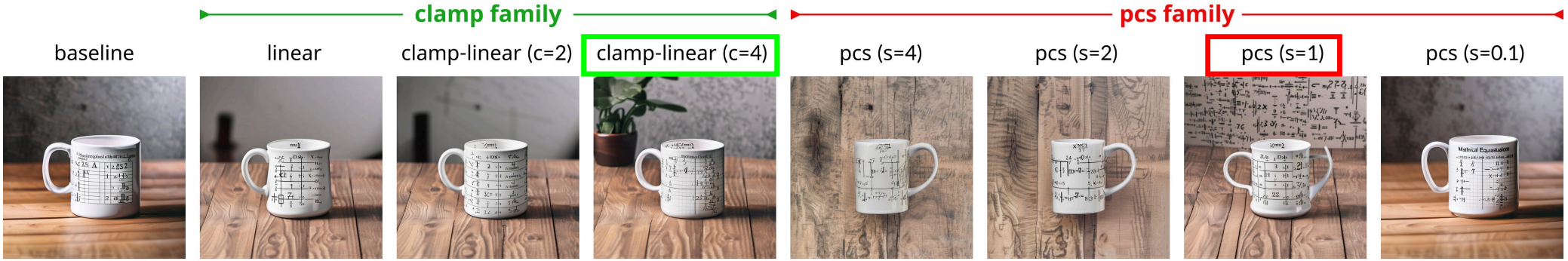
Qualitative Results: Parametrized functions



Textual comprehension, fidelity, attention to detail

SDXL

Qualitative Results: Parametrized functions



Prompt: A mug with mathematical equations put a wooden table.

Prompt: A black car running on the road with a lot of trees on the side.

+ better details (mug)
+ more realistic (car)
+ better textured background (mug)

Conclusion

- Among heuristic functions, monotonically increasing guidance schedulers enhance both performance and diversity
- Well-tuned parameterized functions can achieve *better performa*nce but risk overfitting and require additional time and computational resources for tuning
- The implementation code is 1-line, w/o retraining the model

Low static guidance:

```
w = 2.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✗ Fuzzy images, but many details and textures

High static guidance:

```
w = 14.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✗ Sharp images, but lack of details and solid colors

Dynamic guidance:

```
w0 = 14.0
for t in range(1, T):
  eps_c = model(x, T-t, c)
  eps_u = model(x, T-t, 0)
  # clamp-linear scheduler
  w = max(1, w0*2*t/T)
  eps = (w+1)*eps_c - w*eps_u
  x = denoise(x, eps, T-t)
```

✓ Sharp images with many details and textures, without extra cost.

"full body, a cat dressed as a Viking, with weapons in his paws, on a Viking ship, battle coloring, glow hyper-detail, hyper-realism, cinematic, trending on artstation"

ControlNet

ControlNet

- Paper title “Adding Conditional Control to Text-to-Image Diffusion Models”
- International Conference on Computer Vision (ICCV) 2023
- By Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
- Stanford University

- First appeared online: February 2023
- Published: October 2023

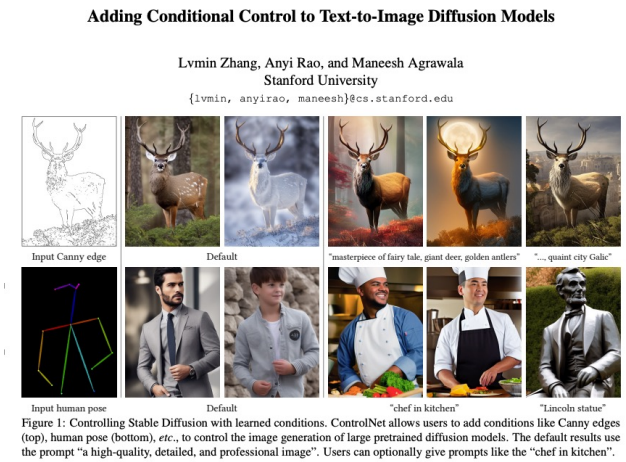
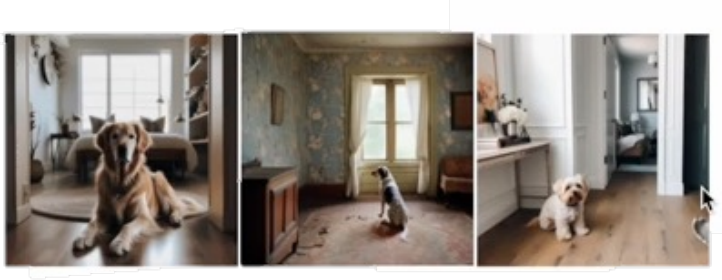


Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt “a high-quality, detailed, and professional image”. Users can optionally give prompts like the “chef in kitchen”.

Introduction

- Basic form of using diffusion models (e.g. Stable Diffusion) is text-to-image
  - Use text prompts as the conditioning to steer image generation so that you generate images that match the text prompt

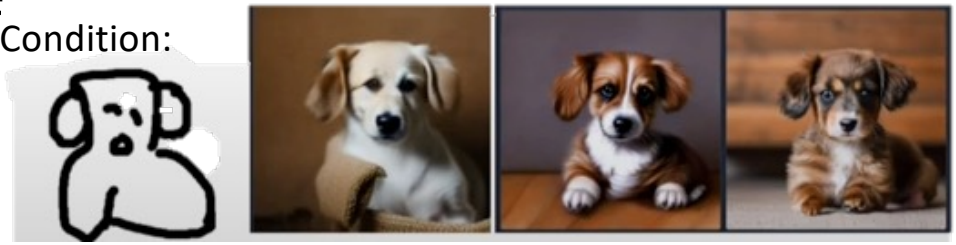
Prompt: “Dog in a room”



- Control the output by giving more input conditions
  - Keep properties from text
  - Adhere to additional properties from condition

Prompt: “Dog in a room”

Condition:



Introduction

- NN architecture that helps you **control** pre-trained diffusion models (such as Stable Diffusion model) by adding extra conditions

Goodies:

✓End-to-end architecture

✓Robust on small dataset (<50k images)

✓As fast as fine-tuning

✓Can be trained on personal devices

✓Can scale to large amounts of data (millions to billions)

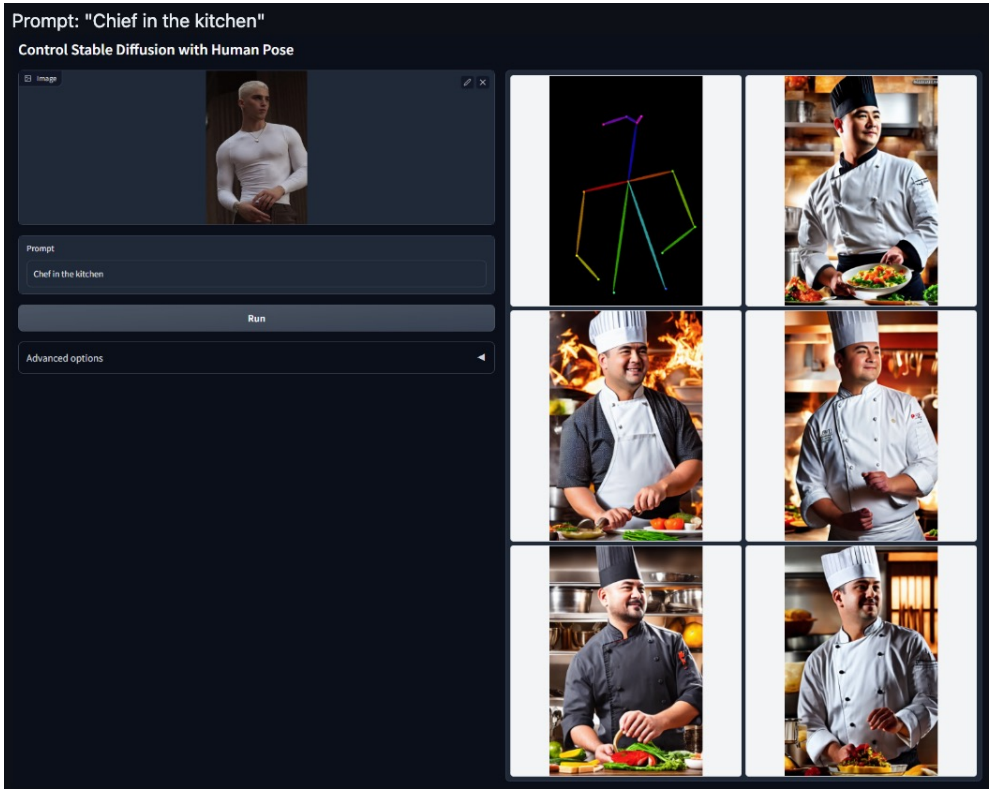
Introduction

- ControlNet adds one more conditioning in addition to the text prompt
  - allows for the manipulation of existing diffusion model architectures
  - think of it as a way to make slight changes to a neural network's structure and add desired properties or characteristics

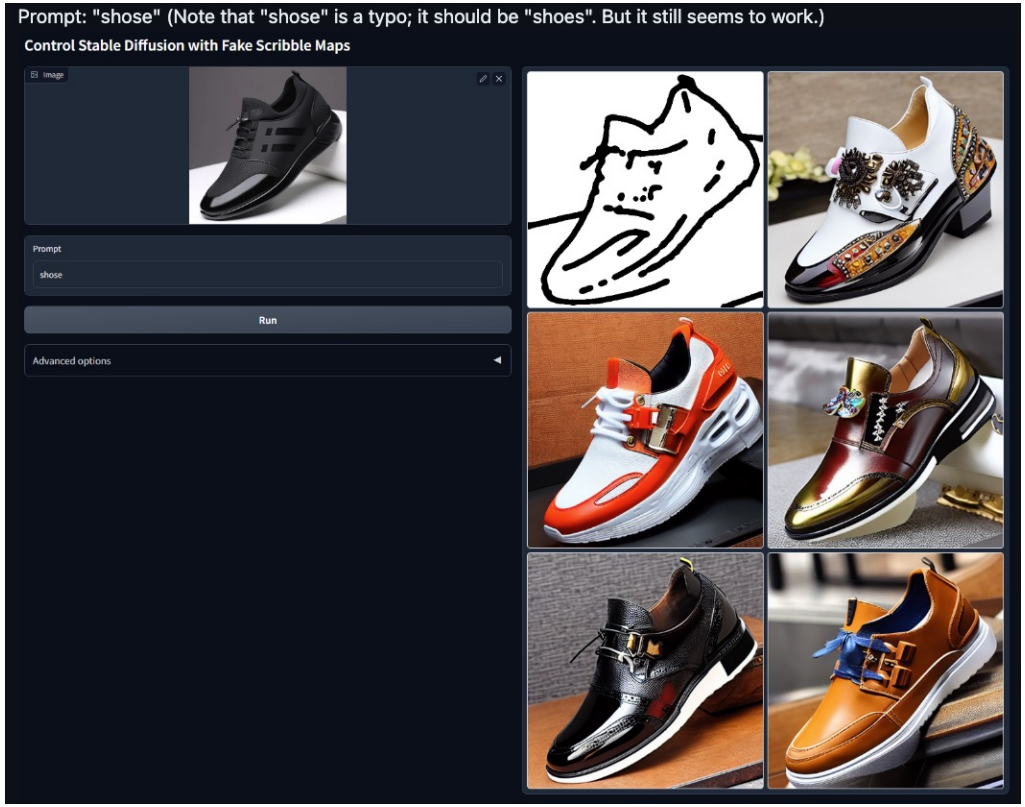
- The extra conditioning can take many forms
  - Segmentation map
  - Depth map
  - Pose
  - Infrared
  - HED map
  - Hough Line
  - Cartoon Line Drawing
  - …

Examples

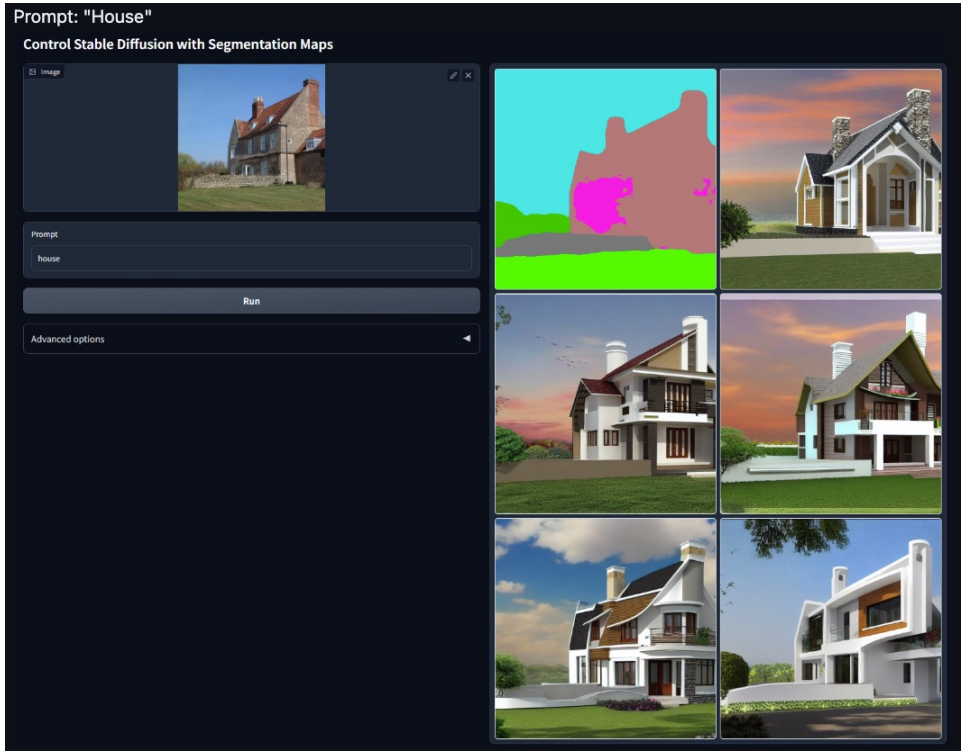
ControlNet examples
Condition: Pose

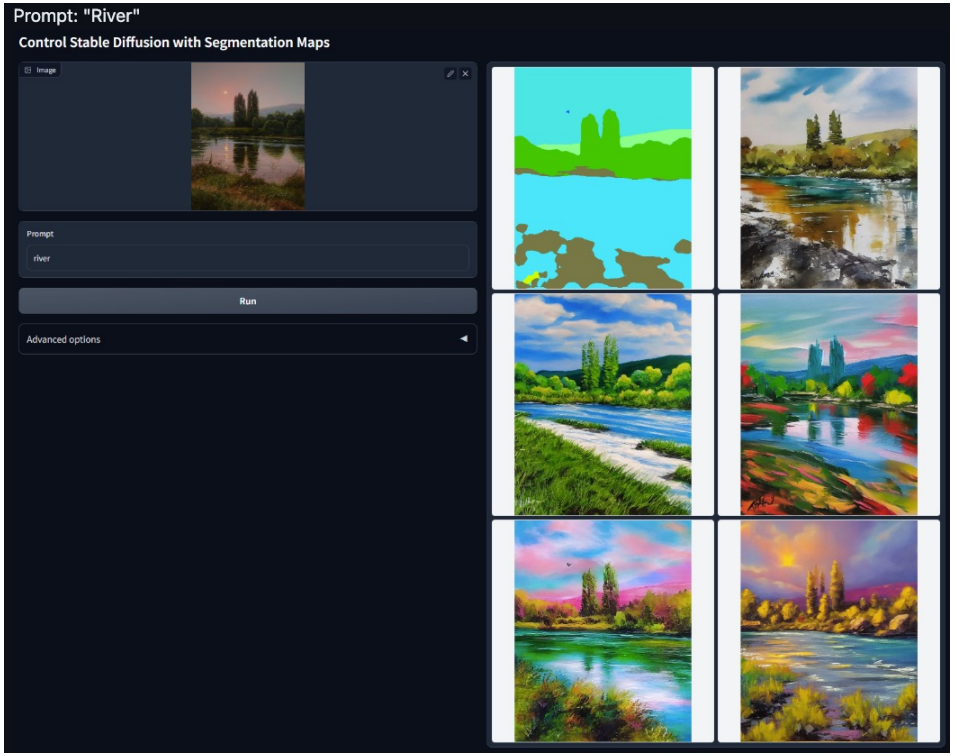


ControlNet examples
Condition: Scribble

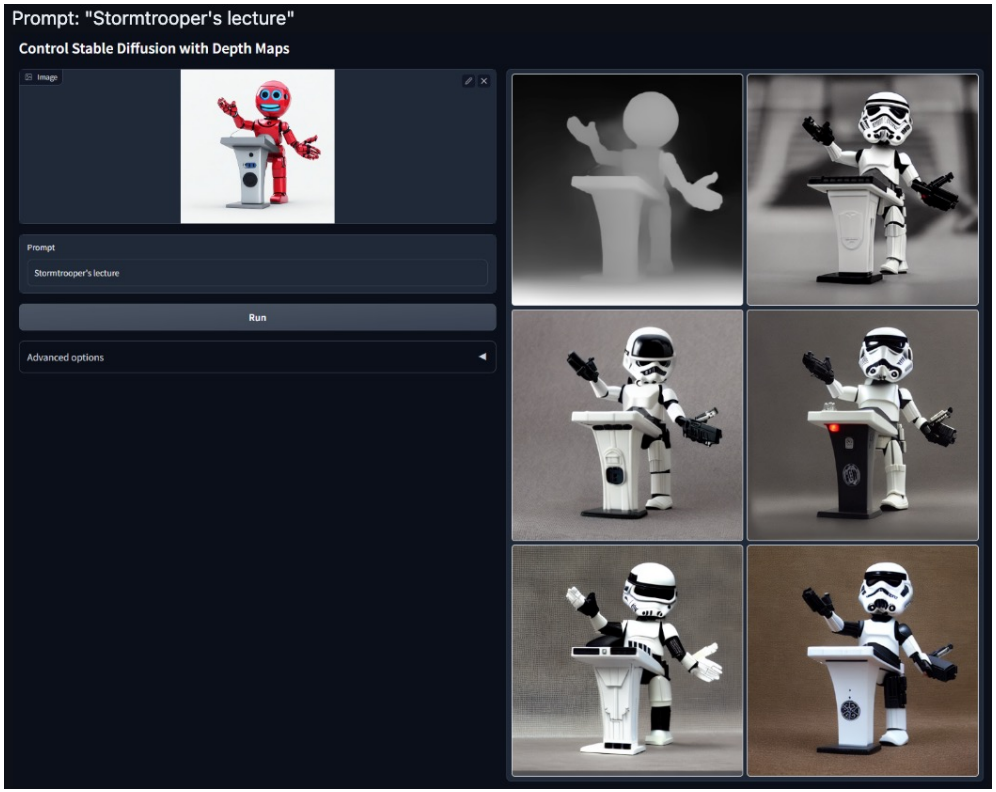


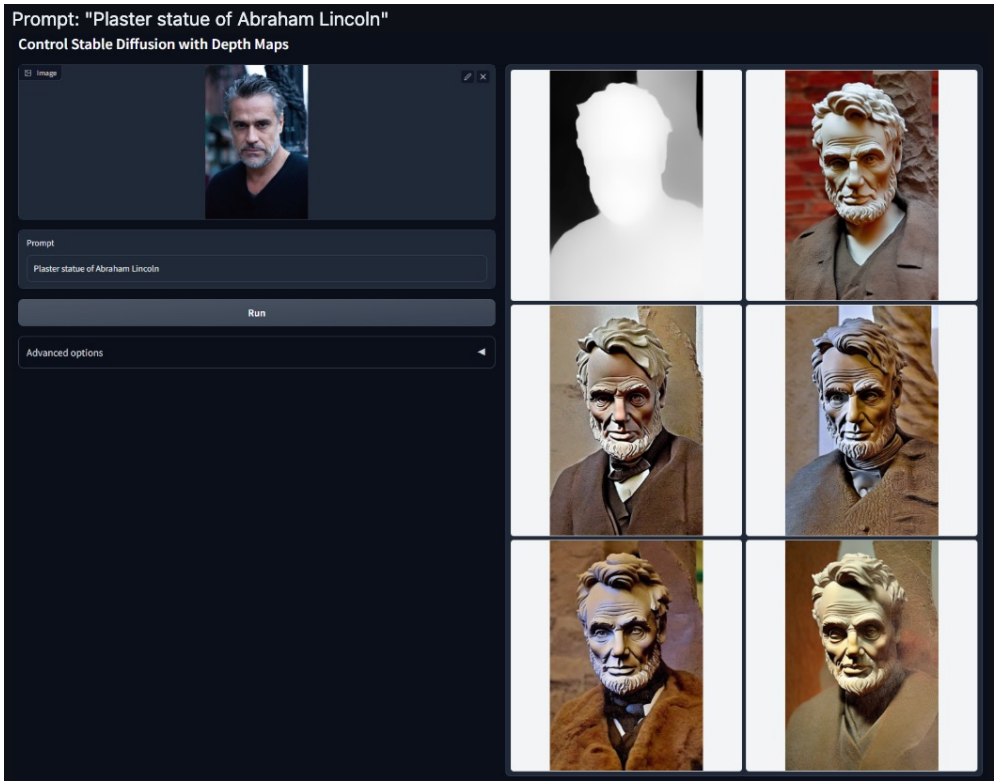
ControlNet examples
Condition: Segmentation map



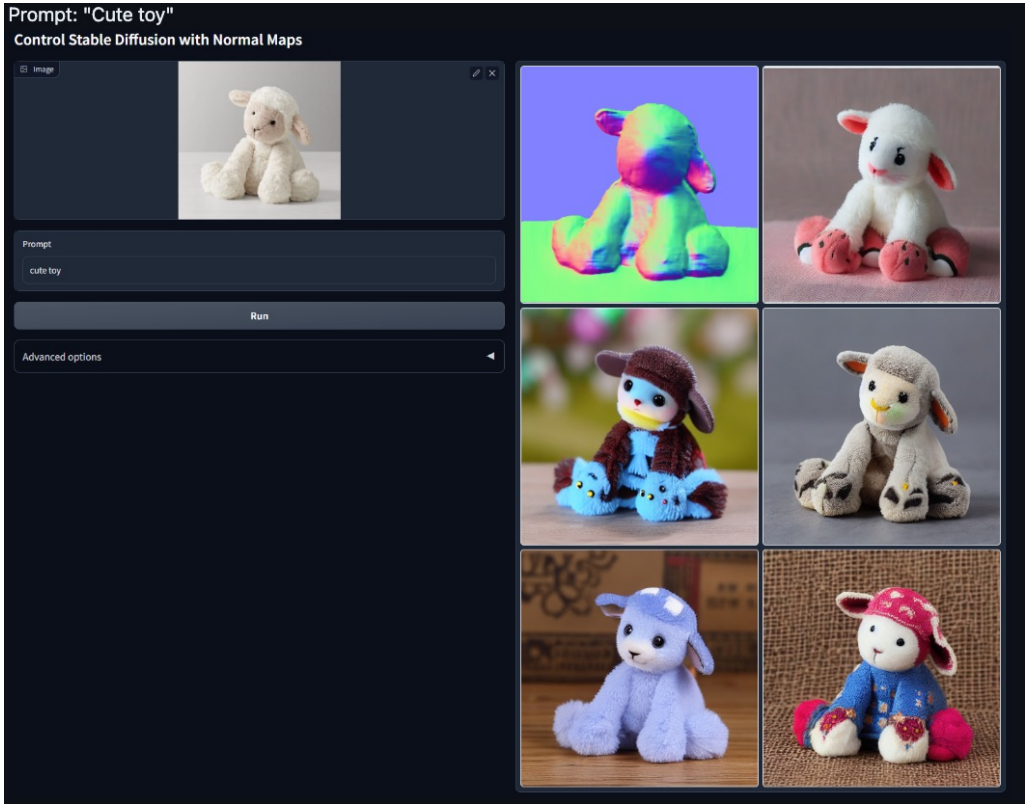


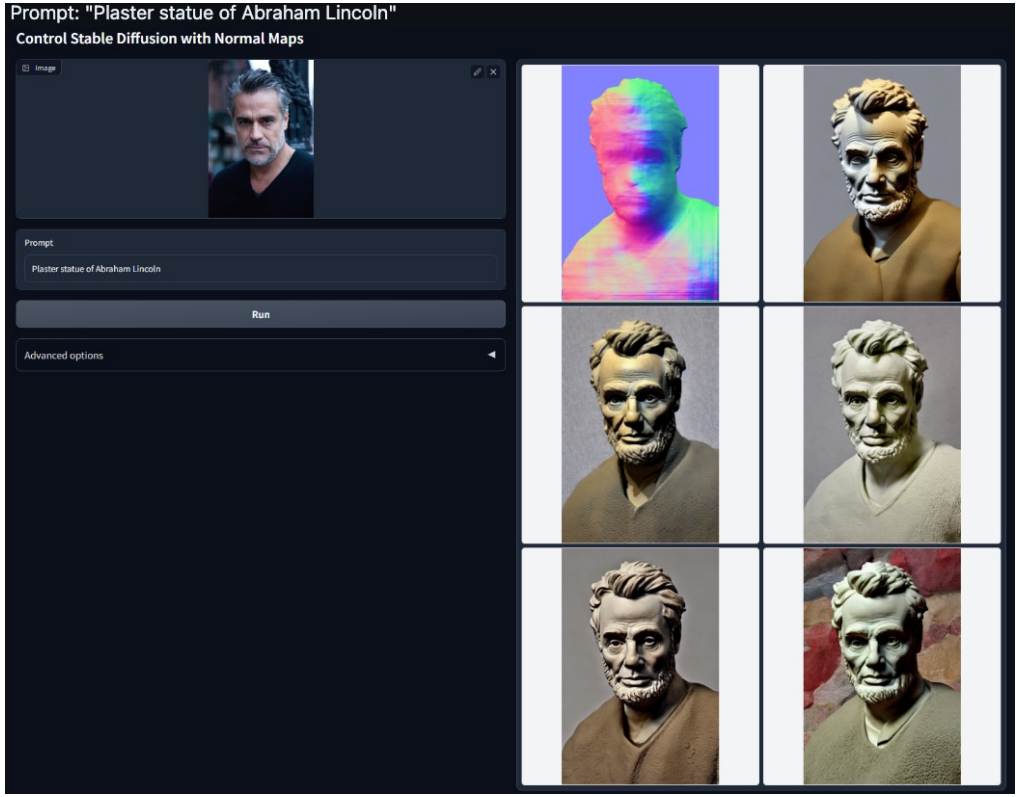
ControlNet examples
Condition: Depth map



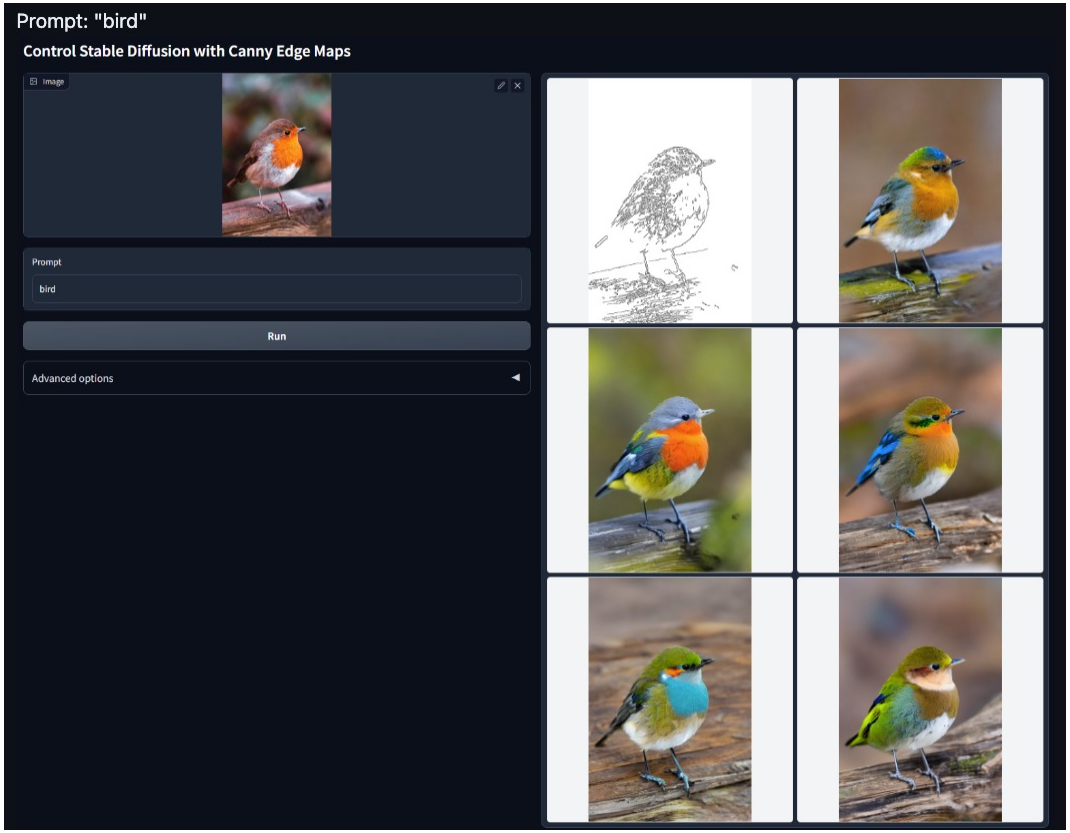


ControlNet examples
Condition: Normal map

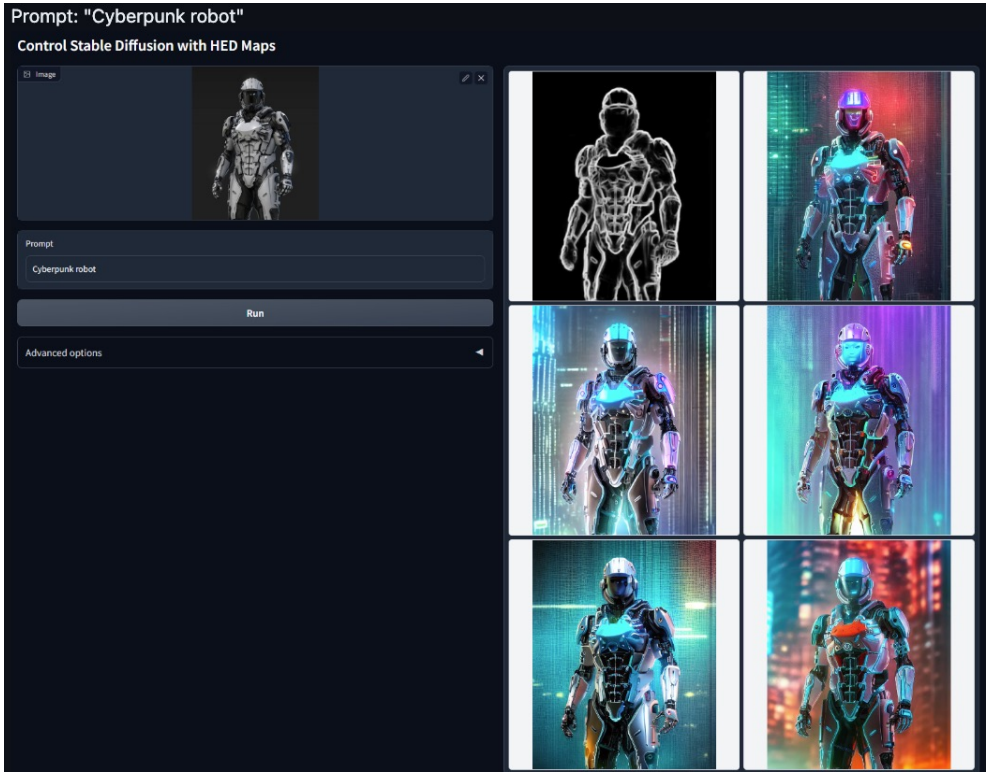




ControlNet examples
Condition: Canny Edge map

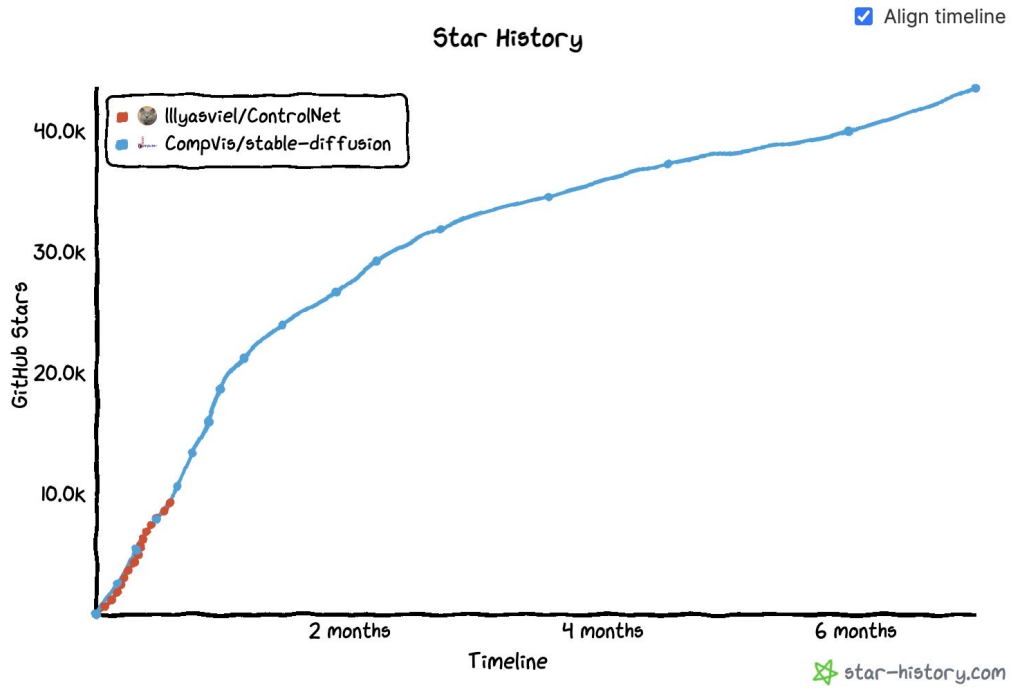


ControlNet examples
Condition: HED Map



Growth & Accessibility

Growth of ControlNet



Co-founder & CEO HuggingFace

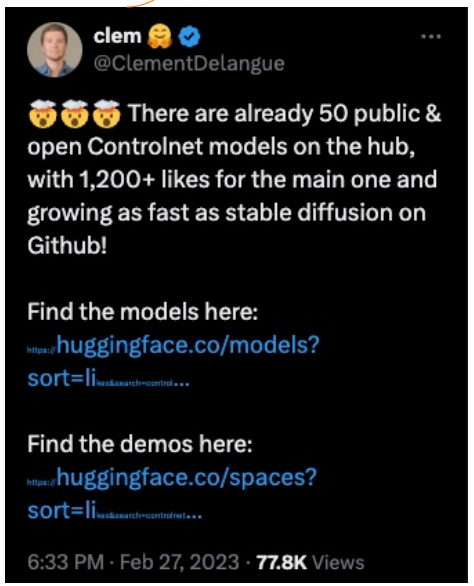
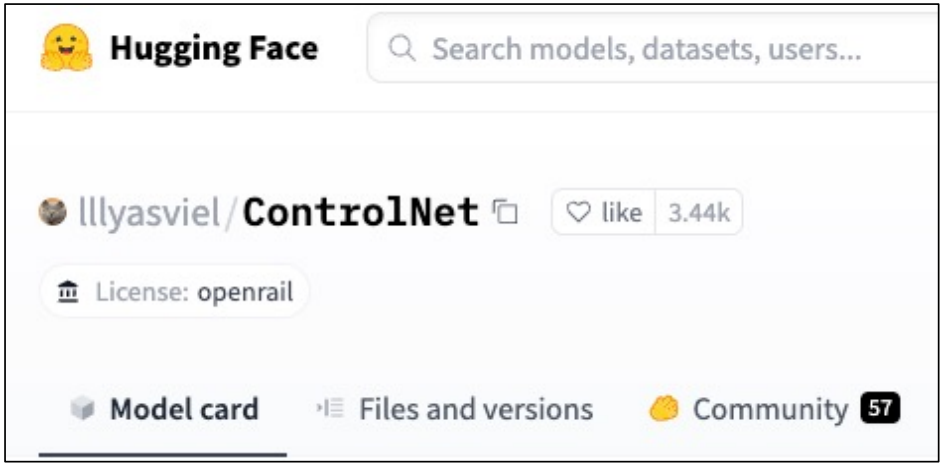
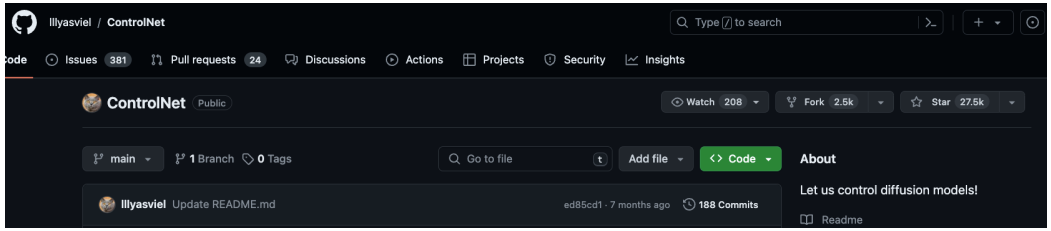


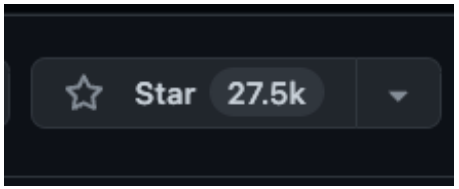
Image source: https://twitter.com/ClementDelangue/status/1630259781742067718/photo/1

ControlNet access



https://huggingface.co/lllyasviel/ControlNet





https://github.com/lllyasviel/ControlNet

Motivation

Motivation

- Can large models be applied to facilitate specific tasks?
- What kind of framework should we build to handle the wide range of problem conditions and user controls?

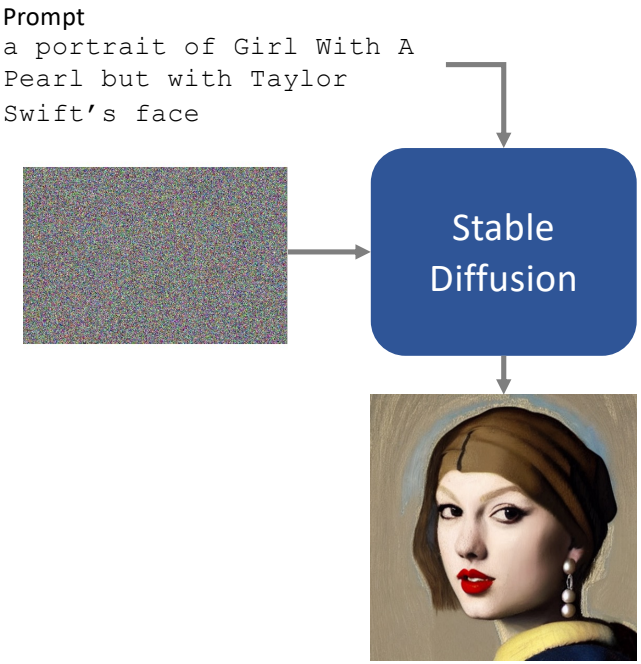
Motivation

- Can large models be applied to facilitate specific tasks?
- What kind of framework should we build to handle the wide range of problem conditions and user controls?
- Three findings:
  - The available data scale in a task-specific domain is not always as large as that in the general image-text domain
  - Large computation clusters are not always available
  - Various image processing problems have diverse forms of problem definitions, user controls, or images annotations

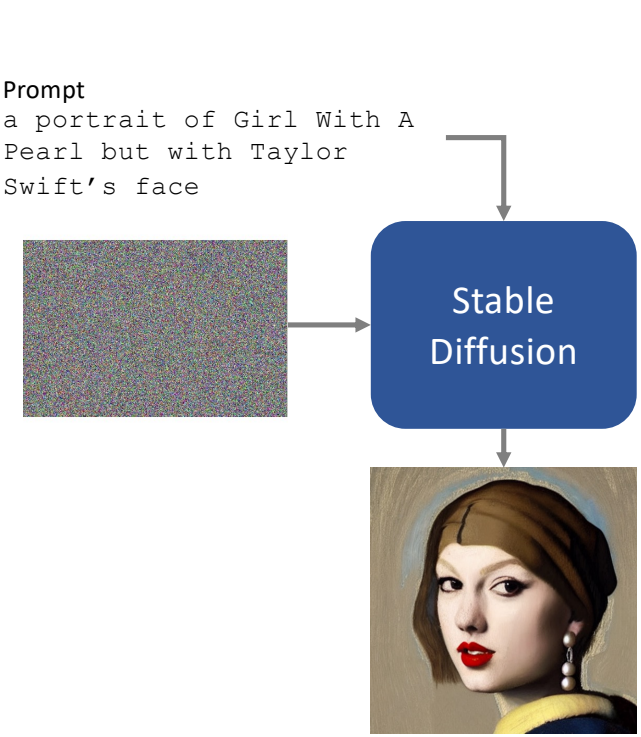
ControlNet Outline

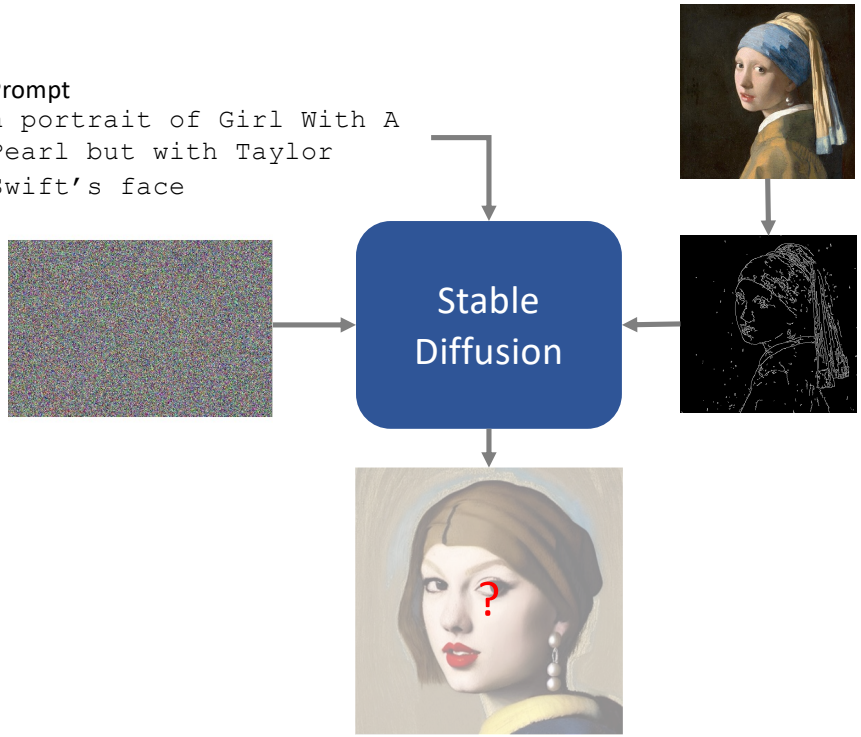
- Introduction
  - ControlNet
  - Examples for condition
  - Growth & Accessibility
  - Motivation
- ControlNet
  - High-level architecture
  - ControlNet Block
  - Architecture
- Experiments
  - Examples
  - Analysis
  - Why does it work?
  - Comparison to state of the art

ControlNet

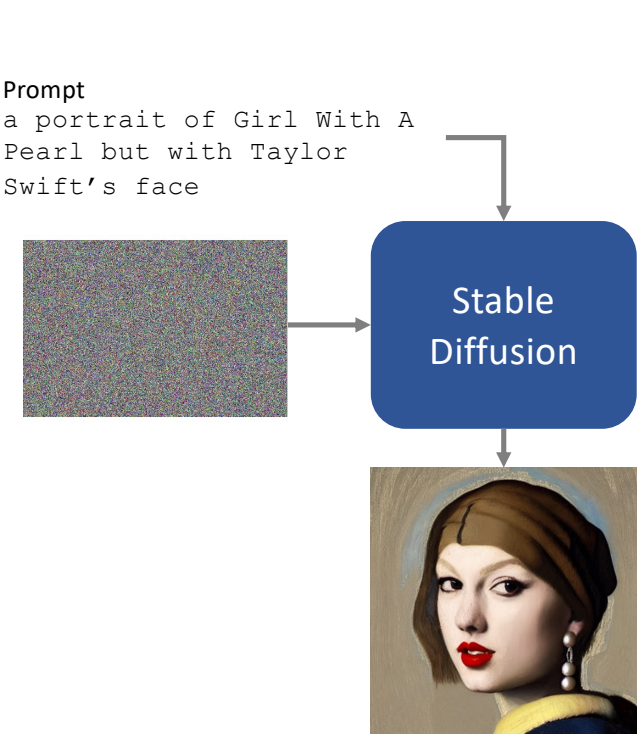

Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
Stable Diffusion

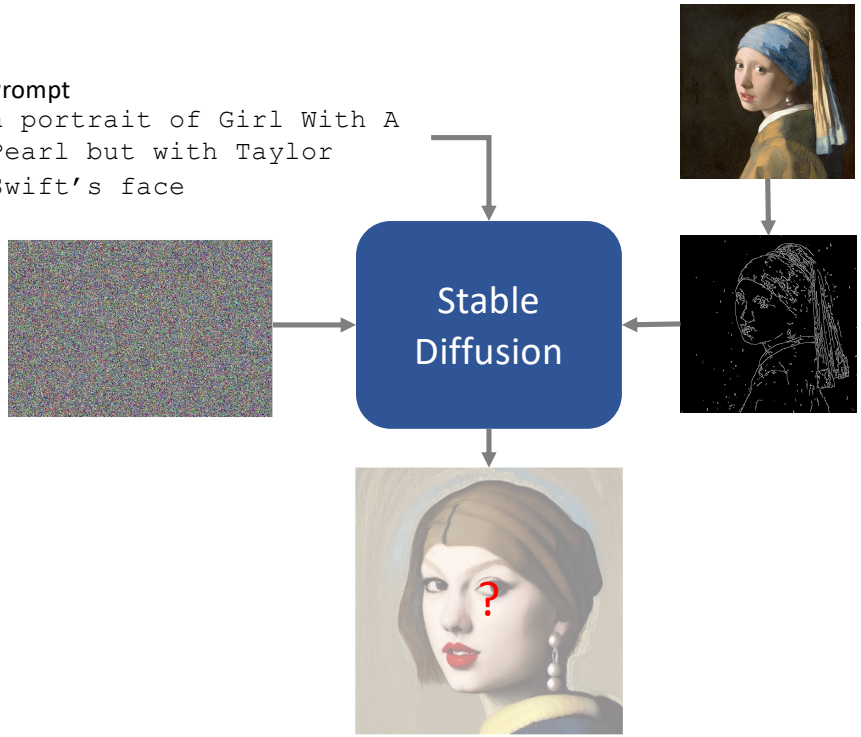
ControlNet


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
Stable Diffusion


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
Stable Diffusion
?

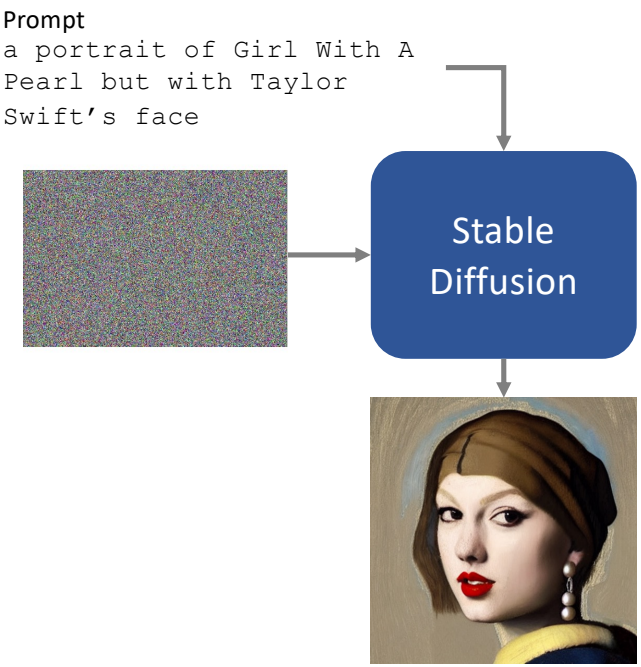
ControlNet

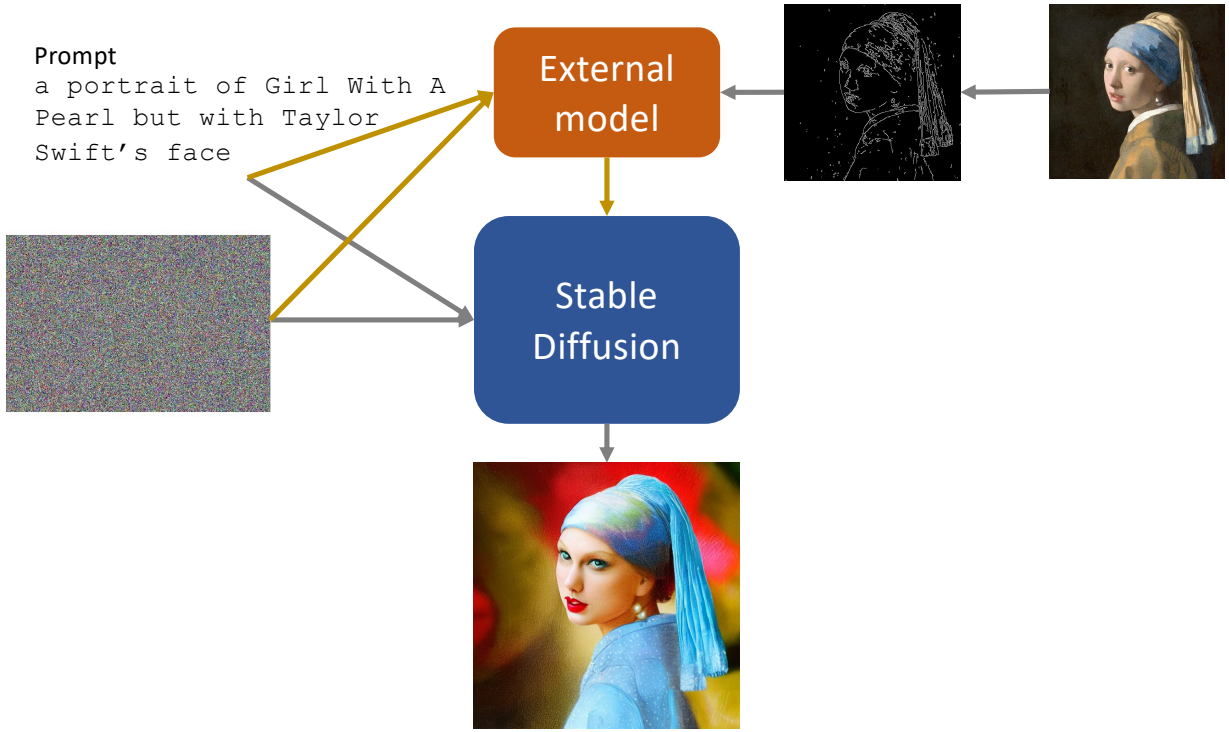




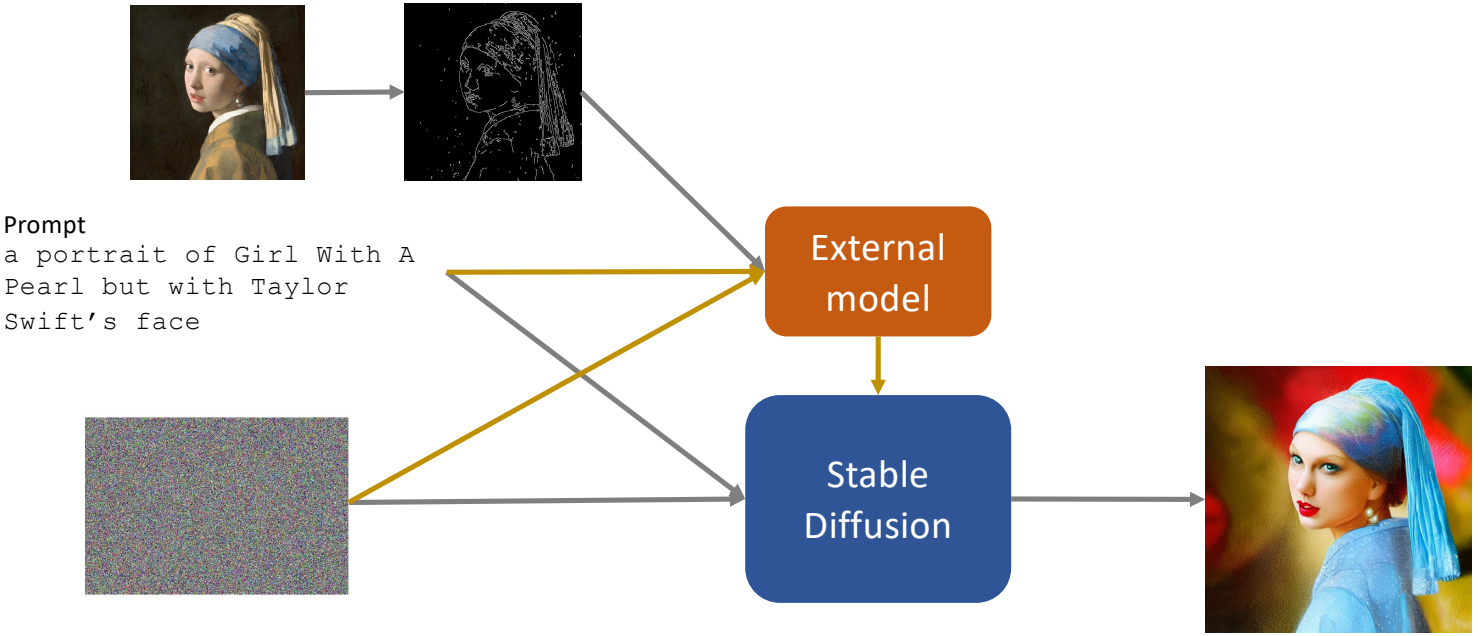
What is the most efficient way to train a model to take in additional conditioning inputs?

ControlNet

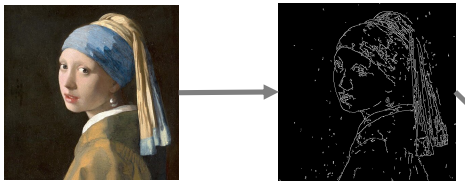

Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
Stable Diffusion


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
External model
Stable Diffusion

ControlNet

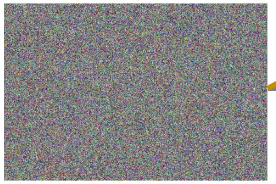

Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
External model
Stable Diffusion

ControlNet



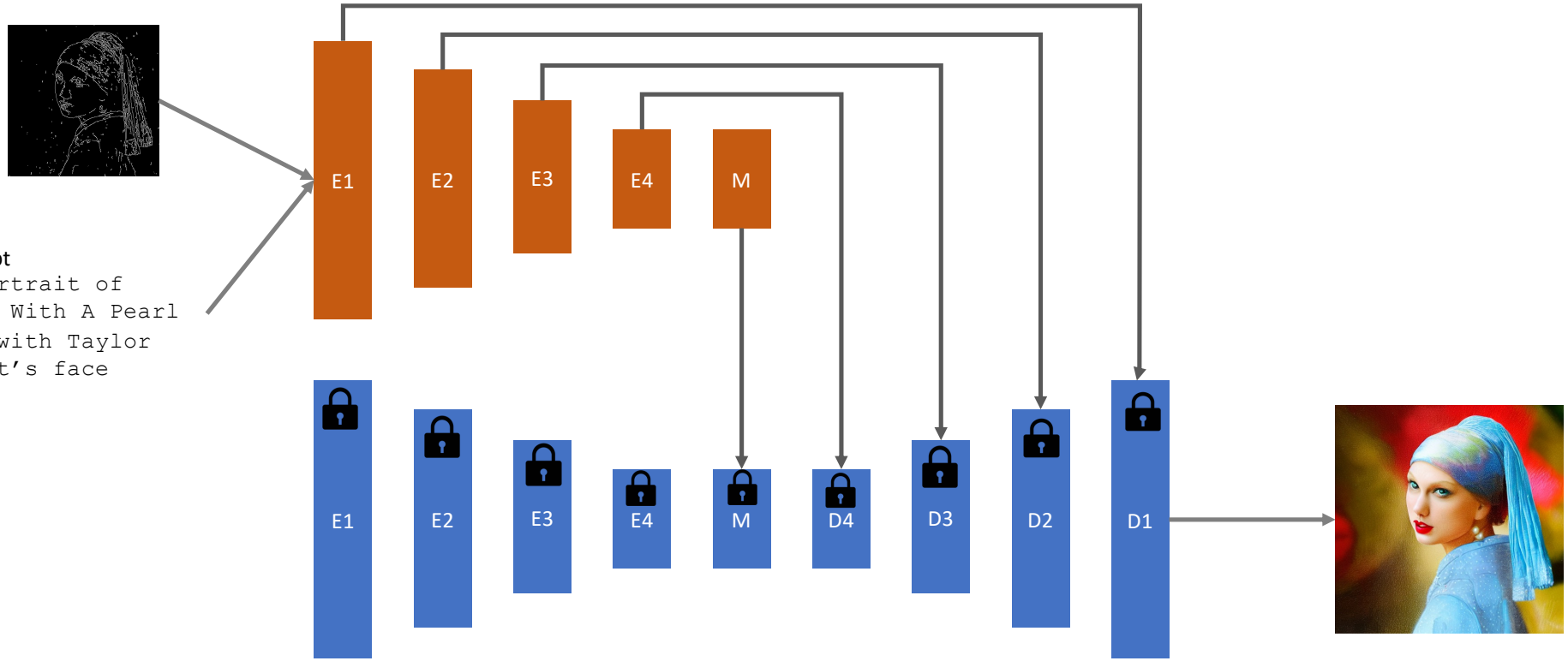
Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face

External model

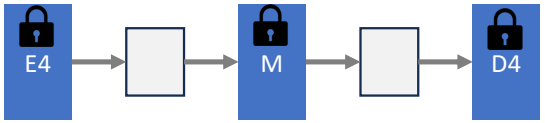


Stable Diffusion

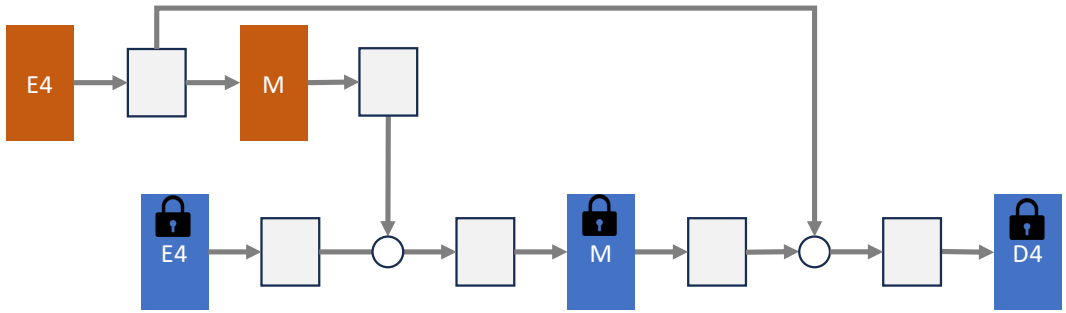
ControlNet


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
E1
E2
E3
E4
M
E1
E2
E3
E4
M
D4
D3
D2
D1

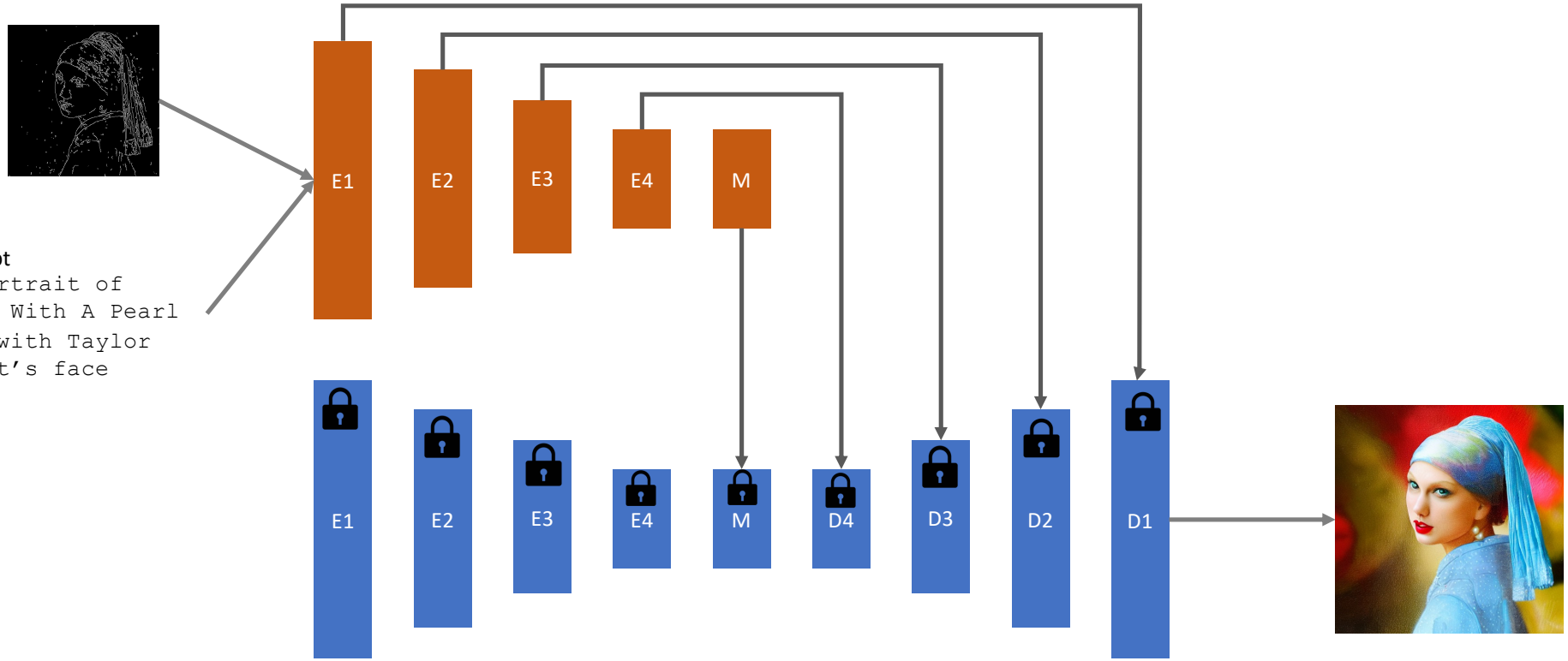
ControlNet



ControlNet

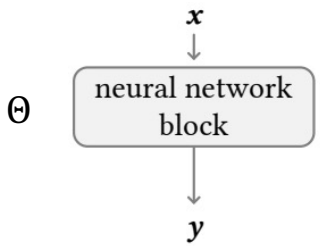


ControlNet


Prompt
a portrait of Girl With A Pearl but with Taylor Swift's face
E1
E2
E3
E4
M
E1
E2
E3
E4
M
D4
D3
D2
D1

ControlNet: Block

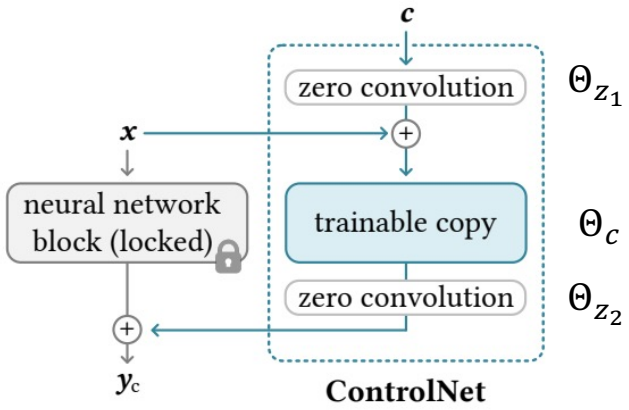
- x : input feature map
- F: Neural Network block with parameters Θ
- y : output feature map



(a) Before

$$y = F(x, \Theta)$$

Manipulate the input conditions of NN blocks to further control the overall behavior of an entire NN



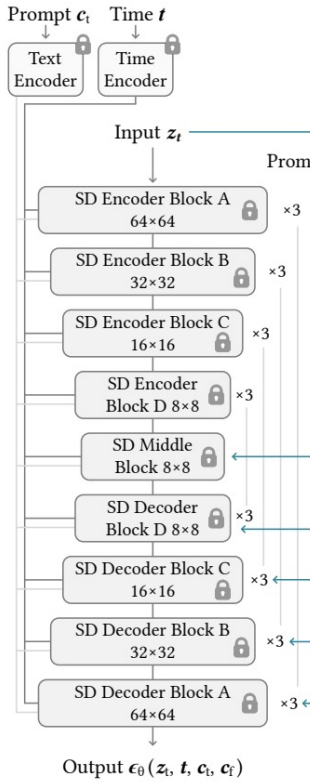
(b) After

$$y = F(x; \Theta) + Z(F(x + Z(c; \Theta_{z_1}); \Theta_c); \theta_{z_2})$$

The direct finetuning or continued training of a large pretrained model with limited data may cause overfitting and catastrophic forgetting

- c: external condition vector
- Θ_c: Trainable copy
- Z: Zero convolution: 1×1 conv. with weights and bias initialized with zeros

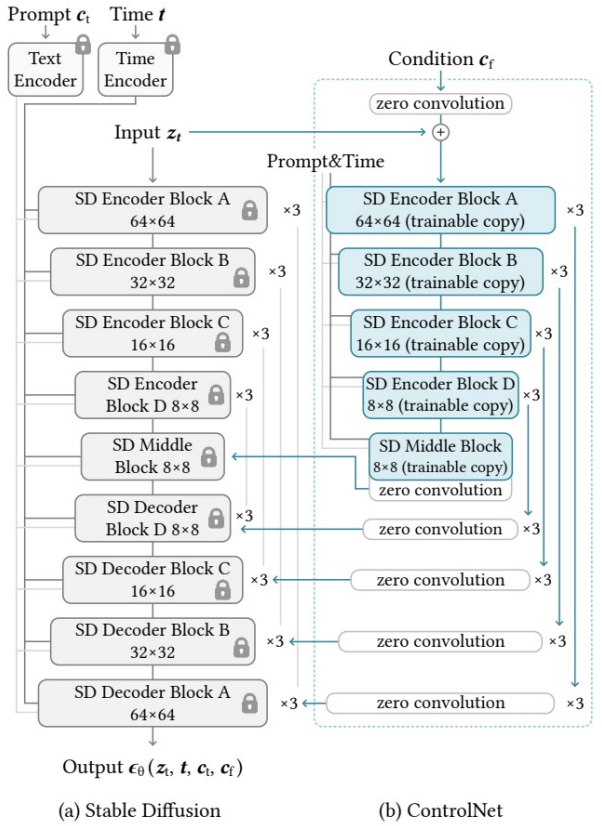
ControlNet: Architecture



(a) Stable Diffusion

ControlNet: Architecture

- Computationally efficient

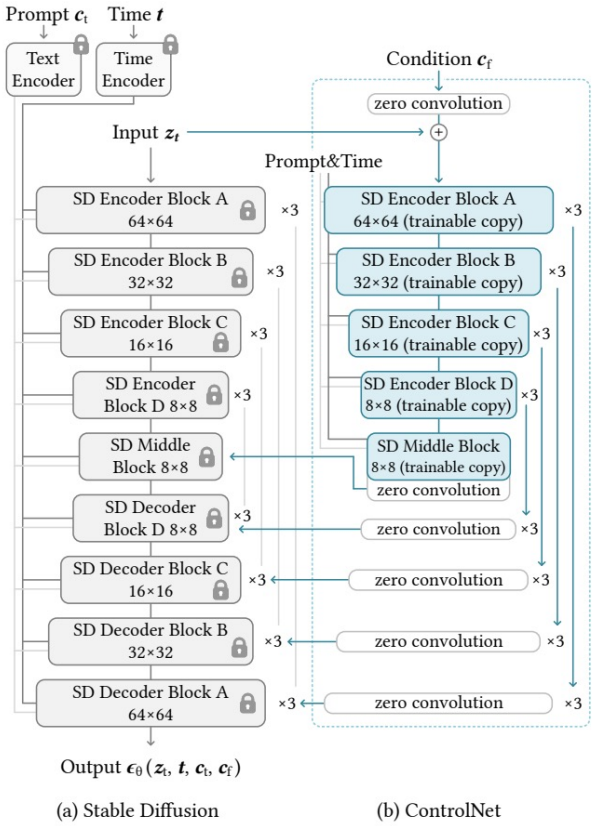


ControlNet: Architecture

- Computationally efficient
- Overall learning objective:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}) \|_2^2 \right]$$

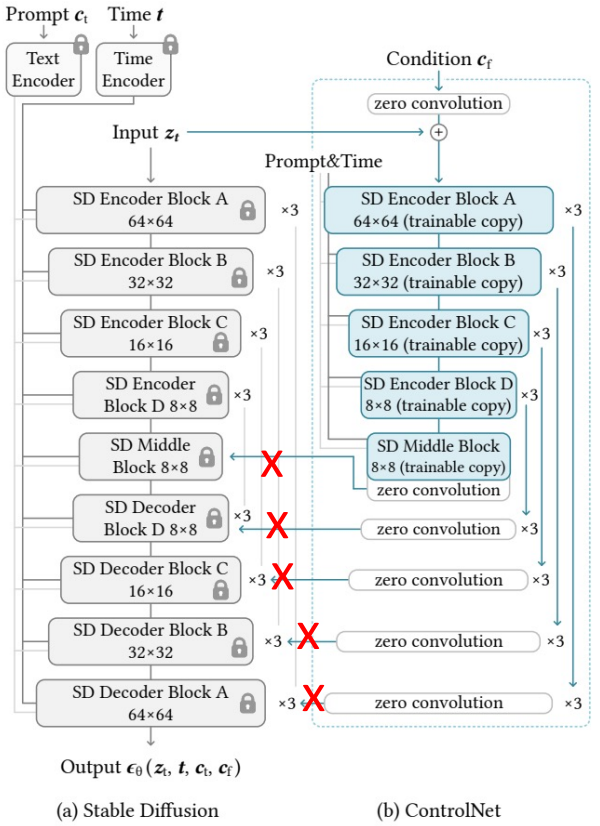
- z_o: original images
- z_t : noisy images
- $t.$: time step
- c_t : text prompts
- c_f : task-specific conditions



ControlNet: Improved Training

- Small-Scale Training
  - Disconnecting the link to decoder 1,2,3,4 and only connecting the middle block

- Large-Scale Training
  - Train ControlNets for a large number of iterations
  - Unlock all weights of Stable Diffusion and jointly train the entire model as a whole



ControlNet Outline

- Introduction
  - ControlNet
  - Examples for condition
  - Growth & Accessibility
  - Motivation
- ControlNet
  - High-level architecture
  - ControlNet Block
  - Architecture
- Experiments
  - Examples
  - Analysis
  - Why does it work?
  - Comparison to state of the art

Examples

ControlNet experiments
Condition: Scribble & Pose

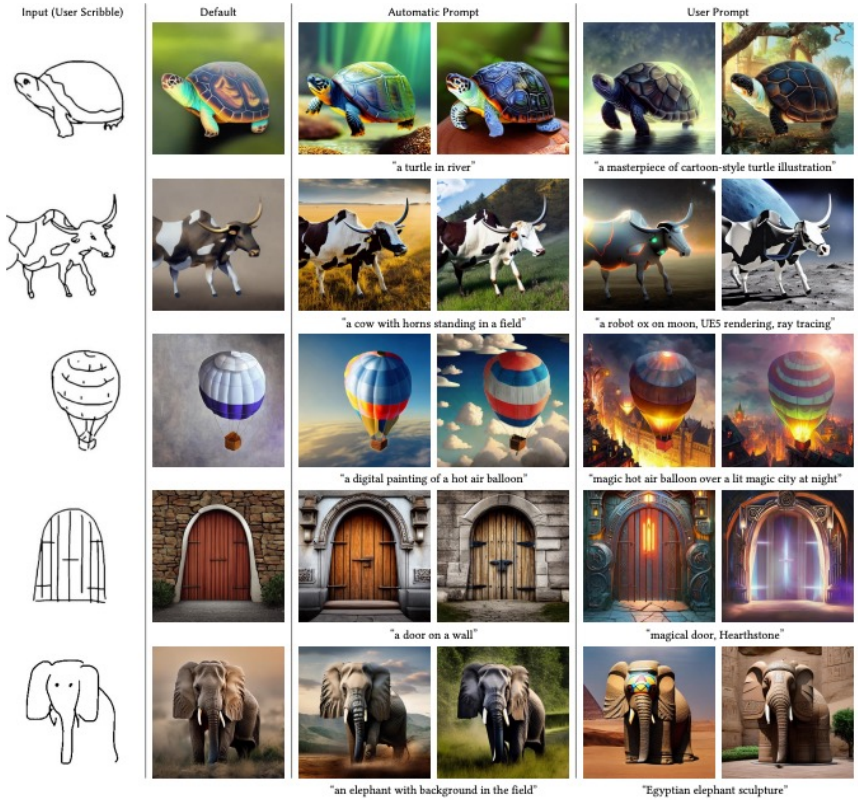


Figure 10: Controlling Stable Diffusion with Human scribbles. The "automatic prompts" are generated by BLIP based on the default result images without using user prompts. These scribbles are from [19].

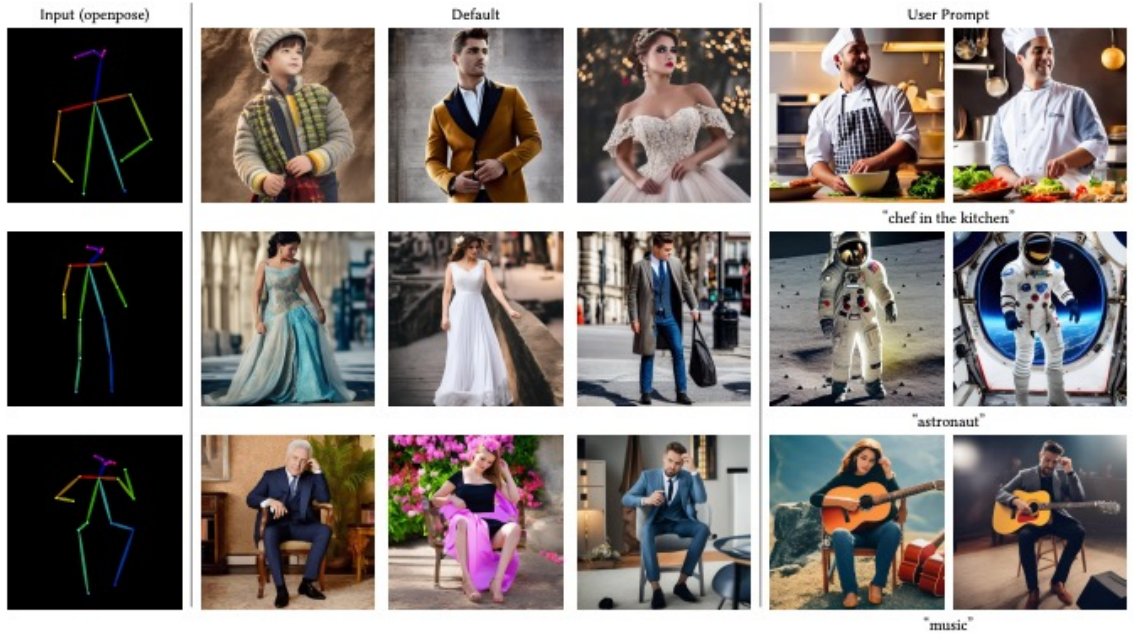
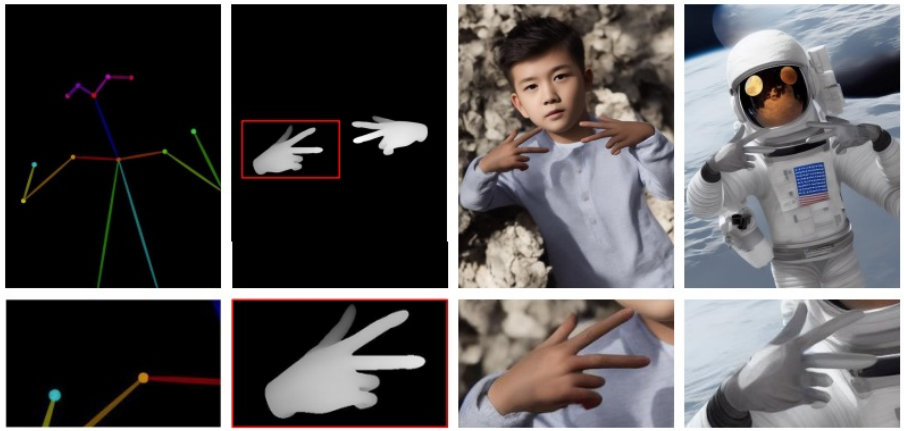


Figure 13: Controlling Stable Diffusion with Openpose. See also the Appendix for source images for Openpose pose detection.

ControlNet experiments
Multiple conditions

- Multiple conditioning images (here Canny edges and pose) to a single instance of Stable Diffusion:
  - Directly add the outputs of the corresponding ControlNets to Stable Diffusion
  - No extra weighting or linear interpolation is necessary for such composition



Multiple condition (pose&depth) "boy" "astronaut"

Figure 6: Composition of multiple conditions. We present the application to use depth and pose simultaneously.

Analysis

ControlNet Analysis

Sudden convergence phenomenon

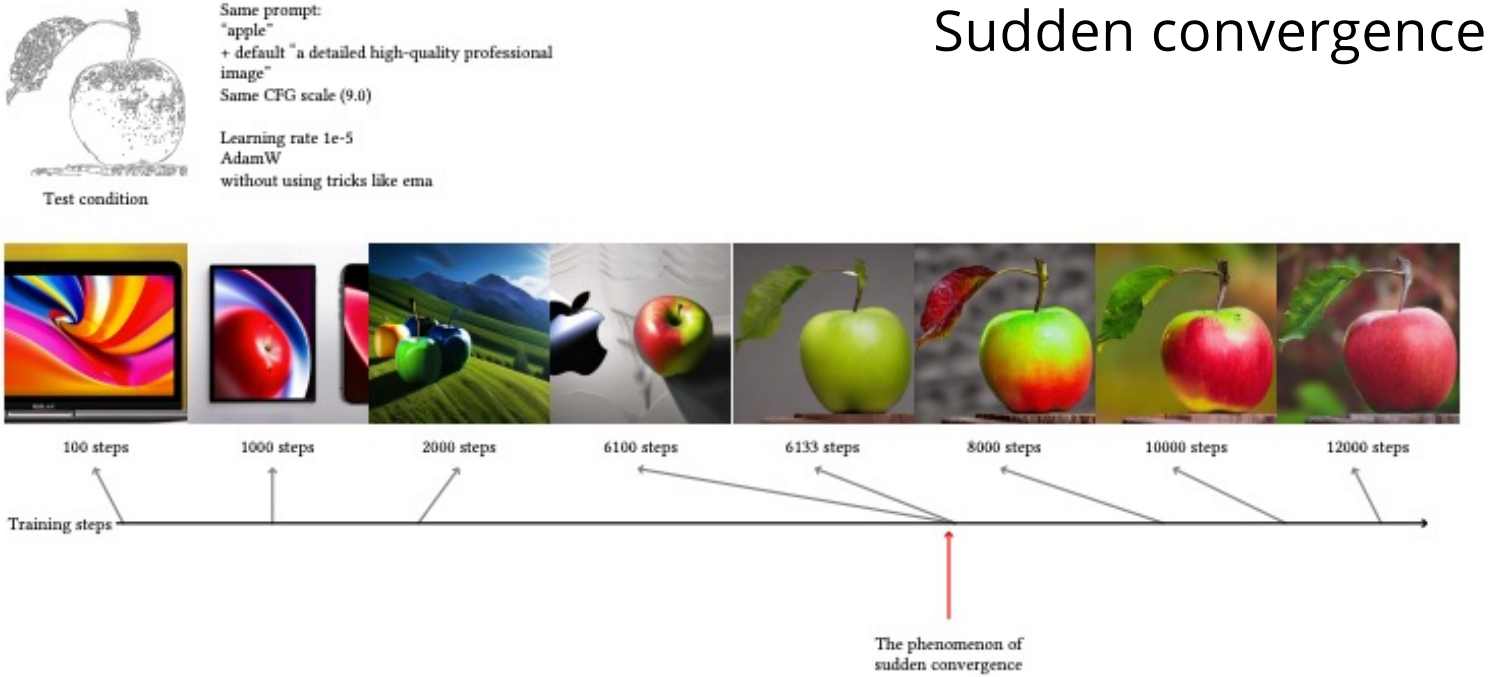


Figure 21: The sudden converge phenomenon. Because we use zero convolutions, the neural network always predicts high-quality images during the entire training. At a certain point in training steps, the model suddenly learns to adapt to the input conditions. We call this "sudden converge phenomenon".

ControlNet Analysis

Sudden convergence phenomenon

Figure 22: Training on different dataset sizes. We show the Canny-edge-based ControlNet trained on different experimental settings with various dataset size.

ControlNet Analysis

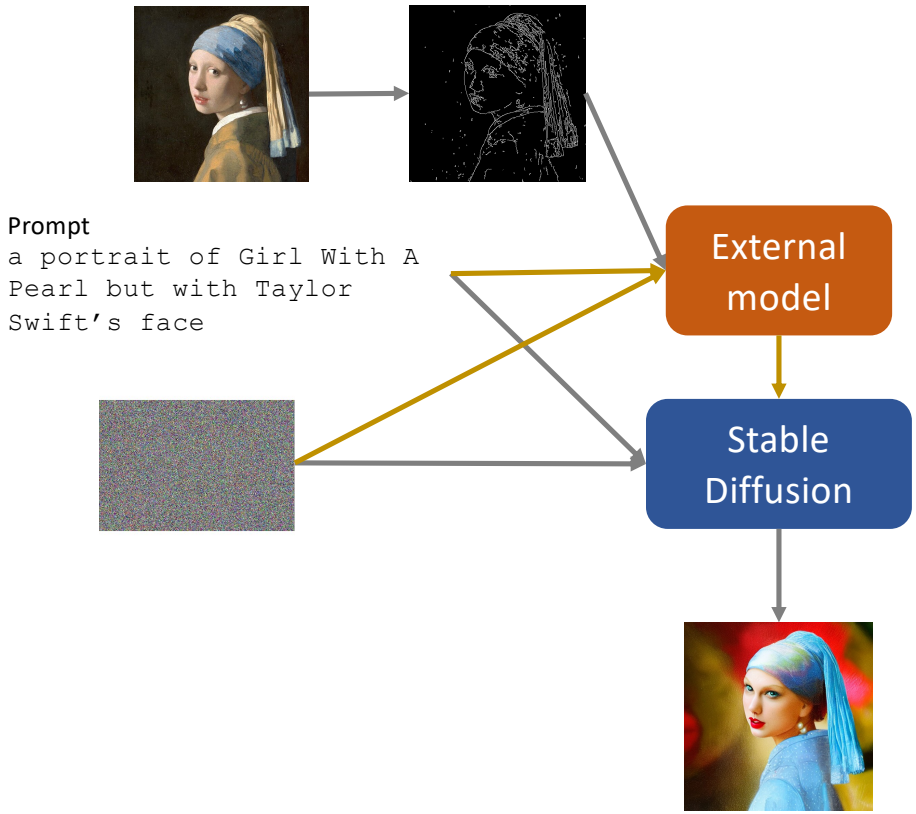
Limitation:

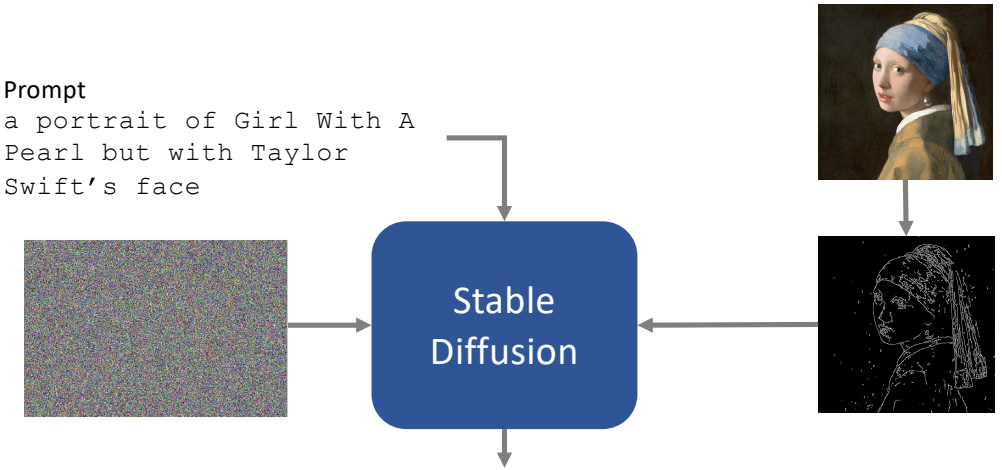
wrong semantic interpretation
→ difficulty in generation

Figure 28: Limitation. When the semantic of input image is mistakenly recognized, the negative effects seem difficult to be eliminated, even if a strong prompt is provided.

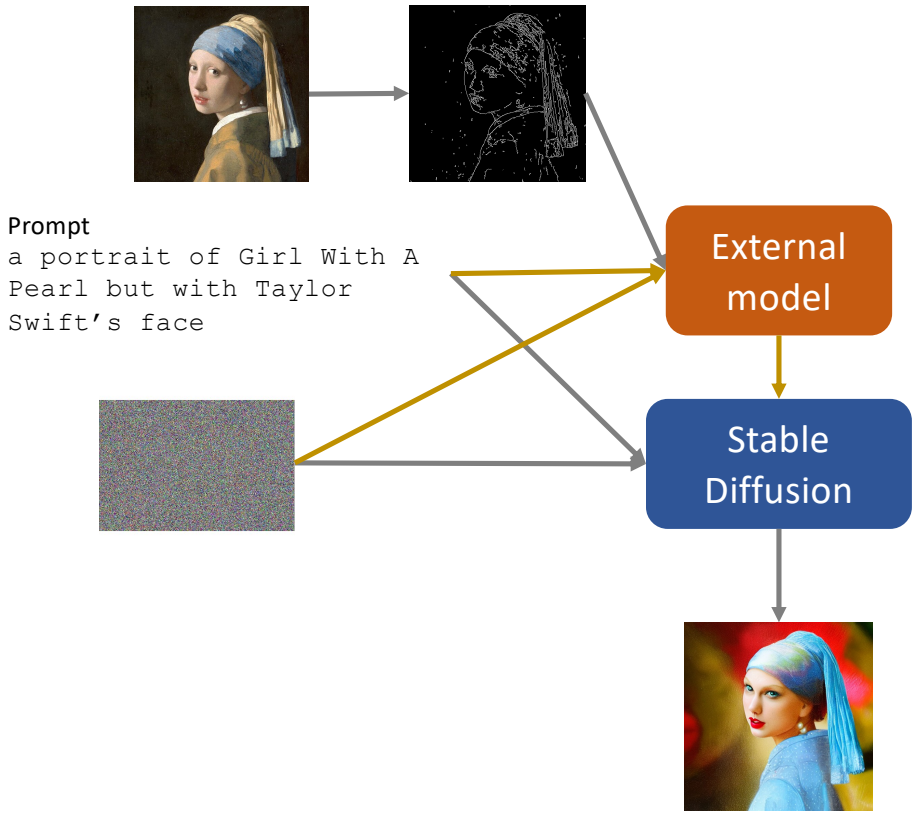
Why does it work?

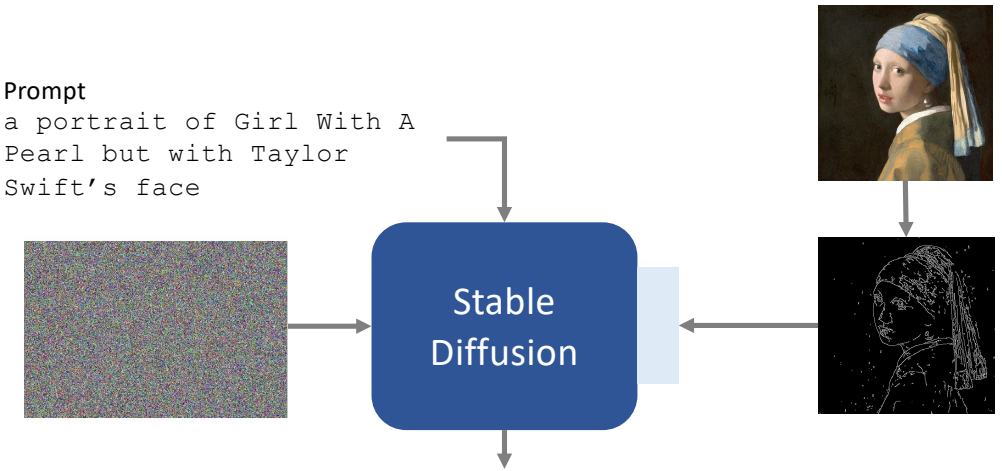
Why does it work?





Why does it work?





Comparison to the state of the art

ControlNet
Comparison to state of the art

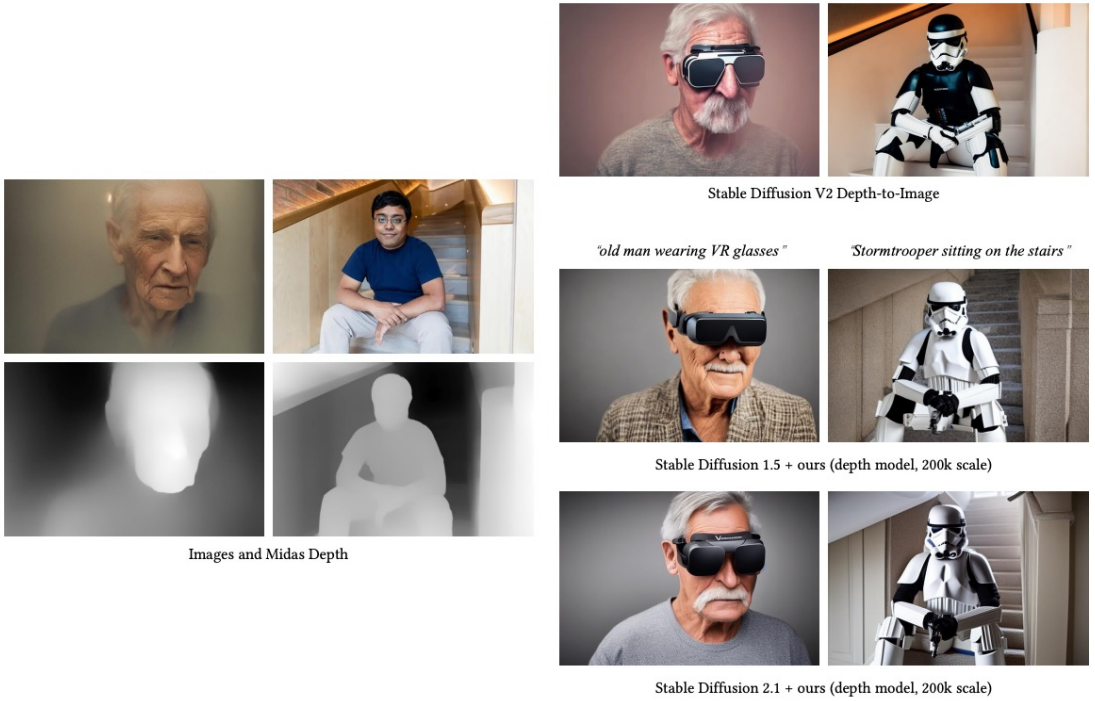


Figure 7: Comparison of Depth-based ControlNet and Stable Diffusion V2 Depth-to-Image. Note that in this experiment, the Depth-based ControlNet is trained at a relatively small scale to test minimal required computation resources. We also provide relatively stronger models that are trained at relatively large scales.

>2000 A100 GPU hours

12M training images

<5 days (<120 hours)
On a consumer GPU: NVIDIA RTX 3090Ti

200k training samples

ControlNet
Comparison to state of the art

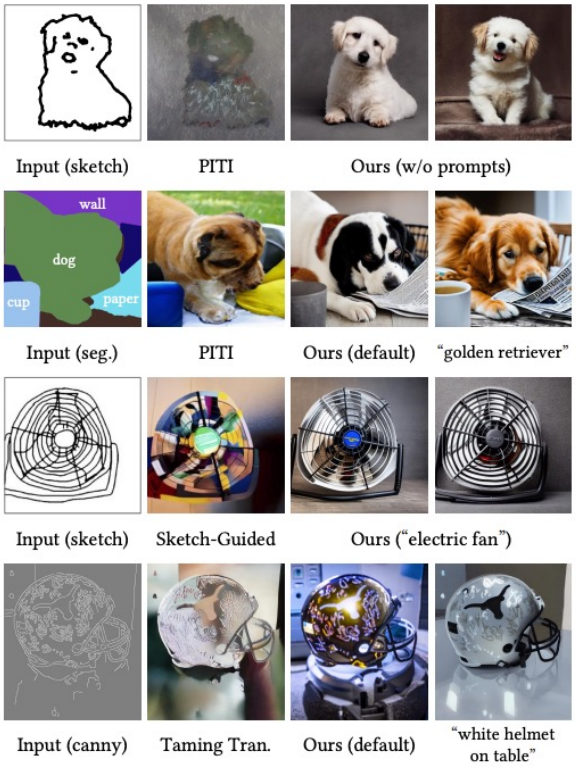


Figure 9: Comparison to previous methods.We present the qualitative comparisons to PITI [89], Sketch-Guided Diffusion [88], and Taming Transformers [19].

ControlNet
Comparison to state of the art

ADE20K (GT)	VQGAN [19]	LDM [72]	PITI [89]	ControlNet-lite	ControlNet
0.58 ± 0.10	0.21 ± 0.15	0.31 ± 0.09	0.26 ± 0.16	0.32 ± 0.12	$\mathbf{0.35} \pm \mathbf{0.14}$

Table 2: Evaluation of semantic segmentation label reconstruction (ADE20K) with Intersection over Union (IoU ↑).

Method	FID ↓	CLIP-score ↑	CLIP-aes. ↑
Stable Diffusion	6.09	0.26	6.32
VQGAN [19](seg.)*	26.28	0.17	5.14
LDM [72](seg.)*	25.35	0.18	5.15
PITI [89](seg.)	19.74	0.20	5.77
ControlNet-lite	17.92	0.26	6.30
ControlNet	15.27	0.26	6.31

Evaluation for image generation conditioned by semantic segmentation. We report FID, CLIP text-image score, and CLIP aesthetic scores for our method and other baselines. We also report the performance of Stable Diffusion without segmentation conditions. Methods marked with "*" are trained from scratch.

ControlNet
User study

- Sample 20 unseen hand-drawn sketches, and then assign each sketch to 5 methods:
  1. PITI
  2. Sketch-Guided Diffusion (SGD) with default edge-guidance scale (β = 1.6),
  3. SGD with relatively high edge-guidance scale (β = 3.2)
  4. ControlNet-lite
  5. ControlNet
- 12 users rank these 20 groups of 5 results
  1. quality of displayed images
  2. fidelity to the sketch
- Obtain
  - 100 rankings for result quality
  - 100 for condition fidelity
- Metric: Average Human Ranking (AHR): users rank each result on a scale of 1 to 5 (lower is worse)

Method	Result Quality ↑	Condition Fidelity ↑
PITI [89](sketch)	1.10 ± 0.05	1.02 ± 0.01
Sketch-Guided [88] ($\beta = 1.6$)	3.21 ± 0.62	2.31 ± 0.57
Sketch-Guided [88] ($\beta = 3.2$)	2.52 ± 0.44	3.28 ± 0.72
ControlNet-lite	3.93 ± 0.59	4.09 ± 0.46
ControlNet	$\mathbf{4.22 \pm 0.43}$	$\mathbf{4.28 \pm 0.45}$

Table 1: Average User Ranking (AUR) of result quality and condition fidelity. We report the user preference ranking (1 to 5 indicates worst to best) of different methods.

Conclusions

- ControlNet: neural network structure that learns conditional control for large pretrained text-to-image diffusion models
- Reuses the large-scale pretrained layers of source models to build a deep and strong encoder to learn specific conditions
- The original model and trainable copy are connected via “zero convolution” layers that eliminate harmful noise during training
- Effectively control Stable Diffusion with single or multiple conditions, with or without prompts
- ControlNet structure: applicable to wider range of conditions & facilitate relevant applications

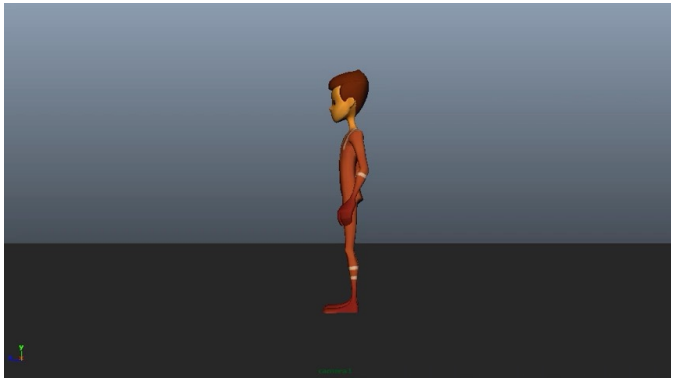
But now Sora is here!

Sora, OpenAI, ElevenLabs

What's next?

Future work: Multimodal Self-supervision

- Inherit motion characteristics
  - Audio
  - Speech
  - Language
  - 3D dynamics
  - Temporal correlation of frames



Running

High-five

Future work: Text-to-Video generation

A confused grizzly bear in a calculus class

A golden retriever eating ice cream on a beautiful tropical beach at sunset, high resolution

A panda playing on a swing set

Audio to visual content generation without passing through text?

Future work:
Memorable Multimodal genAI

- Memorability: **Story, Emotion, Place**
  - Explore this in existing visual data
  - Check how modifying visual content changes this
  - Generate visual content accordingly

Future work: Towards General AI

Thank you

vicky.kalogeiton@polytechnique.edu