Noisy Channel に基づく生成確率による画像生成評価

林 和樹　尾崎 慎太郎　神野 倫行　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学 (NAIST)
hayashi.kazuki.hl4@is.naist.jp
{ozaki.shintaro.ou6, jinno.tomoyuki.jx3}@naist.ac.jp
{kamigaito.h, taro}@is.naist.jp

概要

近年の画像生成（T2I）モデルの進展により，生成画像の表現力や多様性は大きく向上している一方で，長文や複雑な指示を含む生成では，単一指標で出力を評価することが難しく，既存の評価手法は高度化した生成能力に十分対応できていない．本研究では，生成確率に基づく Noisy Channel により T2I 評価を再定式化し，画像のテキスト整合性と視覚的品質を統一的に捉える確率的評価指標を提案する．提案手法は，LVLM の推論能力を教師強制尤度として用いた整合性評価と，自己回帰型画像生成モデルの尤度による品質評価を組み合わせることで，生成結果間の相対比較に依存せず，各画像を独立に評価できる．検証の結果，提案手法は人手による画像選好と高い整合性を示し，既存のスコアリング手法を一貫して上回る性能を達成した．また，評価観点を切り替えることで，同一の確率的枠組みのもとで多様な人手判断を柔軟に捉えられることを確認した．

1　はじめに

近年，画像生成モデルは急速に発展し，Stable Diffusion に代表される拡散モデルの登場により，生成画像の表現力と多様性が向上するとともに，高解像度で写実的な画像を生成できるようになった [1, 2, 3]．これにより，長文や複数要素を含む設定では，同一のプロンプトに対して複数の妥当な生成結果が成立しうる．これは，画像生成が本質的に単一の正解を持たない高自由度な課題であり，生成結果の評価が一意に定まらないためである．その結果，単一の参照や固定的な基準に基づく評価は困難であり，既存の評価枠組みはこのような画像生成タスクの特性に十分対応できていない．

FID や IS などの分布ベース指標は，生成集合全体の傾向は捉えられても，各画像ごとの品質や整合性を直接評価できず，個々の生成結果に現れる微細な誤りや品質差を十分に反映できない [4, 5]．また，CLIPScore などの埋め込み類似度に基づく指標は，評価を単一のスカラー値で表現するため，長文プロンプトや要素間の関係性・構成的要素，推論を要する整合性の検証が難しい [6]．これを補うために，質問生成と VQA による忠実性評価（TIFA）[7] や，物体検出器を用いた構成能力の検証（GenEval）[8]，人手選好データから学習した報酬モデル（ImageReward / PickScore）[9, 10] などが提案されてきた．さらに近年では，大規模視覚言語モデル（LVLM）を判定器として用い，生成画像とプロンプトの整合性や品質を自然言語で評価させる手法も広く用いられている [11, 12]．しかし，これらの手法は，複雑な評価パイプラインや高い計算コストを要し，評価器固有の偏りやプロンプト設計への感度を避けられない．また，ドメインや設定の違いによってスコアの解釈が変動しやすく，各画像を一貫した絶対尺度で評価することは依然として困難である．

本研究では，参照を仮定しない多様な生成結果を対象として，生成確率に基づく雑音のある通信過程 (Noisy Channel) モデルの観点から画像生成評価を再定式化し，テキスト整合性と画像品質を同一の確率的枠組みで捉える評価指標を提案する．提案手法は，LVLM の教師強制尤度による整合性評価と，自己回帰型画像生成モデル (AR) の尤度による画質評価を組み合わせることで，生成結果間の相対比較に依存せず，各画像を独立に評価できる．さらに，評価用テキストテンプレートを切り替えることで，異なる評価観点を同一枠組みで柔軟に検証できる．検証の結果，提案手法は人手による画像選好と高い整合性を示し，既存のスコアリング手法を一貫して上回る性能を達成した．また，評価観点を切り替えることで，同一の確率的枠組みのもとで，観点の異なる人手判断を柔軟に捉えられることを確認した．

2 提案手法

本研究では，テキスト T に基づいて生成された画像 I を対象とし，画像の生成過程 $P(I \mid T)$ を Noisy Channel として捉えることで評価を行う．具体的には，ベイズの定理に基づき，画像 I の評価を条件付き対数確率 $\log P(T \mid I)$ と事前対数確率 $\log P(I)$ の和として定式化する．

2.1 Noisy Channel に基づく定式化

画像生成を $P(I \mid T)$ とみなすと，ベイズの定理より

$$P(I \mid T) = \frac{P(T \mid I)\, P(I)}{P(T)} \tag{1}$$

となる．同一の T を比較する場合，$\log P(T)$ は定数であるため，Noisy Channel に基づくスコアを

$$\mathcal{S}_\alpha(I \mid T) = \log P(T \mid I) + \alpha \log P(I) \tag{2}$$

と定義する．$\log P(T \mid I)$ は整合性，$\log P(I)$ は画像の自然さを表し，$\alpha \geq 0$ は両者の寄与を制御する．

2.2 整合性の推定：$\log P(T \mid I)$

テキスト列 $T = (w_1, \ldots, w_N)$ と画像 I に対し，デコーダ型の LVLM は，各トークンの条件付き確率 $P(w_i \mid w_{<i}, I)$ を与える．本研究では，与えられた画像とテキストの整合性を尤度として評価することを目的とし，教師強制により尤度を計算する．具体的には，$\log P(T \mid I)$ を

$$\log P(T \mid I) \;\approx\; s_{\text{align}}(I, T) \;\triangleq\; \frac{1}{N} \sum_{i=1}^{N} \log P(w_i \mid w_{<i}, I) \tag{3}$$

とする．ここで N はトークン列の長さであり，$1/N$ による正規化は，トークナイズ長に起因するスケール差を抑えるためである．この値は，画像 I が評価用テキスト T に対して一貫して説明可能であるかを測る指標であり，LVLM の内部的な推論により内容的な不整合を確率として捉えることを目的とする．

2.3 画像の視覚的品質の推定：$\log P(I)$

画像の自然さを確率として評価するため，画像 I をトークン列 $Z = (z_1, \ldots, z_M)$ に変換し，自己回帰型画像生成モデルにより

$$P(I) \equiv P(Z) = \prod_{j=1}^{M} P(z_j \mid z_{<j}) \tag{4}$$

と分解する．ここで Z はピクセル列または離散的な画像コード列を表す．$\log P(I)$ を

$$\log P(I) \;\approx\; s_{\text{img}}(I) \;\triangleq\; \frac{1}{M} \sum_{j=1}^{M} \log P(z_j \mid z_{<j}) \tag{5}$$

とする．ここで M は画像トークン列の長さであり，$1/M$ による正規化は解像度や符号化方式によるスケール差を除去するためである．この項は，画像の自然さを確率として捉える役割を担い，整合性項と合わせることで，内容は整合しているが視覚的に破綻した画像を抑制する．

表 1 視覚的妥当性の評価観点を制御するための評価用テキスト $T^{(c)}$ の例．同一の整合性スコア $s_{\text{align}}(I, T)$ を用い，テキストの差し替えにより評価観点 c を導入する．

評価観点 c	評価用テキスト $T^{(c)}$ の例
視覚的妥当性（全体）	The image is visually plausible and does not contain physically impossible or contradictory elements.
形状・構造の自然さ	Object shapes and proportions appear natural and structurally consistent.
照明・外観の自然さ	Lighting, shadows, and surface appearance are visually consistent and realistic.

2.4 評価用テキストによる観点制御

本手法では，生成時に用いられたプロンプト p とは独立に，特定の評価観点において生成画像が満たすべき性質を記述した評価用テキスト T を用いることで，画像の評価観点を柔軟に切り替えた多様な評価を可能とする．評価観点 c に対しては，評価用テキストのテンプレート群に基づき

$$T^{(c)} = g(p; c) \tag{6}$$

を構成する．ここで $T^{(c)}$ は，観点 c において妥当な画像であれば自然に成立すると考えられる評価基準を文章として表現したものである．表 1 に，視覚的妥当性に関する評価用テキスト $T^{(c)}$ の例を示す．

3 実験設定

3.1 評価タスクと評価方法

評価タスクは，単一のプロンプト p に対して生成された 2 枚の画像 $\{I_1, I_2\}$ の優劣を人手で判定する二者画像選好評価であり，各評価ペアには勝者 I^+ と敗者 I^- が付与されている．評価指標として，提案スコア $\mathcal{S}(I \mid T)$ による二者比較の結果が人手選好

表 2　Text-to-Image Alignment におけるペアワイズ正解率（%）．本観点では，生成に用いたプロンプトを評価用テキスト T とし，Noisy Channel に基づく評価スコア $\mathcal{S}_\alpha(I \mid T) = \log P(T \mid I) + \alpha \log P(I)$ により画像を評価する．また，LVLM のアンサンブルの結果 (Ens:) も併せて示す．アンサンブルは候補モデル集合から 3 モデルを選び（C_3），α による統合スコアの平均が最大となる組を用いた．先行研究である PickScore（63.5%）を基準とし，これを上回った値を **bold** で示す．

LVLM	VAR				SphereAR				FlexVAR				LlamaGen			
	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0
Text-to-Image Alignment																
CLIPScore								53.1								
PickScore								63.5								
Gemma-3	59.7	60.2	59.7	58.0	59.7	59.5	59.6	59.6	59.7	**63.7**	63.1	62.9	59.7	59.8	59.2	59.1
Phi-3.5-Vision	**64.8**	62.7	58.9	55.9	**64.8**	**65.0**	**64.5**	**64.2**	**64.8**	**66.7**	**66.4**	**65.7**	**64.8**	63.3	61.6	60.5
Phi-4-MM	**67.1**	62.6	58.8	55.8	**67.1**	**66.3**	**66.5**	**66.5**	**67.1**	**68.7**	**68.2**	**65.9**	**67.1**	63.5	62.2	60.4
mPLUG-Owl3	**65.3**	61.1	57.3	54.4	**65.3**	**65.6**	**65.0**	**64.0**	**65.3**	**67.3**	**66.5**	**66.1**	**65.3**	62.9	61.3	59.1
Qwen3-VL	51.1	50.7	50.6	50.2	51.1	51.0	51.0	51.0	51.1	54.2	54.3	54.2	51.1	51.8	51.9	52.4
LLaMA-3.2-Vision	**68.4**	63.3	59.5	56.1	**68.4**	**67.6**	**67.2**	**66.8**	**68.4**	**68.0**	**66.3**	**64.9**	**68.4**	**64.7**	62.3	60.4
Ens:LLaMA-3.2+Phi-4-MM+Qwen3-VL	**68.1**	**64.4**	59.4	55.9	**68.1**	**67.8**	**68.0**	**67.9**	**68.1**	**70.5**	**68.5**	**66.6**	**68.1**	**64.6**	62.2	60.5
Ens:LLaMA-3.2+mPLUG-Owl3+Phi-4-MM	**68.1**	63.6	58.8	55.3	**68.1**	**67.9**	**68.5**	**68.3**	**68.1**	**70.3**	**68.4**	**66.5**	**68.1**	**64.7**	62.1	60.3
Ens:Gemma-3+LLaMA-3.2+Phi-4-MM	**67.6**	**64.5**	59.8	56.1	**67.6**	**67.6**	**67.4**	**67.5**	**67.6**	**69.9**	**68.3**	**66.5**	**67.6**	**64.4**	62.4	60.6

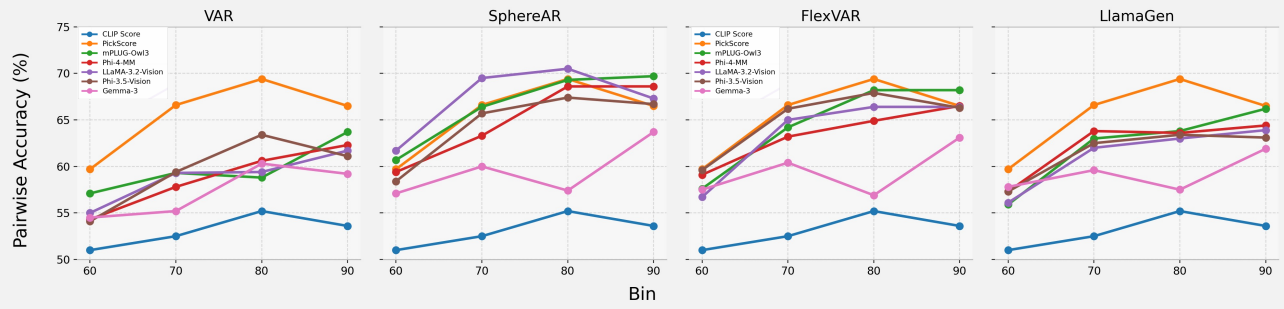

図 1　人間のペアワイズ投票率に基づいてデータを区分し，各区分内で LVLM が人間選好と一致した割合（ペアワイズ正解率）を測定した．例えば 61–70% は “60” 区分として丸めて表示している．各サブ図は異なる自己回帰型画像生成モデル（VAR, SphereAR, FlexVAR, LlamaGen）に対応し，各曲線は異なる LVLM の結果を示す．

と一致する割合を測る正解率を用いる．

$$\mathrm{Acc} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\mathcal{S}(I_n^+ \mid T) > \mathcal{S}(I_n^- \mid T)\right] \tag{7}$$

3.2　評価観点の設定

本研究では，大規模な人手アノテーションに基づく画像生成モデルの評価を行った [13] に従い，同論文で定義された **Text-to-Image Alignment**，**Coherence**，**Preference** の三観点に基づいて検証を行う．評価には，上記の観点に対応するペアワイズ画像選好タスクとして同研究で収集された人手アノテーションデータを用いる．詳細は付録 B に示す．

Text-to-Image Alignment　生成プロンプト p を評価用テキスト T とし，$s_{\mathrm{align}}(I, p)$ を整合性スコアとする．

Coherence / Preference　観点 c に対応する固定文集合 $\mathcal{T}^{(c)} = \{t_m^{(c)}\}_{m=1}^{M}$ を評価用テキストとして用いる．本研究では $M = 10$ とし，人手一致率が高い上位 3 文のみを選択して使用した．評価に用いた文集合およびテンプレートは付録 C に示す．

3.3　使用モデル

$\log P(T \mid I)$ の推定には，デコーダ型の LLaMA [14]，Gemma [15]，Phi [16]，Qwen [17]，mPLUG [18]，といった複数の LVLM を評価器として用い，画像とテキストの整合性を教師強制尤度として算出する．$\log P(I)$ の推定には，自己回帰型画像生成モデルを用いる．本研究では，VAR [19]，FlexVAR [20]，LlamaGen [21]，SphereAR [22] を用い，テキスト非依存な視覚的自然さを確率として評価する．また，比較対象として，埋め込み類似度に基づく既存の画像評価指標である CLIPScore [6]，および人手選好データから学習された報酬モデルである PickScore [10] を用いる．各モデルの詳細は付録 A に記載する．

4　結果と考察

整合性評価　表 2 に，Text-to-Image Alignment におけるペアワイズ正解率を示す．既存の埋め込み類似度指標である CLIPScore や，人手選好データから学習された PickScore と比較して，複数の LVLM を用いた Noisy Channel に基づく評価スコアは，より高

表 3　Coherence および Preference におけるペアワイズ正解率（%）．これらの観点では，評価観点に対応した固定文集合（10 文）から性能が最も高くなる 3 文を選び T として用い，Noisy Channel に基づくスコアにより画像を評価した．先行研究である PickScore を基準とし，これを上回る値を **bold** で示す．

LVLM	VAR				SphereAR				FlexVAR				LlamaGen			
	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0
Coherence																
CLIPScore								51.6								
PickScore								55.5								
Gemma-3	**59.0**	**57.8**	**55.8**	52.7	**59.0**	**58.8**	**58.9**	**59.2**	**59.0**	53.6	**55.6**	**56.5**	**59.0**	**57.4**	**56.1**	55.2
Phi-3.5-Vision	53.6	51.1	49.7	48.9	53.6	54.2	54.5	53.9	53.6	54.8	55.1	54.2	53.6	54.7	54.2	54.0
Phi-4-MM	54.1	50.5	49.4	48.7	54.1	54.1	54.4	54.1	54.1	54.0	54.7	54.1	54.1	55.0	53.8	53.2
mPLUG-Owl3	48.4	47.9	48.2	47.9	48.4	49.1	48.9	49.0	48.4	50.2	50.4	52.0	48.4	52.7	53.1	52.7
Qwen3-VL	**61.3**	53.6	51.0	49.9	**61.3**	**61.4**	**61.4**	**60.7**	**61.3**	**59.9**	**58.4**	**57.2**	**61.3**	**58.6**	**56.2**	55.2
LLaMA-3.2-Vision	54.7	50.2	49.2	48.5	54.7	54.5	54.2	54.0	54.7	54.8	54.2	**56.0**	54.7	54.4	53.4	53.3
Preference																
CLIPScore								50.4								
PickScore								57.8								
Gemma-3	53.5	54.9	55.8	56.3	53.5	53.0	53.0	53.1	53.5	53.0	52.5	53.1	53.5	52.6	51.1	50.2
Phi-3.5-Vision	**59.9**	**58.7**	56.8	56.3	**59.9**	**60.0**	**60.6**	**60.4**	**59.9**	**62.2**	**59.8**	56.9	**59.9**	52.7	50.1	48.8
Phi-4-MM	44.8	50.5	51.9	53.1	44.8	45.4	45.5	45.8	44.8	47.4	47.9	45.6	44.8	44.4	44.9	45.4
mPLUG-Owl3	**62.5**	**60.0**	**58.0**	56.5	**62.5**	**63.4**	**63.5**	**63.6**	**62.5**	**61.0**	**59.8**	57.2	**62.5**	53.5	50.4	49.2
Qwen3-VL	52.4	55.6	55.1	55.5	52.4	52.6	52.6	53.1	52.4	51.4	52.2	50.6	52.4	49.6	49.0	48.4
LLaMA-3.2-Vision	53.9	55.0	54.8	54.5	53.9	54.2	54.1	54.3	53.9	53.6	51.5	49.5	53.9	48.9	47.4	46.8

い一致率を示す傾向が確認される．特に $\alpha = 0$ においても一定の性能が得られており，LVLM の教師強制尤度 $\log P(T \mid I)$ が参照を用いない設定でも人手評価と整合していることが分かる．この結果は，人手選好を直接学習した PickScore のような評価器とは異なり，LVLM の推論能力そのものを評価器として活用できることを示す．AR モデルについては，SphereAR や FlexVAR において $\alpha > 0$ の設定で性能の向上が見られ，特に FlexVAR の $\alpha = 0.3$ では改善が顕著である．これは，画像尤度 $\log P(I)$ が，LVLM のみでは捉えにくい視覚的自然さや破綻を補完し，整合性評価に有効に機能していることを示唆している．さらに，複数の LVLM を用いたアンサンブルは単一モデルを上回る一致率を示し，本設定における最良の性能を達成した．これは，異なる LVLM が持つ補完的な推論を統合することで，より安定した整合性評価が可能となることを示している．

一致率帯別の評価　図 1 に，人手選好率に基づく一致度区分（60–100 %）ごとのペアワイズ正解率を示す．全体として，人手一致度が高い区分ほど提案手法の正解率も高くなる傾向が確認され，人間の判断と整合した挙動を示すことがわかる．一方で，80 %付近では性能がほぼ横ばいとなり，高一致度帯では正解率が収束する傾向が見られる．また，自己回帰型画像モデル（AR）と LVLM の組み合わせによって性能推移は大きく異なり，同一の LVLM を用いた場合でも AR 側の違いにより曲線形状が変化する．これは評価性能が単一のモデル特性ではなく，AR と LVLM の組み合わせに依存して変動することを示している．

観点別評価　表 3 に，Coherence および Preference におけるペアワイズ正解率を示す．Coherence では Gemma-3 や Qwen3-VL が高い一致率を示し，一方で Preference では Phi-3.5-Vision や mPLUG-Owl3 が高い性能を示す．この結果から，評価観点に応じて有効な LVLM が異なることが分かる．これらの結果は，LVLM の指示追従能力を活かすことで，評価観点を容易に追加・切り替えながら，同一のスコアリング枠組みで画像評価を行えることを示している．これは，従来のスコアリング手法では困難であった観点依存の評価を可能にするものである．

5　おわりに

本研究では，画像生成モデルの評価を，生成確率に基づく Noisy Channel モデルとして定式化し，テキスト整合性と視覚的品質を同一の確率的枠組みで評価する手法を提案した．提案手法は，LVLM の教師強制尤度と自己回帰型画像生成モデルの尤度を組み合わせることで，参照を仮定せず各画像を独立して評価できる．検証の結果，提案手法は人手評価と高い整合性を示し，評価観点を切り替えることで，同一の確率的枠組みのもとで多様な人手判断を柔軟に捉えられることを確認した．本手法は，観点可変かつ確率的な画像生成評価の基盤を提供する．

参考文献

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10684–10695, June 2022.

[2] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[3] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.

[4] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 29. Curran Associates, Inc., 2016.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[6] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, 2023.

[8] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[9] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[10] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[11] Jiahui Chen, Candace Ross, Reyhane Askari-Hemmat, Koustuv Sinha, Melissa Hall, Michal Drozdzal, and Adriana Romero-Soriano. Multi-modal language models as text-to-image model evaluators. **arXiv preprint arXiv:2505.00759**, 2025.

[12] Kevin David Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, Ashwinee Panda, and Tom Goldstein. Finegrain: Evaluating failure modes of text-to-image models with vision language model judges. **arXiv preprint arXiv:2512.02161**, 2025.

[13] Dimitrios Christodoulou and Mads Kuhlmann-Jørgensen. Finding the subjective truth: Collecting 2 million votes for comprehensive gen-ai model evaluation, 2024.

[14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[15] Gemma Team. Gemma 3 technical report, 2025.

[16] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[17] Shuai Bai et al. Qwen3-vl technical report, 2025.

[18] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024.

[19] Keyu Tian, Yi Jiang, Zehuan Yuan, BINGYUE PENG, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In **The Thirty-eighth Annual Conference on Neural Information Processing Systems**, 2024.

[20] Siyu Jiao, Gengwei Zhang, Yinlong Qian, Jiancheng Huang, Yao Zhao, Humphrey Shi, Lin Ma, Yunchao Wei, and ZEQUN JIE. FlexVAR: Flexible visual autoregressive modeling without residual prediction. In **The Thirty-ninth Annual Conference on Neural Information Processing Systems**, 2025.

[21] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. **arXiv preprint arXiv:2406.06525**, 2024.

[22] Guolin Ke and Hui Xue. Hyperspherical latents improve continuous-token autoregressive generation, 2025.

表 4　LLaMA-3.2-Vision に対する命令追従学習の有無が，提案スコア $S_\alpha(I \mid T)$ に基づくペアワイズ正解率（%）に与える影響を示す．Text-to-Image Alignment，Coherence，Preference の各観点について，素のモデルと命令追従済みモデルをそれぞれ $\log P(T \mid I)$ の推定器として用いた場合の差分を比較した．

LVLM	VAR				SphereAR				FlexVAR				LlamaGen			
	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0	α=0	α=0.3	α=0.6	α=1.0
Text-to-Image Alignment																
LLaMA-3.2-Vision-Instruct	68.4	63.3	59.5	56.1	68.4	67.6	67.2	66.8	68.4	68.0	66.3	64.9	68.4	64.7	62.3	60.4
LLaMA-3.2-Vision	66.2	63.3	60.2	56.3	66.2	66.2	66.1	65.8	66.2	68.2	66.4	65.6	66.2	64.2	62.7	60.6
Coherence																
LLaMA-3.2-Vision-Instruct	54.7	50.2	49.2	48.5	54.7	54.5	54.2	54.0	54.7	54.8	54.2	56.0	54.7	54.4	53.4	53.3
LLaMA-3.2-Vision	54.1	49.7	48.6	48.7	54.1	54.2	54.2	53.8	54.1	53.7	53.7	54.3	54.1	53.6	53.4	53.5
Preference																
LLaMA-3.2-Vision-Instruct	53.9	55.0	54.8	54.5	53.9	54.2	54.1	54.3	53.9	53.6	51.5	49.5	53.9	48.9	47.4	46.8
LLaMA-3.2-Vision	51.6	54.5	54.2	53.9	51.6	52.9	52.8	52.9	51.6	50.3	49.9	49.7	51.6	47.5	46.4	46.3

表 5　論文中で使用したモデル表記と，Hugging Face 上の識別子との対応を示す．

論文表記	Hugging Face ID
CLIPScore	`openai/clip-vit-large-patch14`
Gemma-3	`google/gemma-3-4b-it`
LLaMA-3.2-Vision	`meta-llama/Llama-3.2-11B-Vision-Instruct`
LLaMA-4-Scout	`meta-llama/Llama-4-Scout-17B-16E-Instruct`
mPLUG-Owl3	`mPLUG/mPLUG-Owl3-7B-240728`
Phi-3.5-Vision	`microsoft/Phi-3.5-vision-instruct`
Phi-4-MM	`microsoft/Phi-4-multimodal-instruct`
Qwen3-VL	`Qwen/Qwen3-VL-8B-Instruct`
PickScore	`uvalkirstain/PickScore_v1`

表 6　本研究で使用した人手アノテーション付きデータセットと，各データセットが対応する評価観点を示す．

データ	評価観点
Rapidata/Flux_SD3_MJ_Dalle_Human_Coherence_Dataset	Coherence
Rapidata/Flux_SD3_MJ_Dalle_Human_Alignment_Dataset	Text-to-Image Alignment
Rapidata/700k_Human_Preference_Dataset_FLUX_SD3_MJ_DALLE3	Preference

表 7　本研究で用いた評価テンプレート一覧．各観点（Coherence / Preference）について英語の固定評価文を 10 文ずつ用い，LVLM による教師強制対数尤度の平均を各観点の評価スコアとして用いる．

評価観点	テンプレート
Coherence	The scene is physically plausible and follows real-world physics. The lighting in the image is consistent and realistic. Object shapes and proportions look natural and well-formed. There are no impossible or contradictory elements in the scene. Textures and materials appear realistic and coherent. The spatial layout and perspective are geometrically consistent. Human figures and body parts, if present, look anatomically plausible. Reflections and shadows match the environment correctly. The image does not contain distorted or melted objects. Overall, the scene appears visually coherent and believable.
Preference	The image has a visually pleasing overall composition. The color palette is harmonious and appealing. The image feels balanced rather than cluttered. The main subject stands out clearly. The image has a clean and polished appearance. The visual style feels intentional and well-executed. The image is comfortable to look at without visual strain. Details are clear without being overwhelming. The image gives a positive visual impression. Overall, the image is aesthetically pleasing.

A　モデルの詳細

本研究では複数の LVLM を公平に比較するため，すべての実験を単一の NVIDIA RTX 6000 Ada GPU 上で実施した．推論は半精度で行い，前処理は各モデル付属の `AutoProcessor` に統一した．教師強制による条件付き尤度 $\log P(T \mid I)$ の計算では，チャットテンプレートとマスク処理を共通化し，モデル差分に依存しない比較を可能とした．使用したモデルと Hugging Face の識別子を表 5 に示す．

B　データセット

本研究で用いた人手選好データセットの詳細および HuggingFace 上での名称を表 6 に示す．これらのデータセットは，いずれも生成画像ペアに対する人手選好率（preference rate）を含んでおり，本研究ではこの選好率に基づきデータを層化し評価に用いた．具体的には，選好率を 60–100 の範囲に限定し，61–70, 71–80, 81–90, 91–100 の 4 つの区分に分類した．各区分から 1,000 件ずつサンプリングし，合計 4,000 件のデータを評価実験に用いた．本設定は，極端な選好差を持たない画像ペアを均等に含むことで，生成モデルの微細な差異を安定に観測することを目的に検証を行った．

C　評価テンプレート一覧

Coherence および Preference の評価に対応する固定テンプレート文集合 $\mathcal{U}^{(c)}$ の具体例を表 7 に示す．

D　Instruction Tuning の影響

Instruction Tuning が整合性項 $\log P(T \mid I)$ の推定に与える影響を検証した．表 4 に示すように，Text-to-Image Alignment，Coherence，Preference のいずれにおいても命令追従モデルの方が一貫して高いスコアを示したが，差は小さく，非 Instruction モデルでも十分に機能した．したがって，観点 c に対して $\log P(T^{(c)} \mid I)$ を用いる際，Instruction Tuning は性能をわずかに改善するが必須ではなく，本手法は非 Instruction モデルでも成立することが確認された．