動画拡散モデルを用いた 複数物体におけるゼロショット動作制御

梶 凌太[1,a)]　柳井 啓司[1,b)]

概要

本研究では，大規模に事前学習された Text-to-Video Diffusion Model の Cross, Spatial, Temporal Attention Map を制御することで，生成動画内の複数のオブジェクトについてゼロショットでコントロールする手法を提案する．条件付けとして，動画全体を表すテキストと各オブジェクトの軌跡を表すバウンディングボックスを用いることで生成動画を制御することができる．

1. はじめに

近年，大規模な動画データで学習された Text-to-Video(T2V) モデルが多数発表され [1][2][3][10][12]，その中でも Sora[2] 等は非常に高い生成品質を達成しており，画像生成に次いで動画生成に対しても注目が集まっている．今後，社会実装に耐えうる性能を持つ動画生成モデルが登場し，教育やコミュニケーションツール，コマーシャル活用など非常に幅広い分野で応用されることが考えられる．

しかしながら，現在の大規模な動画生成モデルの多くは生成品質の向上に重きが置かれており，生成する動画内容についての細かい条件付けができないことが多い．そのため，動画生成モデルの利用者は動画制作のコスト削減の利点がある一方で，望んだ動画を生成するために複数回動画を生成する必要がある．先行研究である PEEKABOO[6] は，作成したい動画のテキストプロンプトとプロンプト中の動かしたいオブジェクトの軌跡を条件として，大規模な動画生成モデルに対してゼロショットで生成動画を制御する手法を提案している．しかしながら，制御する対象が 1 つのオブジェクトに限定されており，社会実装において活用できる幅は少ない．

そこで本研究では，大規模な動画生成モデルをゼロショットで制御可能な手法である PEEKABOO[6] を拡張し，マルチオブジェクトに対する制御をゼロショットで行う手法を提案する．本手法では，PEEKABOO[6] で提案されている Spatial, Temporal, Cross Attention における Masked Attention をマルチオブジェクトに拡張するのに加え，より条件に対して忠実に生成するための Latent Alignment を導入する．

1 電気通信大学 大学院情報理工学研究科 情報学専攻
a) kaji-r@mm.inf.uec.ac.jp
b) yanai@cs.uec.ac.jp

2. 関連研究

近年，Latent Diffusion Models (LDM) [9] 構造を用いたテキストベースの動画生成モデルは目覚ましい発展を遂げている [1][2][3][4][10][12]．VideoLDM [1] は，事前学習された LDM [9] の Spatial Attention，Cross Attention に加え，微調整のための Temporal Attention を追加することで，画像から動画への拡張を可能にしている．この事前学習された空間層に対して微調整用の時間層を追加するパイプラインは，効率的な動画生成モデルの構築手法として幅広く採用されている．しかしながら，これらの大規模な T2V モデルは，生成品質の向上を目的に大量の動画データによって学習されることが多く，時空間制御についてはあまり検討されていない．

時空間制御を探求している先行研究としては，Local な Animation と User-friendly な入力を提案している Follow-your-click [8], ControlVideo [11]，一貫性のある動画を生成するために大規模言語モデルと Text-to-Image モデルを組み合わせた Free-Bloom [5] などがある．これらの手法は追加のトレーニングデータや Reference Video，別の大規模モデルを利用する必要がある場合が多い．それに対して PEEKABOO [6] では，大規模 T2V モデルに対して Attention Map をコントロールすることで，ゼロショットで時空間制御を可能にしている．したがって，本研究では PEEKABOO [6] で提案されている Masked Attention をマルチオブジェクト向けに拡張する．

潜在変数の探求については，Text2Video-Zero [7]，PYoCo [3] などの手法がある．これらの手法は，動画の時間軸にわたって整列された潜在変数を利用することで，ゼロショットで動画へ拡張したり少ないパラメータ数で高品質な動画を生成することを可能としている．本研究では，生成動画における時間方向の一貫性を向上させるために，先行研究の手法 [3] を参考に Latent Alignment を提案する．

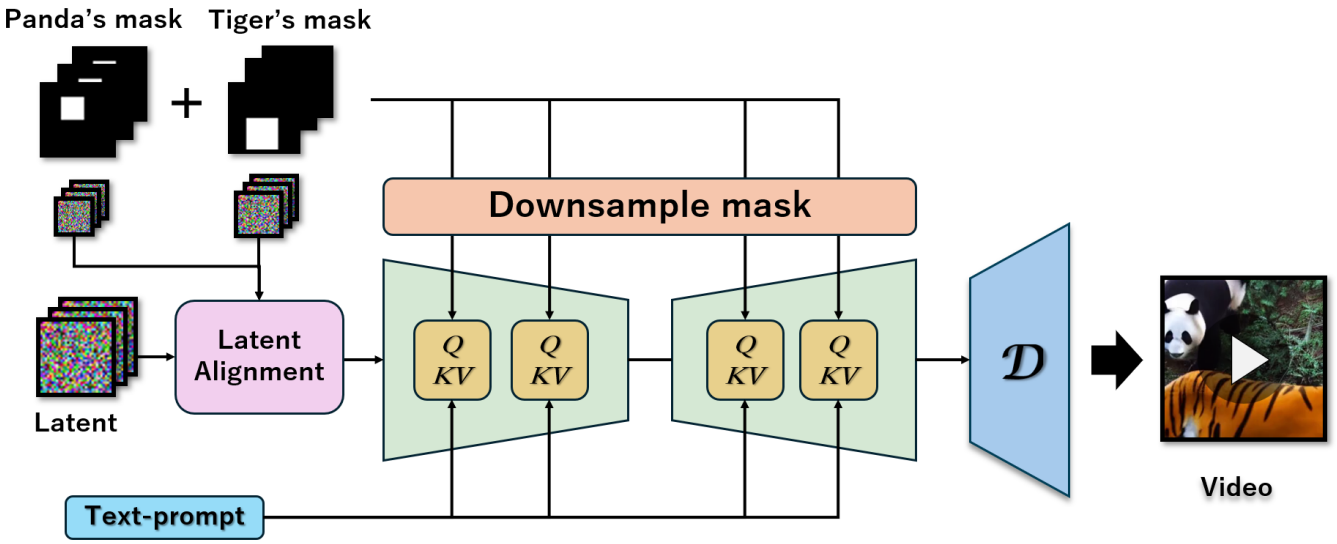

図 1 本手法の概要図

3. 準備: PEEKABOO

Jain ら [6] は生成時に条件付けされるオブジェクトのバウンディングボックス軌跡をもとに，事前学習された T2V モデルの Attention Map を制御することで，ゼロショットで生成動画をコントロールする手法を提案している．具体的には，T2V 内の Spatial Attention，Cross Attention，Temporal Attention それぞれについて，前景画素と背景画素がそれぞれの領域内のみに注目するように Attention Mask を用いて計算を行う．この計算は Diffusion Models (DM) のサンプリングにおける生成ステップの初期に数ステップのみ適応され，その後は自由に生成される．

注目領域を制御する Masked Attention では，任意のクエリ Q，キー K，バリュー V に対して，バイナリの二次元マスク M を用いて式 1 のように計算される．

$$\text{MaskedAttention}(Q,K,V,M) = \text{softmax}(\frac{QK^T}{\sqrt{d}} + \mathcal{M})V$$
$$\text{where} \quad \mathcal{M}[i,j] = \begin{cases} -\infty & \text{if } M[i,j]=0 \\ 0 & \text{if } M[i,j]=1 \end{cases} \tag{1}$$

ここで加算マスク $\mathcal{M}$ は，マスク M の値が 0 の時大きな負の値を持つため，各アテンションの計算において注目領域を制御することができる．また，スケール値 $\sqrt{d}$ は各クエリの系列長である．系列長について，テキストプロンプトの長さを l_{text}，ビデオの長さを l_{video}，潜在変数の空間方向のサイズを l_{latent} とする．また，ピクセルまたはテキストトークンを入力とし，それが動画の前景に対応する場合は 1 を，そうでない場合は 0 を返す関数 $\text{fg}(\cdot)$ を定義する．2 次元バイナリマスク M の計算方法については，Spatial, Temporal Attention では前景と背景をそれぞれの領域内にのみ注目するようにし，Cross Attention では条件付けされたバウンディングボックス内にオブジェクトが生成されるよう，前景と背景それぞれについて関係するテキスト Embedding のみ取り込むことで，オブジェクトの位置，大きさ，動きを制御することができる．

Spatial, temporal attention mask

Spatial Attention では各フレーム f に対して，$l_{latent} \times l_{latent}$ の 2 次元行列であるマスク M^f_{SA} を計算する．各ピクセルペアに対して，両方のピクセルが前景もしくは背景の場合に 1 となる．形式的には，式 2 のようになる．

$$M^f_{SA}[i,j] = \text{fg}(M^f_{input}[i]) * \text{fg}(M^f_{input}[j]) + (1-\text{fg}(M^f_{input}[i])) * (1-\text{fg}(M^f_{input}[j])) \tag{2}$$

Temporal Attention では，座標 i のピクセルについて，$l_{video} \times l_{video}$ の 2 次元行列であるマスク M^i_{TA} を計算する．各フレームペアについて，両方のフレームが前景もしくは背景の場合に 1 となる．形式的には，式 3 のようになる．

$$M^i_{TA}[f,k] = \text{fg}(M^f_{input}[i]) * \text{fg}(M^k_{input}[i]) + (1-\text{fg}(M^f_{input}[i])) * (1-\text{fg}(M^k_{input}[i])) \tag{3}$$

Cross attention mask

Cross Attention では, 各フレーム f に対して, $l_{video} \times l_{text}$ の二次元行列であるマスク M^f_{CA} を計算する．ピクセルとテキストトークンのペアについて互いに前景もしくは背景の場合に 1 となる．形式的には，式 4 のようになる．

$$M_{CA}^{f}[i,j] = \mathrm{fg}(M_{input}^{f}[i]) * \mathrm{fg}(T[j]) + (1-\mathrm{fg}(M_{input}^{f}[i])) * (1-\mathrm{fg}(T[j])) \quad (4)$$

4. 手法

提案手法の全体図を図 1 に示す．本手法は，生成したい動画の入力テキストと各オブジェクトのバウンディングボックス軌跡をマスク画像に変換したものを条件として動画を生成する．本手法は，先行研究 PEEKABOO [6] の Masked Attention を複数オブジェクトに拡張した Multi Masked Attention と，条件付けへの忠実性を高めるための Latent Alignment の 2 つの主要なコンポーネントから構成される．

4.1 Multi Masked Attention

PEEKABOO[6] の Masked Attention では前景と背景について互いのピクセルが参照しないよう Attention Mask を設計しているが，本手法では前景と背景の分離に加え，前景内のオブジェクト同士も分離した Attention Mask を設計する．形式的には，条件付ける n 個の各オブジェクト $\mathbf{O} = [o_1, o_2, ..., o_n]$ について，入力マスク $M_{input}^{o_i}$ を他のオブジェクトの入力マスクと重ならない $\hat{M}_{input}^{o_i}$ に変換した後，式 2,3,4 より Attention Mask を生成する．各オブジェクトの入力マスク $\hat{M}_{input}^{o_i}$ は式 5 で表される．

$$\hat{M}_{input}^{o_i} = M_{input}^{o_i} \odot \prod_{j=1}^{i-1} (1 - M_{input}^{o_j}) \quad (5)$$

4.2 Latent Alignment

Video diffusion model を用いた動画生成では，生成時における初期潜在変数が生成動画の品質に影響を与えるため，ノイズの設計についてさまざまな研究が行われている [3][7]．本研究では，一般に動画フレーム間のコンテンツは類似しており，類似したコンテンツ間の潜在変数は類似しているという観察から，入力マスクに対してより忠実に複数オブジェクトを制御するための Latent Alignment を導入する．Latent Alignment では DM のサンプリングに用いる潜在変数に対して，各オブジェクトの入力マスク箇所についてフレーム間で類似した潜在変数を用いる．

具体的には，初期潜在変数 $\mathbf{z}$ とは別に各オブジェクトのベースノイズとなる $\mathbf{z}_{o_i}$ をサンプリングする．

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{z}_{o_i} \sim \mathcal{N}\left(\mathbf{0}, \frac{\alpha^2}{1+\alpha^2}\mathbf{I}\right) \quad (6)$$

この時，各オブジェクトのベースノイズのサイズは各オブジェクトの入力バウンディングボックスサイズと同様である．その後，式 7 により表されるフレーム f の $\mathbf{z}_{o_i}^{f}$ を計算したのち，各オブジェクトの入力マスク $\hat{M}_{input}^{o_i}$ に基づいて初期潜在変数 $\mathbf{z}$ に統合される．

$$\epsilon_{o_i}^{f} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{1+\alpha^2}\mathbf{I}\right), \quad \mathbf{z}_{o_i}^{f} = \mathbf{z}_{o_i} + \epsilon_{o_i}^{f} \quad (7)$$
$$\mathbf{z}[\hat{M}_{input}^{o_i}] = \mathbf{z}_{o_i}$$

ここで，α は各フレーム間でのベースノイズの割合をコントロールするハイパーパラメータであり，$\alpha \to \infty$ の時各フレームのオブジェクトノイズはベースノイズに等しくなり，$\alpha = 0$ の時，各フレームのオブジェクトノイズは独立となる．

5. 実験

本手法では，事前学習済みの T2V モデルに ModelScope[10] を活用し，モデル内部の 3D U-Net を構成する Spatial Attention, Cross Attention, Temporal Attention に対して Multi Masked Attention を適用した．生成される動画の解像度は 16 フレーム，256×256 であり，生成サンプラーには DDIMSampler を用いた．また， [6] に従い Multi Masked Attention は生成ステップ 40 ステップのうちの最初の 2 ステップだけ行い，残りの 38 ステップは Attention Map の制約なしで生成した．Latent Alignment のパラメータ α については経験的に 0.2 に設定した．

提案手法の評価方法については，各生成結果について定性評価を行った．5.1 ではテキストと入力マスクによる生成結果を，5.2 では Multi Masked Attention と Latent Alignment についての Ablation study の結果について述べる．

5.1 定性評価

入力したテキストとバウンディングボックス軌跡および生成結果を図 2 に示す．左に示しているのが，各オブジェクトに対応した初期位置のバウンディングボックスとその後の軌跡の方向であり，生成結果の下に記述してある入力テキストプロンプト内のオブジェクトと色が対応している．図 2 の上二行の生成結果を見ると，二匹のカエルと二隻の船について入力時に条件づけた軌跡通りに生成されていることが分かる．三行目の鴨および四行目のパンダとトラの生成結果については，初期フレーム時には片方のオブジェクトが動画中に表れないものの，おおむね条件づけた通りに生成できていることが分かる．

5.2 Ablation study

二つのコンポーネントである，Masked Attention と Latent Alignment についての Ablation study の結果を図 3 に示す．

また，図 3 中において MA は Masked Attention，LA は Latent Alignment である．Masked Attention なしの結果を見てみると，動画中にパンダしか登場しておらず，また Mask の軌跡と全く違う動作をしているため，Masked

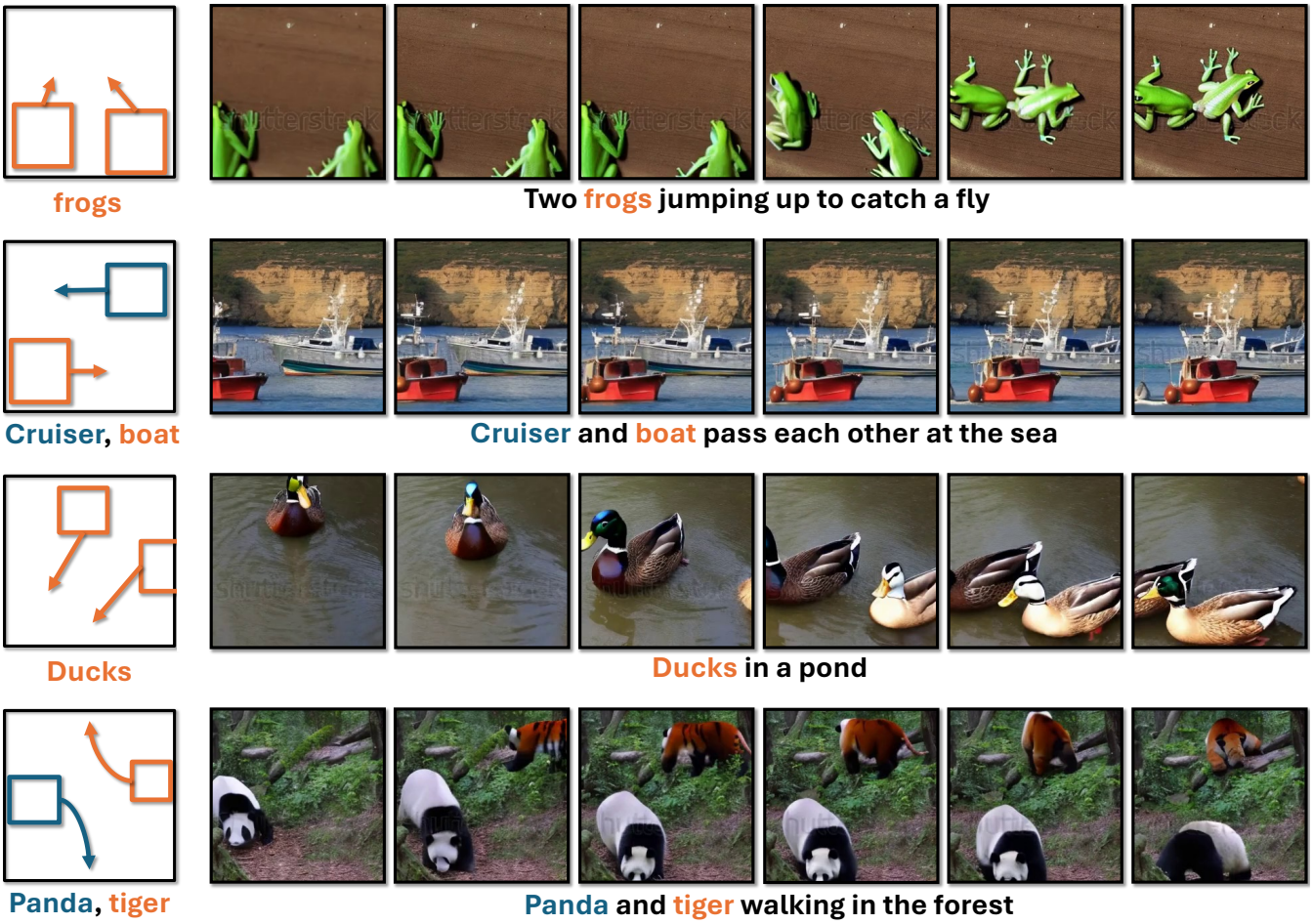

図 2　提案手法の生成結果

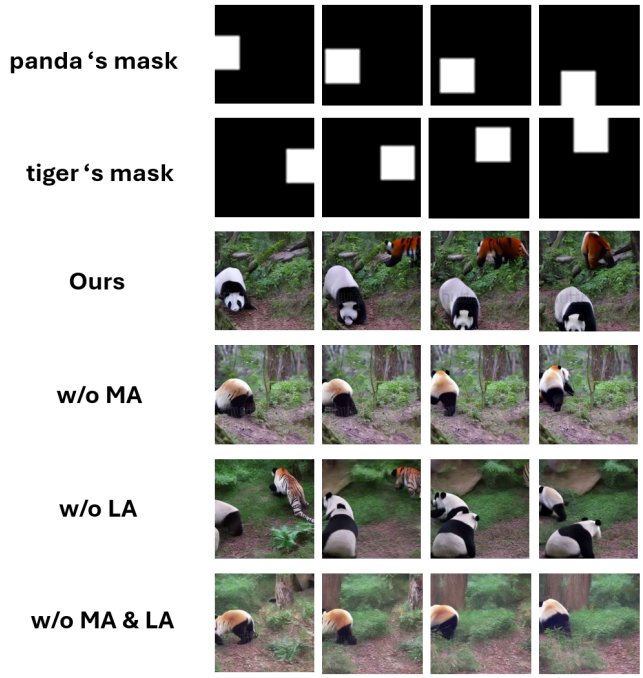

図 3　Ablation study 結果

Attention は必須のコンポーネントであることが分かる．次に Latent Alignment なしの結果を見てみると，Masked Attention によりわずかにガイダンスできているものの，動画後半ではタイガーがいなくなりパンダが増えるなど，Masked Attention のみでは条件付けに不十分であることが分かる．

6.　おわりに

本研究では，大規模に事前学習された Text-to-Video Diffusion Model の Cross, Spatial, Temporal Attention Map を制御することで，ゼロショットで複数のオブジェクトをコントロールする手法を提案した．複数オブジェクトに対応するための Multi Masked Attention に加え，入力された条件に対してより忠実に生成するための Latent Alignment を導入することで，安定して複数オブジェクトの異なる動きの制御が可能となった．しかしながら，Latent Alignment はオブジェクトのベースノイズを生成する際のランダム性が生成品質に大きく影響を与えるものであったり，Multi Masked Attention はガイダンス後の自由生成によりオブジェクト同士が混ざって生成されてしまうなど，提案手法については未だ複数の課題を抱えている．そのため，今後はこれらのマルチオブジェクトに起因する課題の解決に取り組む考えである．

参考文献

[1] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S. and Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 22563–22575 (2023).

[2] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R. and Ramesh, A.: Video generation models as world simulators (2024).

[3] Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.-B., Liu, M.-Y. and Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models, *Proc.of IEEE International Conference on Computer Vision*, pp. 22930–22941 (2023).

[4] Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D. and Dai, B.: AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning, *Proc. of International Conference on Learning Representation* (2023).

[5] Huang, H., Feng, Y., Shi, C., Xu, L., Yu, J. and Yang, S.: Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator, *Proc. of Advances in Neural Information Processing Systems*, Vol. 36 (2024).

[6] Jain, Y., Nasery, A., Vineet, V. and Behl, H.: Peekaboo: Interactive video generation via masked-diffusion, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 8079–8088 (2024).

[7] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S. and Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators, *Proc.of IEEE International Conference on Computer Vision*, pp. 15954–15964 (2023).

[8] Ma, Y., He, Y., Wang, H., Wang, A., Qi, C., Cai, C., Li, X., Li, Z., Shum, H. Y., Liu, W. and Chen, Q.: Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts, *arXiv preprint arXiv:2403.08268* (2024).

[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-resolution image synthesis with latent diffusion models, *Proc.of IEEE Computer Vision and Pattern Recognition*, pp. 10684–10695 (2022).

[10] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X. and Zhang, S.: Modelscope text-to-video technical report, *arXiv preprint arXiv:2308.06571* (2023).

[11] Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W. and Tian, Q.: ControlVideo: Training-free Controllable Text-to-video Generation, *Proc. of International Conference on Learning Representation* (2023).

[12] Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y. and Feng, J.: Magicvideo: Efficient video generation with latent diffusion models, *arXiv preprint arXiv:2211.11018* (2022).