# FVD：映像生成の新しい指標

## 抄録

近年の深層生成モデルの進歩により、高画質画像の合成は目覚ましい進歩を遂げている。画像処理や表現学習への応用が成功したのに続き、次のステップとして重要なのが動画である。映像の生成モデルを学習するのは、オブジェクトの視覚的表現に加えて、シーンの時間的ダイナミクスを捉えるモデルを必要とする、はるかに難しいタスクである。最近のビデオの生成モデルはある程度の成功を収めているが、現在の進歩は、視覚的品質、時間的一貫性、サンプルの多様性を考慮した定性的な測定基準の欠如によって妨げられている。そこで我々は、FIDに基づくビデオの生成モデルのための新しいメトリックであるFrechet Video Distance (FVD)を提案する。我々は、FVDが生成されたビデオに対する人間の定性的判断とよく相関することを確認した大規模な人間研究に貢献する。

## 1 はじめに

近年の深層生成モデルの進歩は、高品質な画像の合成 [1] [2] に目覚ましい成功をもたらしている。次の挑戦は、ビデオ制作を考えることだ。これは、オブジェクトの視覚的表現に加えて、シーンの時間的ダイナミクスを捉えるモデルを必要とする、はるかに難しいタスクである。ビデオの生成モデルは、欠落フレームの予測[3]、インスタンスのセグメンテーションの改善[4]、あるいは推論[5]を行うことによる複雑な（関係性のある）推論タスクなど、多くのアプリケーションを可能にする。

近年大きな進歩が見られるが、ビデオ生成モデルはまだ発展途上であり、一般的に数秒以上のビデオを合成することはできない[6]。優れたダイナミクスモデルを学習することは、実世界の動画を生成する上で依然として大きな課題である。しかし、ビデオ合成の進歩を定性的に測定するためには、視覚的品質、時間的一貫性、生成サンプルの多様性を考慮した測定基準が必要である。

我々は、ビデオの生成モデルのための新しいメトリックであるFrechet Video Distance (FVD)を提供する。FVDは、画像への適用に成功している (FID) [7]の根底にある原理に基づいている。各フレームの品質に加え、映像コンテンツの時間的一貫性を捉える特徴表現を紹介する。ピーク信号対雑音比（PSNR）や構造的類似度（SSIM）[8]のような一般的なメトリクスとは異なり、FVDはビデオ上の分布を考慮するため、フレームレベルのメトリクスの欠点を回避することができます[9]。FVDは、生成されたビデオに対する人間の定性的判断とよく一致することを確認した大規模な人間研究を含む、FVDを評価するための広範な実験に貢献しています。

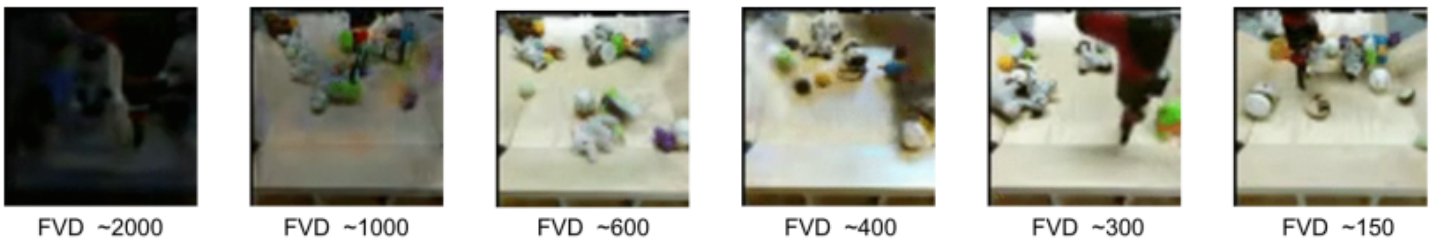


**図1**：様々なモデルによって生成されたビデオをFVDに従ってランク付け（低い方が良い）

# 2 フレシェ動画距離

ビデオの生成モデルは、観察されたデータが生成された基礎となるデータ分布をとらえなければならない。実世界のデータ分布 $P_R$ と、生成モデル $P_G$ によって定義された分布との間の距離は、自然な評価指標である。通常、どちらの分布も解析的な表現ができないため、多くの一般的な距離関数をそのまま適用することができない。例えば、$P_R$ と $P_G$ の間の一般的なフレシェ距離（または2-ワッサーシュタイン距離）は、$d\left(P_R、\ P_G\right) = \min_{X,Y} E\left|X - Y\right|^2$と定義される。ここで、最小化は、それぞれ分布 $P_R$ と $P_G$ を持つすべての確率変数 $X$ と $Y$ にわたって行われる。この式は，一般的な場合には解くのが難しいが、$P_R$ と $P_G$ が多変量ガウスである場合には、閉形式の解がある[10]。

$$d(P_R, P_G) = \left|\mu_R - \mu_G\right|^2 + Tr\left(\Sigma_R + \Sigma_G - 2(\Sigma_R\Sigma_G)^{\frac{1}{2}}\right) \tag{1}$$

ここで，$\mu_R$ と $\mu_G$ は平均であり，$\Sigma_R$ と $\Sigma_G$ は $P_R$ と $P_G$ の共分散行列である。多変量ガウシアンは、基本的なデータ分布の正確な表現であることはめったにないが、適切な特徴空間を使用する場合、これは合理的な近似である。実世界の画像上の分布について、Heuselら(2017)は、学習された特徴埋め込みを用いて、$P_R$ と $P_G$ の間の距離を以下のように計算した[11]。$P_R$ と $P_G$ からの最初のサンプルは、ImageNet で訓練された Inception ネットワーク [12] に供給され、隠れ層の1つにおけるそれらの特徴表現（活性化）が記録される。次に、フレシェ開始距離（FID）[7:1]が、実サンプルと生成サンプルの記録された応答から得られた平均と共分散を用いて、式1により計算される。

事前に訓練されたニューラルネットワークによって学習された特徴表現は、メトリックの品質に大きく影響する。ImageNet で学習する場合、学習された特徴量は画像内のオブジェクトに関する推論に有用な情報を露出するが、それ以外の情報内容は抑制される可能性がある。同様に、ネットワークの異なるレイヤーは、異なる抽象化レベルで特徴をエンコードする。動画に適した特徴表現を得るためには、任意の時点における視覚的提示に加えて、一連のフレームにわたる視覚的内容の時間的一貫性を考慮した、事前に訓練されたネットワークが必要である。この研究では、事前に訓練された Inflated 3D Convnet (I3D) [13] のいくつかのバリエーションを調査し、得られたメトリックを Frechet Video Distance (FVD)と名付けた。同様のFIDの適応は、[14]が彼らのvid2vidモデルを評価するために使用した。ここでは、ビデオの一般的な指標としてFVDを紹介し、広範な実証研究に焦点を当てる。

I3D ネットワークは、Inception アーキテクチャをシーケンシャルデータに一般化したもので、人間中心の YouTube ビデオで構成される Kinetics データセット[15]の行動認識を行うように訓練されている。行動認識には、視覚的文脈と時間発展を同時に考慮する必要があるが、I3D はこのタスクに優れていることが示されている。我々は、Kinetics-400 と Kinetics-600 で事前に訓練された I3D ネットワークによって学習された2つの異なる特徴表現 (logits、avg. pool) を探索した。

式1を使用することの潜在的な欠点は、学習された特徴空間上のガウス分布を推定する際の誤差が大きくなる可能性があることである。Binkowskiら(2018) [16] は、画像の場合の代替案として最大平均不一致 (Maximum Mean Discrepancy) [17]を使用することを提案しており、動画の文脈でもこのバリエーションを検討する。MMD はカーネルベースのアプローチであり、特定の形式を仮定することなく2つの経験分布間の距離を計算する手段を提供する。Binkowskiら(2018) [16:1]は、多項式カーネル $k(a, b) := \left(a^T b + 1\right)^3$ を使用することを提案し、これを I3D ネットワークの学習された特徴に適用して、カーネルビデオ距離 (KVD) を求める。

# 3 実験

以下では、FVDに関するノイズの多い研究、および人間による研究の結果を紹介する。サンプルサイズと分解能に対する感度を分析した追加実験は、付録BとDにある。

## 3.1 ノイズ調査

実映像にノイズを加えることで、FVDが基本的な歪みに対してどの程度敏感かをテストする。個々のフレームに付加される静的ノイズと、一連のフレーム全体を歪ませる時間的ノイズを考える。これらの歪み（詳細は付録Aに記載）を最大6つの異なる強度で適用し、BAIR [18]、Kinetics-400 [13:1]、HMDB51 [19]のデータセットのビデオと、それらのノイズの多い対応するビデオとの間で FVD と KVD を計算した。埋め込み候補として、Kinetics-400 データセットで事前に訓練されたI3Dモデルの一番上のプーリング層と logits 層、および拡張 Kinetics-600 データセットで事前に訓練された I3D モデルのバリアントの同じ層を考慮した。ベースラインとして、動画に対する FID の素朴な拡張と比較した。この拡張では、(ImageNet で事前に訓練された) Inception ネットワークが各フレームに対して個別に評価され、その結果得られた埋め込み（またはそのペアごとの差分）が平均化されて、各動画に対する単一の埋め込みが得られる。この "FID "スコアは式1に従って計算される。

すべてのバリエーションは、注入されたさまざまな歪みをある程度検出することができたが、事前に訓練されたInceptionネットワークは、予想通り、時間的な歪みを検出するのに劣っていた。図2は、Kinetics-400 で事前に訓練された I3D モデルの logits 層が、一連のノイズ強度と最も良い平均順位相関を持つことを示している。ノイズ調査でのスコアの概要は、図3で見ることができる。

## 3.2 人的評価

| Metric | eq.FVD | eq.SSIM | eq.PSNR | | spr.FVD | spr.SSIM | spr.PSNR |
|---|---|---|---|---|---|---|---|
| FVD | N/A | **74.9%** | **81.0%** | | **71.9%** | **58.4%** | **63.5%** |
| SSIM | 51.5% | N/A | 44.6% | | 61.8% | 51.2% | 45.9% |
| PSNR | **56.3%** | 21.4% | N/A | | 54.1% | 37.0% | 44.8% |

**表1**：ある指標（eq.）に対して固定値を持つモデル、または広い範囲（spr.）に値を広げたモデルを考慮した場合の、人間の判断と指標の一致度

生成モデルの性能に関する重要な基準の1つは、人間の観察者によって判断されるサンプルの視覚的忠実度である[20]。そこで、いくつかの条件付き動画生成モデルを訓練し、人間の評価者に異なるシナリオで生成された動画の品質を比較してもらった。

**結果** 人的評価試験の結果は表1に、KVD と Avg.FID の追加結果は付録Cにある。私たちは、FVD がテストされた他のすべてのメトリクスと比較して優れた選択肢であることを発見した。eq.FVD と spr.FVD で得られた結果は、ユーザーが実際に FVD をどのように体験するかを決定するため、重要な意味を持つ。spr.FVD の列から、他のどの指標も FVD によって引き起こされたランキングを改善することはできないと結論づけることができる。また eq.FVD の列は、他のどの指標も FVD の点で同等である優れたモデルを確実に区別することはできないと教えてくれる。

一方、FVDは、他のメトリクス（eq.SSIM、eq.PSNR）では識別できないモデルを識別することができ、人間の判断とよく一致する（それぞれ74.9%、81.0%の一致）。同様に、FVDは、他の指標（SSIM、PSNR）によって引き起こされるランキングを一貫して改善する。

# 結論

動画生成モデルの新しい評価指標であり、動画生成モデルのより良い評価への重要な一歩であるフレシェ動画距離（Frechet Video Distance：FVD）を紹介した。我々の実験では、FVD が静的ノイズや時間的ノイズを含むように修正されたビデオを正確に評価できることが確認された。さらに重要なことに、最近のいくつかの生成モデルから生成されたビデオを対象とした大規模な人間による研究により、FVD が人間の判断と一致するという点で、SSIM と PSNR を一貫して上回ることが明らかになった。

# A ノイズ調査

FVDが静的ノイズを検出できるかどうかをテストするために、一連のビデオフレームの各フレームに以下の歪みのいずれかを加えた。

1. フレーム内のランダムな位置に描かれた黒い四角形
2. フレームにガウシアンスムージングカーネルを適用するガウシアンブラー
3. 観測フレームと標準ガウシアンノイズの間を補間するガウシアンノイズ
4. フレーム内の各ピクセルを一定の確率で黒か白にするソルト＆ペッパーノイズ

時間的ノイズは以下のように加えた。

1. ランダムに選ばれたフレームの数をシーケンス内の隣接するフレームと局所的に入れ替える
2. シーケンス全体にわたってランダムに選ばれたフレームのペアの数を、グローバルに入れ替える
3. 複数の異なる動画に対応するフレームのシーケンスをインターリーブして新しい動画を得る
4. ある動画から別の動画にフレーム数後に切り替えて新しい動画を得る

それぞれのタイプに固有の最大6つの異なる強さで、これらの歪みを適用した、 例えば、黒い四角形の大きさ、スワップの回数、インターリーブするビデオの数などである。

HMDB[19:1]、BAIR[18:1]、Kinetics-400[15:1]を用いてノイズ調査を行った。利用可能なサンプルの合計 90%（訓練とテスト）が比較に使用された。異なるノイズ強度と、我々が考慮する様々なノイズタイプのパラメータ値との対応付けは、表2で見ることができる。

| **Noise type** | **Parameter** | **Int.1** | **Int.2** | **Int.3** | **Int.4** | **Int.5** | **Int.6** |
|---|---|---|---|---|---|---|---|
| Black rectangle | size relative to image | $15\%$ | $30\%$ | $45\%$ | $60\%$ | $75\%$ | $\mathrm{N/A}$ |
| Gaussian blur | sigma of Gaussian kernel | $1$ | $2$ | $3$ | $4$ | $5$ | $\mathrm{N/A}$ |
| Gaussian noise | mixing factor | $15\%$ | $30\%$ | $45\%$ | $60\%$ | $75\%$ | $\mathrm{N/A}$ |
| Salt & Pepper | prob. of applying noise | $0.1$ | $0.2$ | $0.3$ | $0.4$ | $0.5$ | $\mathrm{N/A}$ |
| | | | | | | | |

| Noise type | Parameter | Int.1 | Int.2 | Int.3 | Int.4 | Int.5 | Int.6 |
|---|---|---|---|---|---|---|---|
| Local swap | nr. of swaps | 4 | 8 | 12 | 16 | 20 | 24 |
| Global swap | nr. of swaps | 4 | 8 | 12 | 16 | 20 | 24 |
| Interleaving | nr. of sequences | 2 | 3 | 4 | 5 | 6 | N/A |
| Switching | nr. of frames until switch | 1 | 2 | 3 | 4 | 5 | N/A |

**表2**：ノイズの種類によって異なるノイズ強度の概要

図2は、FVDの様々な実装（およびFIDベースのベースライン）とノイズ強度のシーケンスとの相関の概要を示している。Kinetics-400 データセットで訓練された I3D モデルの対数は、様々なノイズタイプにおいて、ノイズ強度とよく相関していることがわかる。その性能は図3で見ることができる。

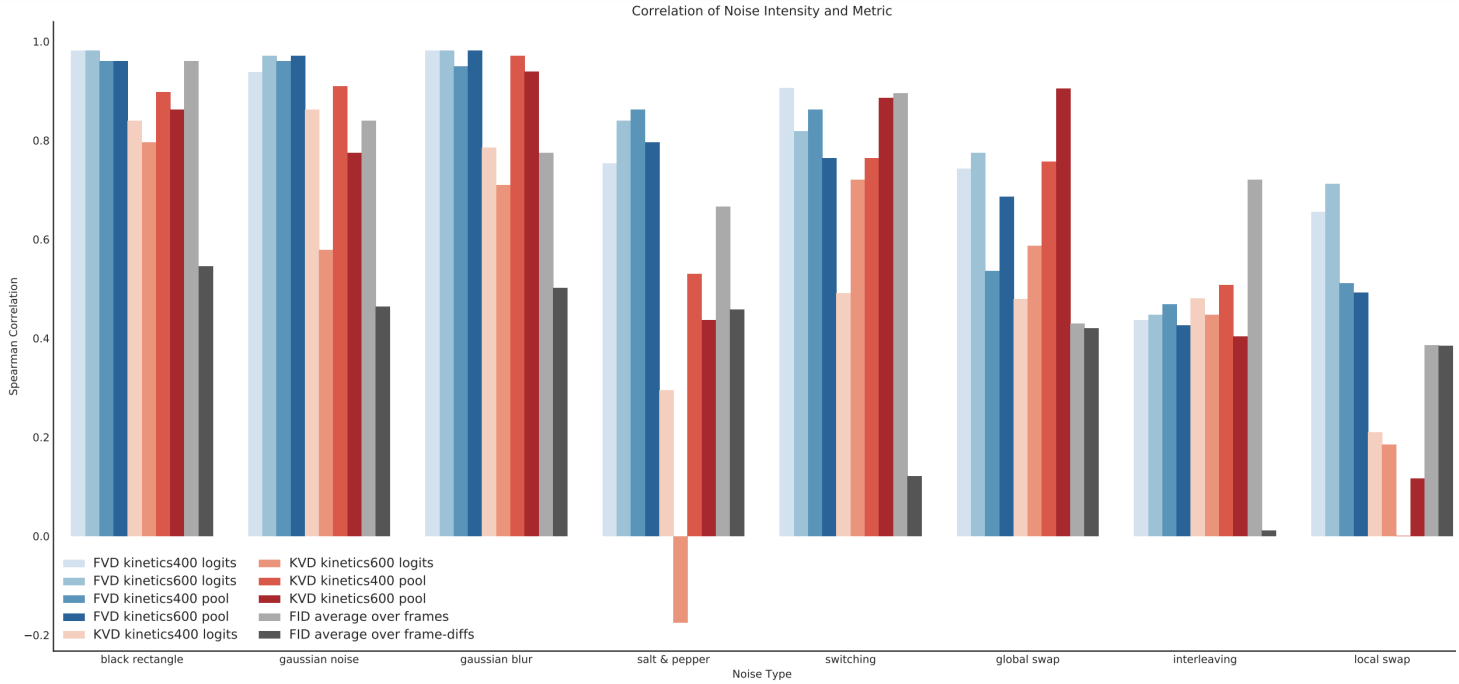


**図2**：ノイズ強度と測定値の相関

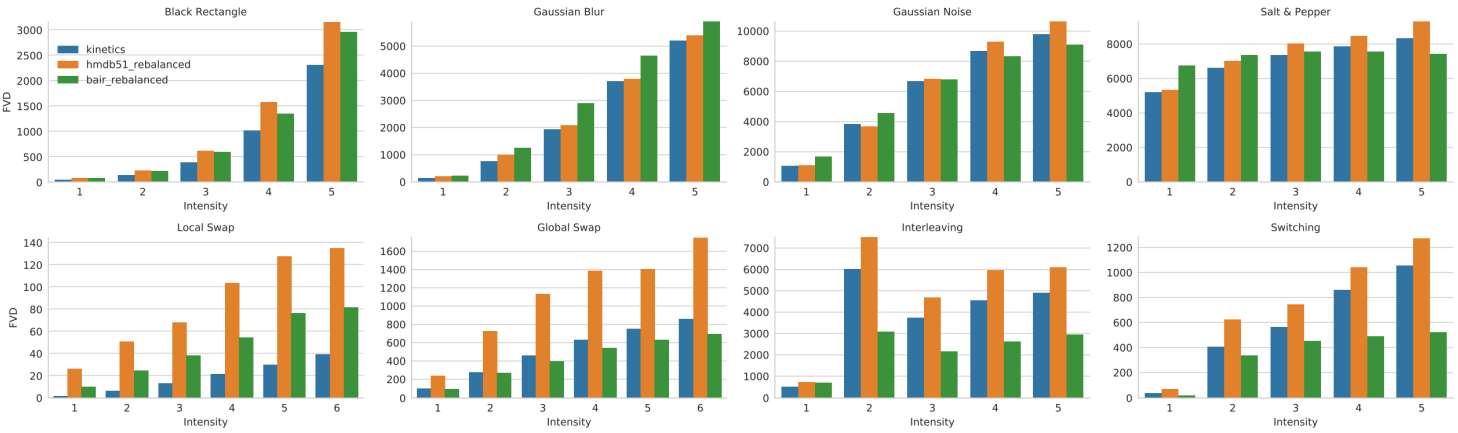


**図3**：Kinetics-400 で学習した I3D モデルの logits アクティベーションを埋め込みとして使用し、異なるデータセットに様々なタイプのノイズを加えた場合の FVD の挙動

# B サンプルサイズがFVDに及ぼす影響

我々は、FVD が生成されたビデオの分布とターゲット分布との間の真の基礎的距離を計算することができる精度を検討する。式1に従って FVD を計算するには、利用可能なサンプルから $\mu_R$、$\mu_G$、$\Sigma_R$、$\Sigma_G$ を推定する必要がある。サンプルサイズが大きければ大きいほど、これらの推定値は良くなり、FVD は分布間の真の基礎的距離をより良く反映することになる。正確な生成モデルの場合、これらの分布は通常かなり近くなり、推定プロセスからのノイズが主に結果に影響する。この効果は、FID [21] [16:2]についてよく研究されており、図4には FVD について描かれている。基礎となる分布が同一であっても、パラメータ $\mu_R$、$\mu_G$、$\Sigma_R$、$\Sigma_G$ の推定値がノイジーであるため、FVD は通常ゼロより大きくなることがわかる。また、一定のサンプル数では標準誤差（50回の試行で測定）は小さく、正確な比較が可能であることもわかる。したがって、モデル間で FVD 値を比較する際には、同じサンプルサイズを使用することが重要である。

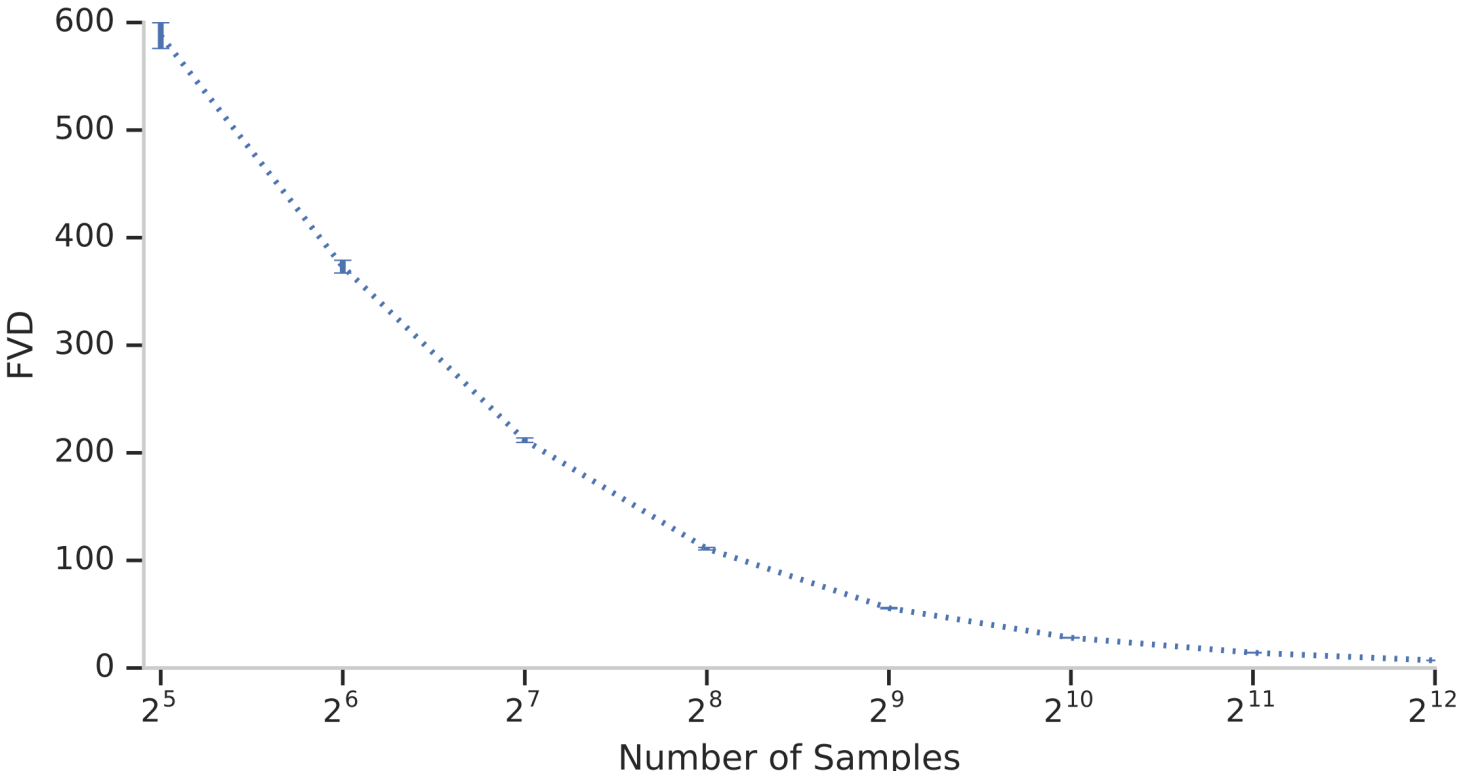


**図4**：BAIR ビデオプッシングデータセットから無作為に抽出された2つの重複しないビデオのサブセット間の FVD（エラーバーは50回の試行における標準誤差）

# C 人的評価

CDNA [22]、SV2P [6:1]、SVP-FP [23]、SAVP [11:1] を BAIR データセットで学習した。可能なハイパーパラメータの設定を幅広く使用し、学習の様々な段階でモデルパラメータを含めることで、3000以上の異なるモデルを得ることができた。生成されたビデオは、コンテキストの2フレームと進行中の14の出力フレームを組み合わせて得られる。先行研究に従い、各入力コンテキスト（コンディショニングフレーム）に対して100個の動画を生成し、これらの動画の中でフレーム平均した値が最も良いものを返すことで、PSNR と SSIM スコアを求める。FVD を計算する際、ターゲット分布を推定するために256のビデオシーケンス（モデルは未見）を考慮する。

私たちは、訓練されたモデルのさまざまなサブセットに基づいて、いくつかの人的実験を行う。特に、2つの異なるシナリオに従ってモデルを選択する。

**One Metric Equal** 単一の指標によって区別できないモデルを検討し、人間の評価者や他の競合する指標が、生成されたビデオの品質という点で、これらのモデルをどの程度区別できるかを評価する。我々は、与えられたメトリックに対してほぼ等しい値を持ち、そのメトリックの全体的な分布の最良の四分位に近い10個のモデルを選択する。すなわち、検討中のメトリックによって決定されるように、残りのモデルの25%より悪く、残りのモデルの75%より良いモデルを選択する。各指標について、最初の4〜5桁の有効数字まで同じ値を持つモデルを選ぶことができた。

**One Metric Spread** 2つ目の設定では、ある指標に対して大きく異なるスコアを持つモデルが、人間が判断した生成ビデオの主観的品質とどの程度一致するかを検討する。我々は、その指標の全体分布の10%と90%のパーセンタイルの間に等距離にある10モデルを選んだ。この場合、（メトリックが正確であれば）検討対象のモデル間で、生成されたビデオの品質に関して明確な違いがあるはずであり、競合するメトリックと比較して、検討対象のメトリックの人間の判断との高い一致を示唆している。

人間による評価には、各選択モデルから生成された3つのビデオを使用した。人間の評価者は、2つのモデルのビデオを見せられ、どちらがより良く見えるかを識別するか、あるいはその品質が区別できないと報告するよう求められる。各ペアの比較ビデオは、最大3人の独立した評価者に見せられ、3人目の評価者は、最初の2人の評価者の意見が一致しない場合にのみ質問された。評価者は、どちらのビデオが優れているかについて、事前に何も知らされていなかった。我々は、これらの人間の評価と、検討中の様々なメトリクスによって決定された評価との対応関係を計算した。

| Metric | eq.FVD | eq.SSIM | eq.PSNR | eq.KVD | | spr.FVD | spr.SSIM | spr.PSNR | spr.KVD |
|---|---|---|---|---|---|---|---|---|---|
| FVD | N/A | **74.9%** | **81.0%** | **63.0%** | | **71.9%** | 58.4% | 63.5% | **63.1%** |
| SSIM | 51.5% | N/A | 44.6% | 43.6% | | 61.8% | 51.2% | 45.9% | 50.2% |
| PSNR | **56.3%** | 21.4% | N/A | 48.8% | | 54.1% | 37.0% | 44.8% | 54.1% |
| KVD | 40.6% | 70.4% | 73.8% | N/A | | 69.4% | 56.8% | **63.8%** | 59.1% |
| Avg.FID | 35.5% | 71.2% | 52.0% | 43.5% | | 62.4% | **62.7%** | 57.6% | 51.2% |
| | | | | | | | | | |
| raters | 79.3% | 77.8% | 84.4% | 74.3% | | 83.3% | 69.9% | 72.5% | 74.1% |

**表3**：ある指標（eq.）に対して固定値を持つモデル、または広い範囲（spr.）に値を広げたモデルを考慮した場合の、人間の判断と指標の一致度。FVDは、主観的な品質に基づいて生成されたビデオを判定するのに優れている。

表3は、KVD と平均 FID メトリクスの人間による評価結果と、主要論文にある他のすべてのメトリクスの結果を示している。平均 FID は、フレシェ距離を計算する前に、各フレームの Inception embedding を平均することによって計算される。

- **平均 FID** 我々は、Avg FIDはFVDと比較して、ほとんどのシナリオで著しくパフォーマンスが悪いことを発見した。SSIMでは、わずかに良いパフォーマンスを達成している。これは、（SSIMによって決定された各デシルからサンプリングされた）幅広いビデオを主にフレームレベルの品質に基づいて判断することが好ましいこ

とを示唆している。一方、同じような画質の動画をSSIMで比較した場合、フレームレベルの画質だけでなく、時間的なコヒーレンスで判断した方が有利であり、平均 FIDの方が劣ることがわかった。
- **KVD** KVD は FVD と高い相関があるが（spearman: 0.9）、ほとんどのシナリオにおいて、人間の判断との一致という点では FVD より若干悪い。

一般的に、表3から、FVD はテストされた他のすべてのメトリクスと比較して優れた選択であると結論づけることができる。eq.FVDとspr.FVDで得られた結果は、ユーザーが実際にFVDをどのように体験するかを決定するため、重要な意味を持つ。spr.FVD の列から、他のどの指標も FVD によって引き起こされたランキングを改善することはできないと結論づけることができる。eq.FVD の列は、他のどの指標も FVD の点で同等である優れたモデルを確実に区別することはできないと教えてくれる。

表3は、評価者間の一致率も示している。これらは、あるビデオペアについて、最初の2人の評価者が合意した比較の割合として計算され、すべての比較で平均され、最終的なパーセンテージが得られる。ほとんどの場合、評価者は生成されたビデオの比較に自信を持っていることがわかる。

# D FVD の解決法

付録Bの結果は、一定のサンプルサイズであればFVDの結果の再現性が高いことを示しているが、FVDのわずかな差がどの程度まで意味のあるものであるかは考慮していない。この疑問に答えるため、人間の評価者は、200/400（base200/base400）のFVDを持つ無作為に選ばれたモデルによって生成されたビデオと、10、20、50、100、200、300、400、500FVDポイント悪いモデルによって生成されたビデオを比較するよう求められた。それぞれの場合において、これらのFVDスコアで利用可能なモデルから5つのモデルを選択し、各モデルについて3つのビデオを作成し、合計1,800の比較を行った。あるビデオの比較について、評価者は2つのビデオのどちらがよりよく見えるか、あるいは同じような品質かどうかを判断するよう求められた。それぞれのペアについて、最大3人の人間の評価者に意見を求めた。

図5を見ると、FVDの差が50より小さい場合、人間の評価者との一致度はランダムに近く（しかし決して悪化しない）、2つのモデルのFVDが50ポイント以上離れると急激に増加することがわかる。したがって、50FVD以上の差は、通常、人間が知覚できる生成ビデオの品質の差に対応すると結論づけることができる。

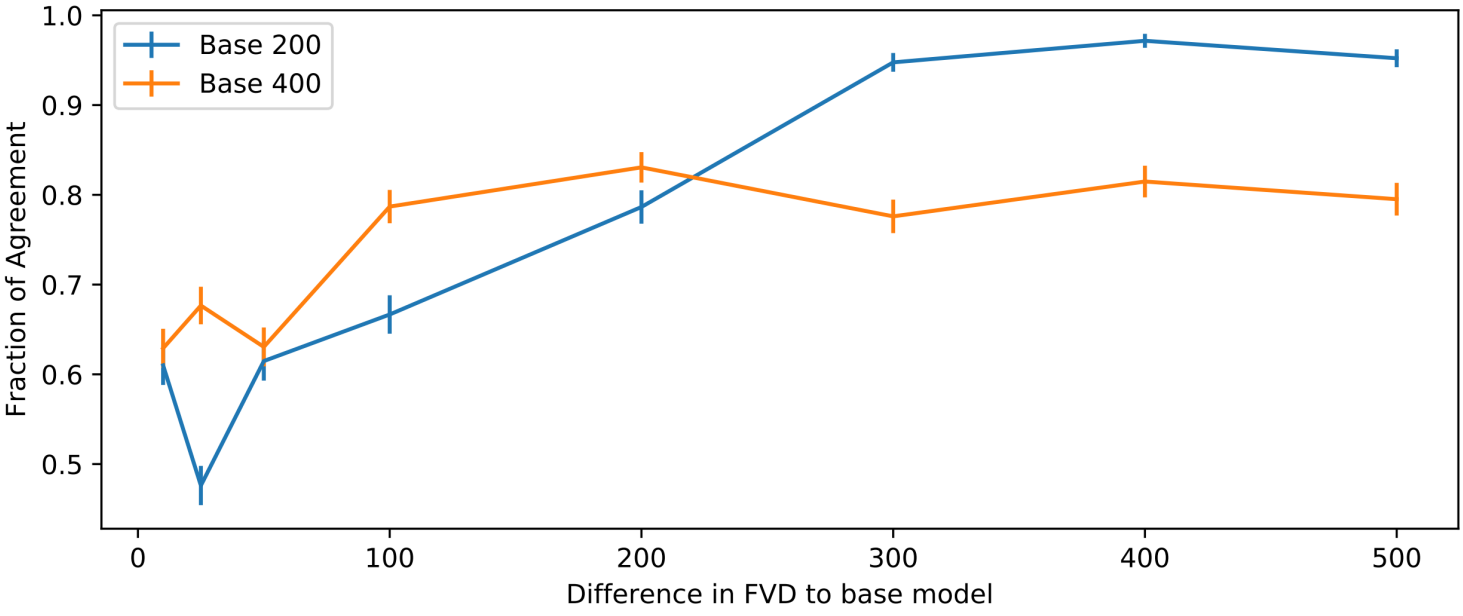

**図5**：2つのモデルのどちらが優れているかについて、FVDに同意した人間の評価者の割合（モデル間のFVDの差の関数として）。エラーバーは標準誤差であり、ビデオペアが同程度の品質であると評価者が判断した場合は、FVDに同意していないとカウントされる。

# 参考


1. Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. ICLR, 2018. ↩
2. Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. arXiv, 2018. ↩
3. Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. CVPR, 2018. ↩
4. Emanuela Haller and Marius Leordeanu. Unsupervised object segmentation in video by efficient selection of highly probable positive features. ICCV, 2017. ↩
5. Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In ICML, pp. 430–438, 2016. ↩
6. Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. ICLR, 2017. ↩ ↩
7. 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. NIPS, 2017. ↩ ↩
8. Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004. ↩
9. Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. Telecommunication Systems, 2012. ↩
10. D.C Dowson and B.V Landau. The frechet distance between multivariate normal distributions. Journal of Multivariate Analysis, 12(3):450 – 455, 1982. ISSN 0047-259X. ↩
11. Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. arXiv, 2018. ↩ ↩
12. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. CVPR, 2016. ↩
13. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. CVPR, 2017. ↩ ↩
14. Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Nikolai Yakovenko, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In NIPS, pp. 1152–1164, 2018. ↩
15. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. arXiv, 2017. ↩ ↩
16. Mikołaj Binkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. ICLR, 2018. ↩ ↩ ↩
17. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander Smola. A kernel two-sample test. Journal of Machine Learning Research, 13(Mar):723–773, 2012. ↩

18. Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In Conference on Robot Learning, pp. 344–356, 2017. ↩ ↩
19. Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In High Performance Computing in Science and Engineering 12. Springer, 2013. ↩ ↩
20. L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. ICLR, 2016. ↩
21. Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. NIPS, 2018. ↩
22. Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. NIPS, 2016. ↩
23. Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. ICML, 2018. ↩