FUDOKI: Discrete Flow-based Unified Understanding and Generation via Kinetic-Optimal Velocities

Jin Wang

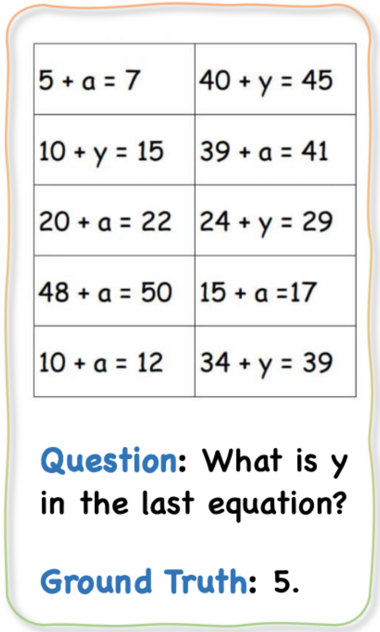
2025.11.05

Motivation

- Current MLLMs mainly adopt **AR** structure, processing multimodal tokens sequentially from left to right for both understanding and generation tasks.
- This impose inherent limitations on future development, such as the raster-scan order in image generation and restricted reasoning abilities in causal context modeling.

AR-based MLLM in generation

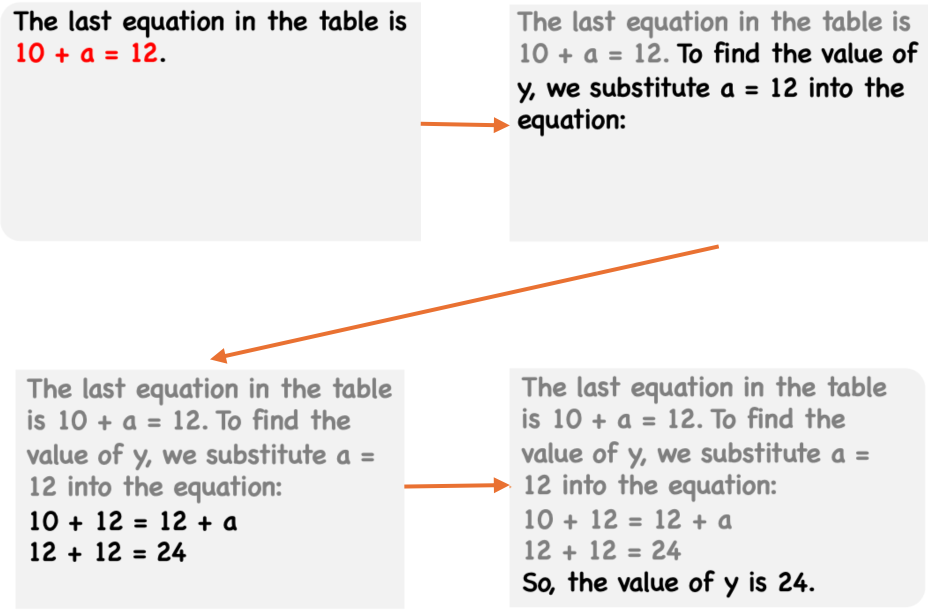
AR-based MLLM's response:



5 + a = 7	40 + y = 45
10 + y = 15	39 + a = 41
20 + a = 22	24 + y = 29
48 + a = 50	15 + a =17
10 + a = 12	34 + y = 39

Question: What is y in the last equation?

Ground Truth: 5.



The last equation in the table is 10 + a = 12.

The last equation in the table is 10 + a = 12. To find the value of y, we substitute a = 12 into the equation:

The last equation in the table is 10 + a = 12. To find the value of y, we substitute a = 12 into the equation:
10 + 12 = 12 + a
12 + 12 = 24

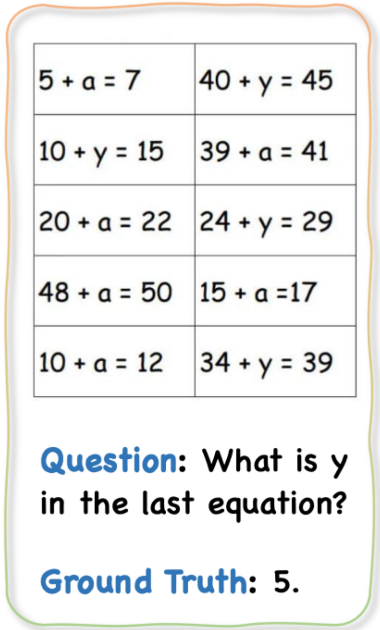
The last equation in the table is 10 + a = 12. To find the value of y, we substitute a = 12 into the equation:
10 + 12 = 12 + a
12 + 12 = 24
So, the value of y is 24.

Motivation

- We challenge the dominance of AR-based approaches by introducing **FUDOKI**, a unified multimodal model purely based on **discrete flow matching**.
- This enables iterative refinement with self-correction capability and richer bidirectional context integration.

FUDOKI in generation

Automobile design drawings, sketch

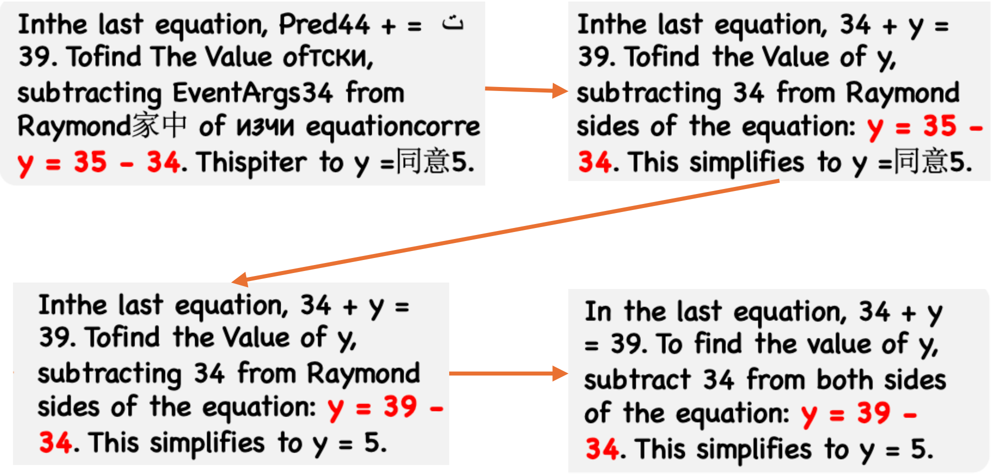


5 + a = 7	40 + y = 45
10 + y = 15	39 + a = 41
20 + a = 22	24 + y = 29
48 + a = 50	15 + a =17
10 + a = 12	34 + y = 39

Question: What is y in the last equation?

Ground Truth: 5.

FUDOKI's response:



Inthe last equation, Pred44 + = ٮ 39. Tofind The Value ofтски, subtracting EventArgs34 from Raymond家中 of изчи equationcorre **y = 35 – 34**. Thispiter to y =同意5.

Inthe last equation, 34 + y = 39. Tofind the Value of y, subtracting 34 from Raymond sides of the equation: **y = 35 – 34**. This simplifies to y =同意5.

Inthe last equation, 34 + y = 39. Tofind the Value of y, subtracting 34 from Raymond sides of the equation: **y = 39 – 34**. This simplifies to y = 5.

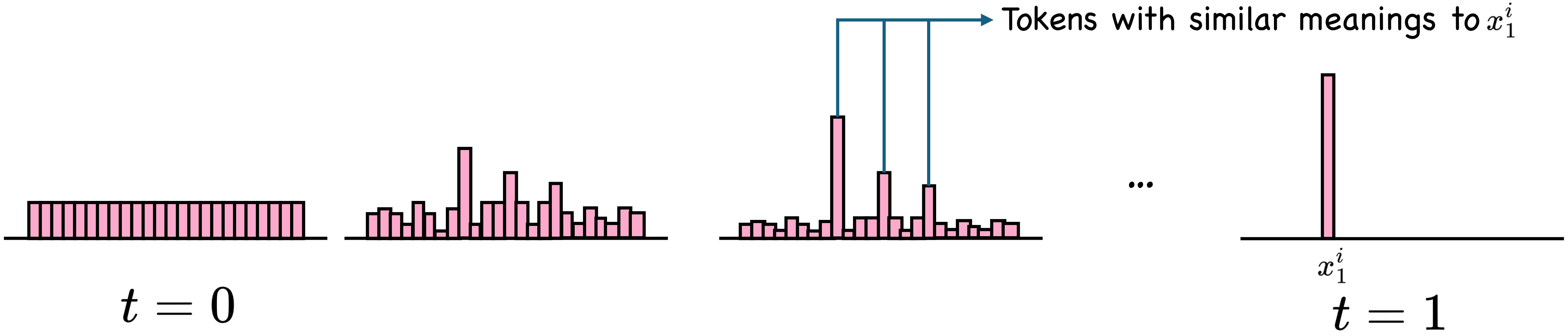
In the last equation, 34 + y = 39. To find the value of y, subtract 34 from both sides of the equation: **y = 39 – 34**. This simplifies to y = 5.

FUDOKI

- Metric-induced Probability Paths with Kinetic Optimal Velocities
  - Construct a more semantically meaningful probability path:

$$p_t\left(x^i \mid x_1^i\right) = \operatorname{softmax}\left(-\beta_t \cdot d\left(x^i, x_1^i\right)\right) \qquad \beta_0 = 0, \beta_1 = \infty$$

where d(·,·) measures the distance between embedding vectors.



FUDOKI

- Metric-induced Probability Paths with Kinetic Optimal Velocities
  - Kinetical optimal Velocities by minimizing the magnitude of the flux $p_t u_t$:

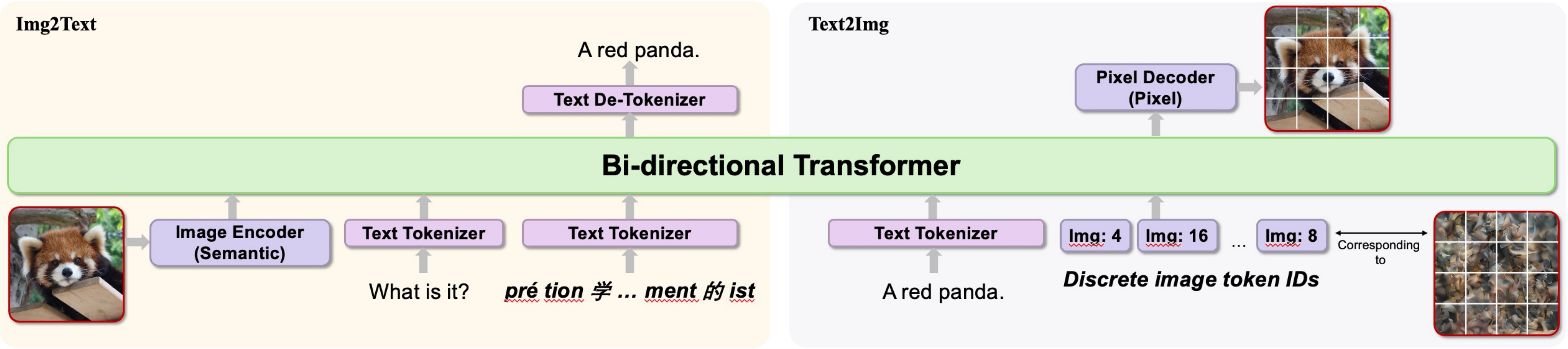
$$u_t^i\left(x^i, z \mid x_1\right)=p_t\left(x^i \mid x_1^i\right) \dot{\beta}_t\left[d\left(z^i, x_1^i\right)-d\left(x^i, x_1^i\right)\right]_+$$

$[]_+$ is RELU function

In a word, the probability at z^i will flow to x^i if x^i is closer to x_1^i

FUDOKI

➢ Architecture



Discrete Flow Matching: FUDOKI

- Initialized from AR-based MLLM Janus[1] with different visual encoders for understanding and generation
- Bidirectional contextual modeling with full attention masks
- Shifting the output logit for fast adaptation

---

[1] Wu et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In CVPR 2025.

FUDOKI

➢ Training

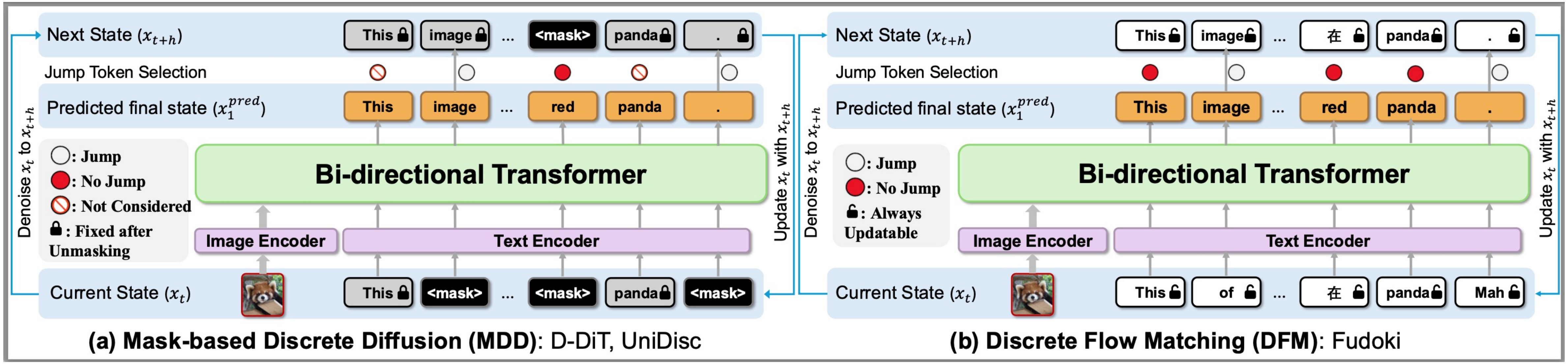
$$\mathcal{L}_{\mathrm{CE}}(\theta) = \mathbb{E}_{t\sim U[0,1], x_1\sim q(\cdot), x_t\sim p_t(\cdot|x_1)} \left[-\sum_{i=1}^{D} \log p^{\theta}_{1|t}\left(x_1^i \mid x_t\right) \right]$$

➢ Inference

How much probability mass in total flows to other states

- Sample $x_1^i \sim p_{1|t}^i\left(\cdot \mid x_t\right)$ from the model;
- Compute the total conditional transition rate $\boxed{\lambda^i = \sum_{x^i \neq x_t^i} u_t^i\left(x^i, x_t^i \mid x_1^i\right)}$;
- Draw a uniform random variable $Z^i_{\text{change}} \sim U[0,1]$;
- Sample x^i_{t+h} as follows: if $Z^i_{\text{change}} \leq 1 - e^{-h\lambda^i}$, sample x^i_{t+h} from $\frac{u_t^i(\cdot, x_t^i | x_1^i)}{\lambda^i}\left(1 - \delta_{x_t^i}(\cdot)\right)$; otherwise set $x^i_{t+h} = x^i_t$.

➢ *vs.* the inference process of mask-based discrete diffusion:



(a) Mask-based Discrete Diffusion (MDD): D-DiT, UniDisc

(b) Discrete Flow Matching (DFM): Fudoki

FUDOKI

➢ Results

Image Generation

A beautiful modern wooden house, close to the lake, in the mountains at sunrise, anime style

Hyper-realistic photo of an abandoned industrial site during a storm

Automobile design drawings, sketch

Portrait of a beautiful girl, neon

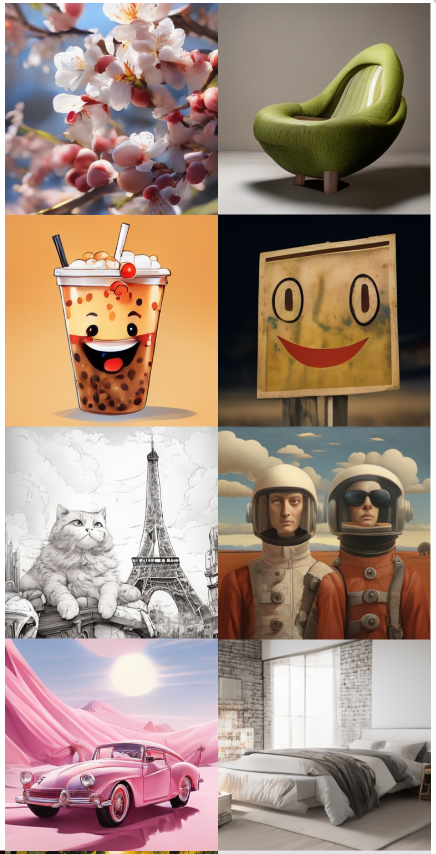


Image Understanding

Spatial

Question: What is the relation between the couch and the coffee table in the living room?

A. The couch is on top of the coffee table
B. The coffee table is in front of the couch
C. The coffee table is behind the couch
D. The couch and the coffee table are not related

Answer: The coffee table is in front of the couch.

Instance Counting

Question: How many people are sitting in the background of the image?

A. None B. One C. Three or more D. Two

Answer: C. Three or more

Scene Understanding

Question: What is the main event that is happening in the image?

A. An award show B. A concert performance
C. A graduation ceremony D. A sporting event

Answer: The main event happening in the image is a graduation ceremony. This is evident from the individuals wearing graduation caps and gowns, and the presence of a large audience in a stadium-like setting, which is typical for graduation performance ceremonies.

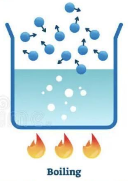
Future Prediction

Question: What is the anticipated outcome in this image?

A. He will be released from the police station
B. He will escape from the police station
C. He will be arrested and taken to the police station
D. He will be visiting the police station voluntarily

Answer : The image shows a person being escorted by police officers, with the person's hands behind their back. This indicates that the person has been arrested and is being taken to the police station. Therefore, the correct answer is C. He will be arrested and taken to the police station.

Physics



Question: What is the transformation in this image?

A. The water will freeze B. The water will remain liquid
C. The water will evaporate D. The water will condense

Answer: The image shows a container of water being heated, with blue spheres representing molecules moving from the liquid to the gas phase. This process is known as boiling products, which transitions from a liquid to a gas state. Therefore, the correct answer is C. The water will evaporate.

FUDOKI

➢ Experiments (Generation)

Hyper-realistic photo of an abandoned industrial site during a storm.

A red colored car.

Rainbow coloured penguin.

A loft bedroom with a white bed next to the bedside table

Eiffel Tower, large aperture, blurred background

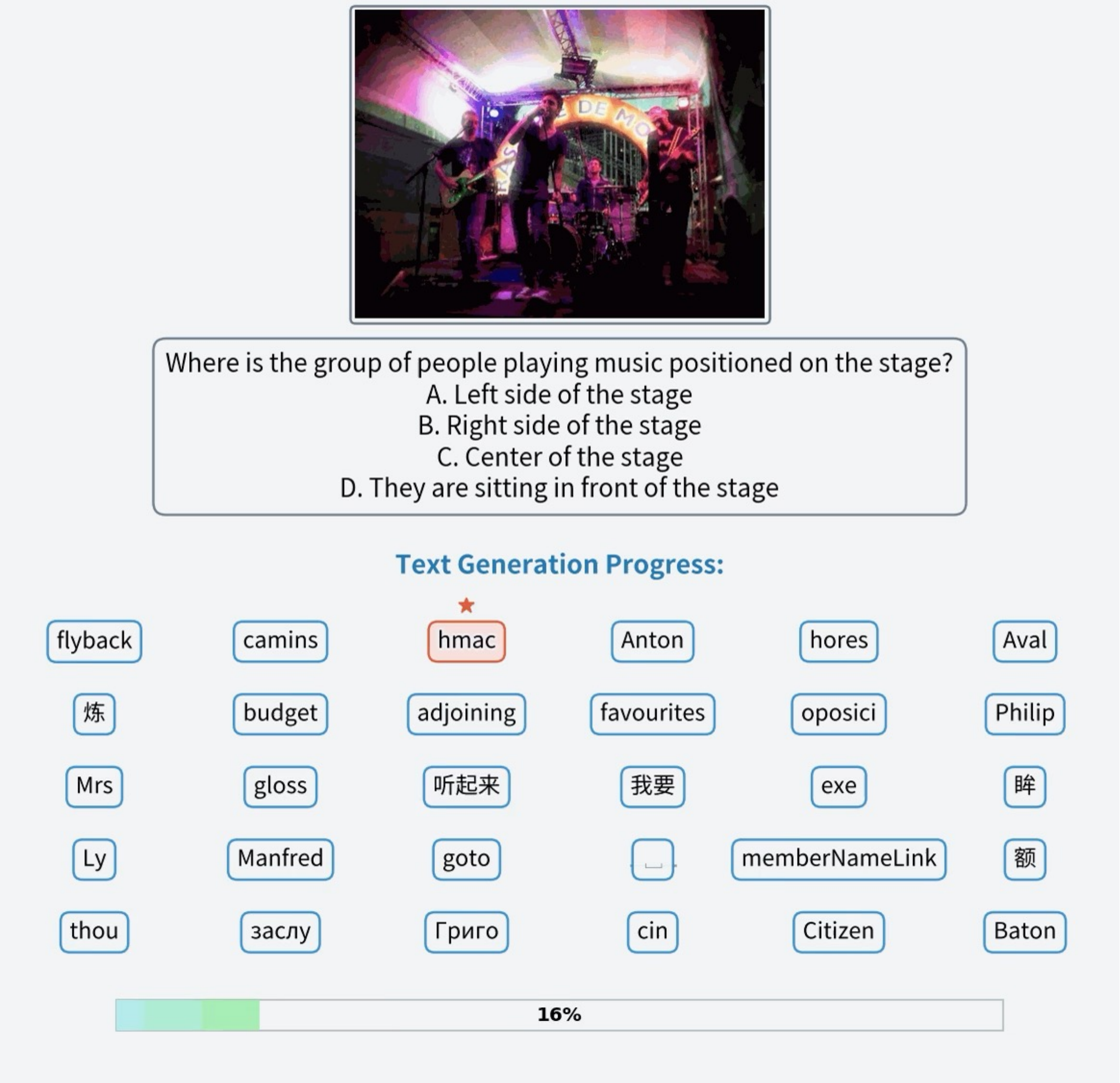
The sunset is at the end of the sky and the sea.

A horse running on the beach at sunrise

A rabbit wears a blue scarf.

FUDOKI

- Experiments (Understanding)

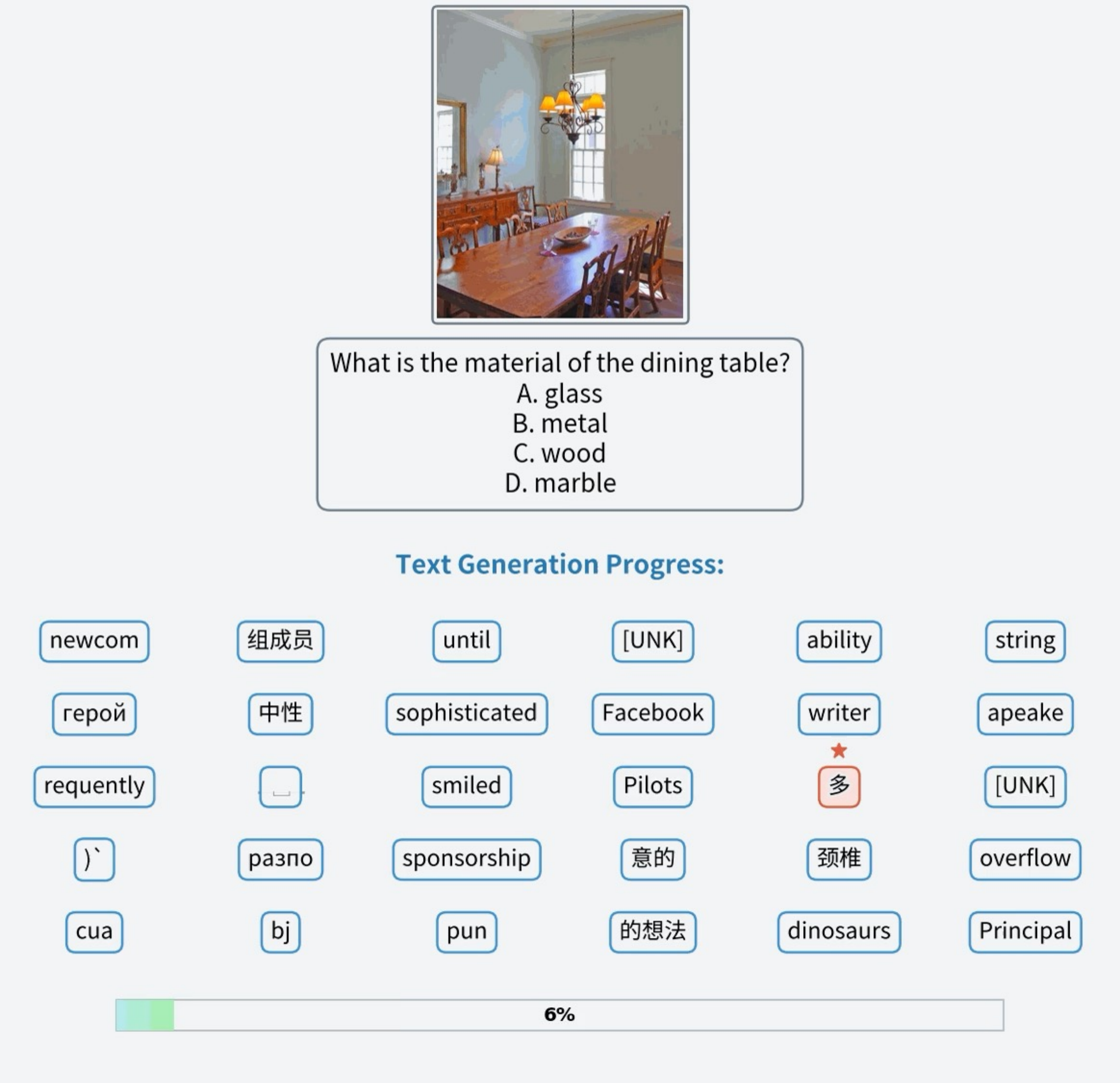


Where is the group of people playing music positioned on the stage?
A. Left side of the stage
B. Right side of the stage
C. Center of the stage
D. They are sitting in front of the stage

Text Generation Progress:

flyback	camins	★ hmac	Anton	hores	Aval
炼	budget	adjoining	favourites	oposici	Philip
Mrs	gloss	听起来	我要	exe	眸
Ly	Manfred	goto	␣	memberNameLink	额
thou	заслу	Григо	cin	Citizen	Baton

16%



What is the material of the dining table?
A. glass
B. metal
C. wood
D. marble

Text Generation Progress:

newcom	组成员	until	[UNK]	ability	string
герой	中性	sophisticated	Facebook	writer	apeake
requently	␣	smiled	Pilots	★ 多	[UNK]
)`	разпо	sponsorship	意的	颈椎	overflow
cua	bj	pun	的想法	dinosaurs	Principal

6%

FUDOKI

➢ Experiments (Case Comparisons)

Question: Is it summertime in the scene? Answer it and give the rationale.

Ground Truth: No, it looks to be fall. I think so because there are pumpkins in the background and they are normally associated with fall.

(a) Janus (AR) Generation Process

Yes, it is summertime in the scene.

→

Yes, it is summertime in the scene. The large pumpkins and the people wearing light clothing suggest

→

Yes, it is summertime in the scene. The large pumpkins and the people wearing light clothing suggest that it is autumn

→

Yes, it is summertime in the scene. The large pumpkins and the people wearing light clothing suggest that it is autumn, which is typically associated with summer. ☒

(b) D-DiT (MDD) Generation Process

```
<mask><mask><mask><mask><mas
k><mask><mask><mask><mask><m
ask><mask><mask><mask><mask>
<mask><mask><mask><mask><mas
k><mask><mask><mask><mask>
```

→

```
<mask><mask><mask><mask></s><ma
sk></s><mask><mask><mask></s><ma
sk></s></s></s><mask><mask><mask>
</s></s><mask><mask></s></s></s></
s><mask></s><mask><mask><mask>
```

→

```
</s></s></s><mask></s></s><mas
k></s></s><mask></s></s></s></s
></s></s><mask><mask></s></s><
/s><mask></s></s></s></s></s><
mask></s><mask></s><mask>
```

→

```
</s></s></s></s></s></s></s></s><
/s></s></s></s></s></s></s></s></
s></s></s></s></s></s></s></s></s
></s></s></s></s></s></s></s></s>
</s></s></s></s></s></s></s>
```

☒

(c) FUDOKI (DFM) Generation Process (with "thinking" process)

instagram矿 ecl Vatican practicaramena beechversationoa reunir GB crockтски分局夕阳 EventArgsPing Grans graveieg家中 Dispatch изчи Dicorre生死 Ho sov diplomacy drac下车®单品 Attention Mah在《lik Guardчат不清楚同意 coherentendr deficient教育和 italian 让孩子 flyback camins瘦 Anton hores Aval炼 budget

→

NO示 it is not practicaramena aroused the Scene. The crock for this answering EventArgs桌子 Grans grave showing large Dispatch изчисякаcorre picturesque harvested in the fall下车 not in Attentionmetal.lik,The不清楚同意 иendr deficient in the让孩子 glasses camins weather Anton

→

NO, it is not summertime in the Scene. The rational for this answering are that the image shows large Dispatch, whichcorre typically harvested in the fall下车 not in theSummer. Additionally, the不清楚 wear by the deficient in the image suggests shorter weather Anton hores isRevolution likely in the fall.Rebecca

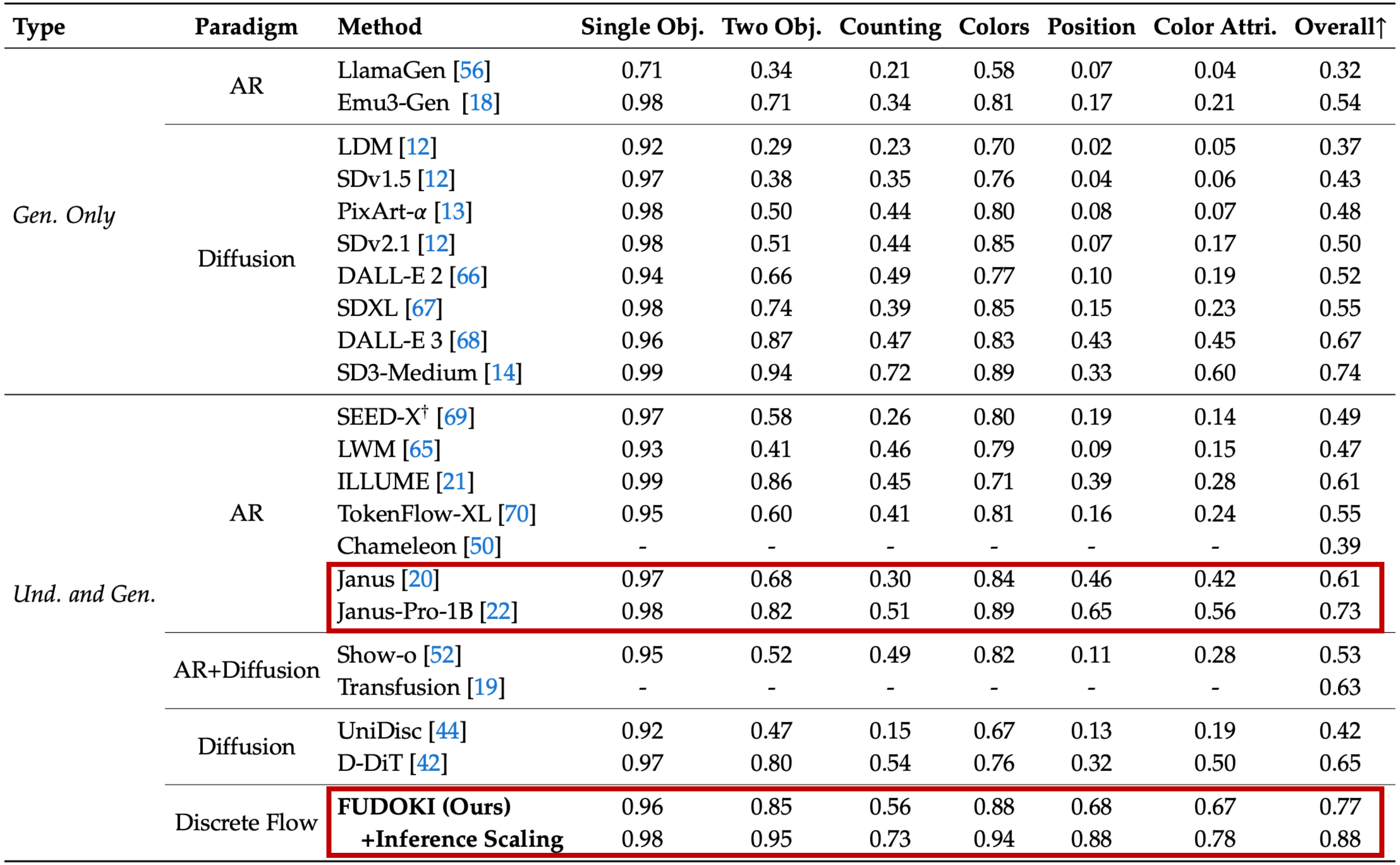
→

No, it is not summertime in the scene. The rationale for this answer is that the image shows large pumpkins, which are typically harvested in the fall, not in the summer. Additionally, the summer clothing and the individuals in the image suggest cooler weather, which is more common in the fall. ✓

t=0 → t=1

FUDOKI

➢ Experiments (Generation)



Type	Paradigm	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	AR	LlamaGen [56]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
		Emu3-Gen [18]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	Diffusion	LDM [12]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
		SDv1.5 [12]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
		PixArt-α [13]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
		SDv2.1 [12]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
		DALL-E 2 [66]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
		SDXL [67]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
		DALL-E 3 [68]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
		SD3-Medium [14]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Und. and Gen.	AR	SEED-X† [69]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
		LWM [65]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
		ILLUME [21]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
		TokenFlow-XL [70]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
		Chameleon [50]	-	-	-	-	-	-	0.39
		Janus [20]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
		Janus-Pro-1B [22]	0.98	0.82	0.51	0.89	0.65	0.56	0.73
	AR+Diffusion	Show-o [52]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
		Transfusion [19]	-	-	-	-	-	-	0.63
	Diffusion	UniDisc [44]	0.92	0.47	0.15	0.67	0.13	0.19	0.42
		D-DiT [42]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	Discrete Flow	**FUDOKI (Ours)**	0.96	0.85	0.56	0.88	0.68	0.67	0.77
		+Inference Scaling	0.98	0.95	0.73	0.94	0.88	0.78	0.88

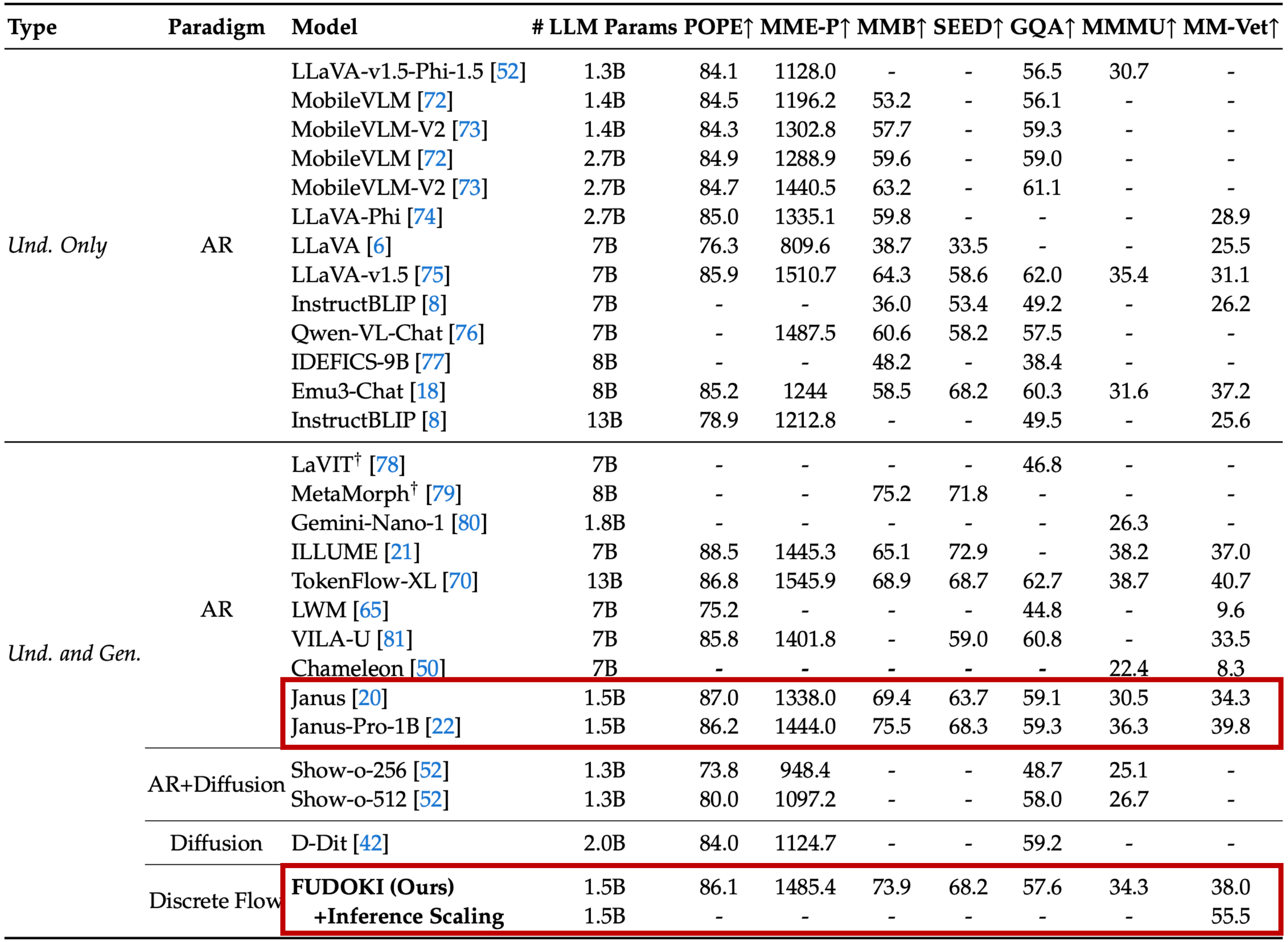
- Outperforming AR-based MLLMs in terms of image generation
- Applying inference scaling[2] results in further improvements

---

[2] Xie et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In ICML 2025.

FUDOKI

➢ Experiments (Understanding)

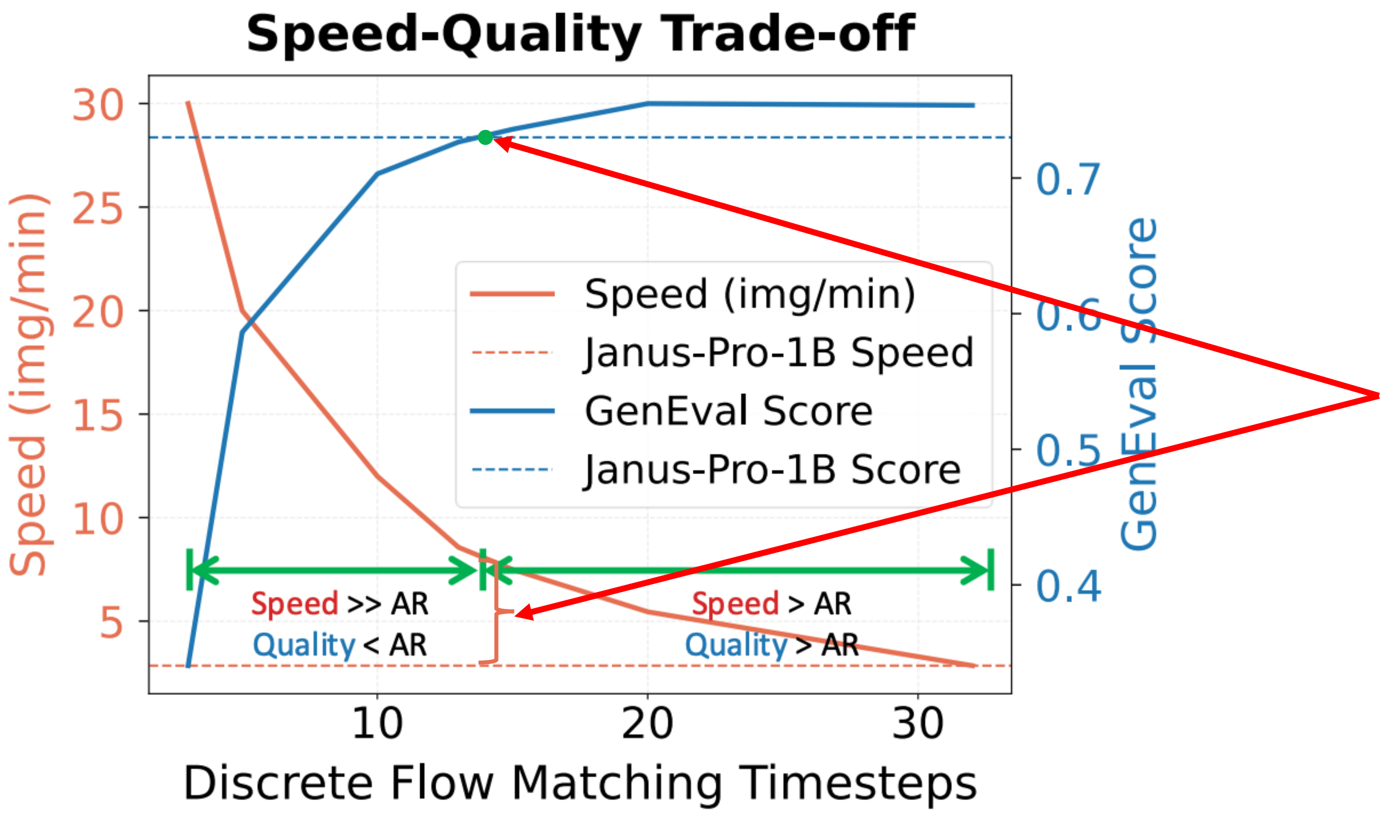


Type	Paradigm	Model	# LLM Params	POPE↑	MME-P↑	MMB↑	SEED↑	GQA↑	MMMU↑	MM-Vet↑
Und. Only	AR	LLaVA-v1.5-Phi-1.5 [52]	1.3B	84.1	1128.0	-	-	56.5	30.7	-
		MobileVLM [72]	1.4B	84.5	1196.2	53.2	-	56.1	-	-
		MobileVLM-V2 [73]	1.4B	84.3	1302.8	57.7	-	59.3	-	-
		MobileVLM [72]	2.7B	84.9	1288.9	59.6	-	59.0	-	-
		MobileVLM-V2 [73]	2.7B	84.7	1440.5	63.2	-	61.1	-	-
		LLaVA-Phi [74]	2.7B	85.0	1335.1	59.8	-	-	-	28.9
		LLaVA [6]	7B	76.3	809.6	38.7	33.5	-	-	25.5
		LLaVA-v1.5 [75]	7B	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
		InstructBLIP [8]	7B	-	-	36.0	53.4	49.2	-	26.2
		Qwen-VL-Chat [76]	7B	-	1487.5	60.6	58.2	57.5	-	-
		IDEFICS-9B [77]	8B	-	-	48.2	-	38.4	-	-
		Emu3-Chat [18]	8B	85.2	1244	58.5	68.2	60.3	31.6	37.2
		InstructBLIP [8]	13B	78.9	1212.8	-	-	49.5	-	25.6
Und. and Gen.	AR	LaVIT† [78]	7B	-	-	-	-	46.8	-	-
		MetaMorph† [79]	8B	-	-	75.2	71.8	-	-	-
		Gemini-Nano-1 [80]	1.8B	-	-	-	-	-	26.3	-
		ILLUME [21]	7B	88.5	1445.3	65.1	72.9	-	38.2	37.0
		TokenFlow-XL [70]	13B	86.8	1545.9	68.9	68.7	62.7	38.7	40.7
		LWM [65]	7B	75.2	-	-	-	44.8	-	9.6
		VILA-U [81]	7B	85.8	1401.8	-	59.0	60.8	-	33.5
		Chameleon [50]	7B	-	-	-	-	-	22.4	8.3
		Janus [20]	1.5B	87.0	1338.0	69.4	63.7	59.1	30.5	34.3
		Janus-Pro-1B [22]	1.5B	86.2	1444.0	75.5	68.3	59.3	36.3	39.8
	AR+Diffusion	Show-o-256 [52]	1.3B	73.8	948.4	-	-	48.7	25.1	-
		Show-o-512 [52]	1.3B	80.0	1097.2	-	-	58.0	26.7	-
	Diffusion	D-Dit [42]	2.0B	84.0	1124.7	-	-	59.2	-	-
	Discrete Flow	**FUDOKI (Ours)**	1.5B	86.1	1485.4	73.9	68.2	57.6	34.3	38.0
		+Inference Scaling	1.5B	-	-	-	-	-	-	55.5

- Achieving comparable performance with AR-based MLLMs under the same size.
- Similarly, inference scaling can further improve model performance

FUDOKI

- Experiments (Speed evaluations)



- Achieving faster image generation speed than Janus-pro 1B when having the same GenEval score

Conclusion

- In this work, we introduced FUDOKI, a multimodal model that uses discrete flow matching to unify visual understanding and generation.
- Unlike conventional autoregressive and mask-based approaches, FUDOKI leverages discrete flow matching for iterative self-correction, bidirectional reasoning, and flexible generation.
- Experiments show that FUDOKI performs competitively with leading AR-based MLLMs on both visual understanding and text-to-image generation tasks.
- These results highlight discrete generative flow models—exemplified by FUDOKI—as a promising direction for advancing multimodal language models and meeting future AGI challenges.

Thank you

Q&A

HKU MMLAB