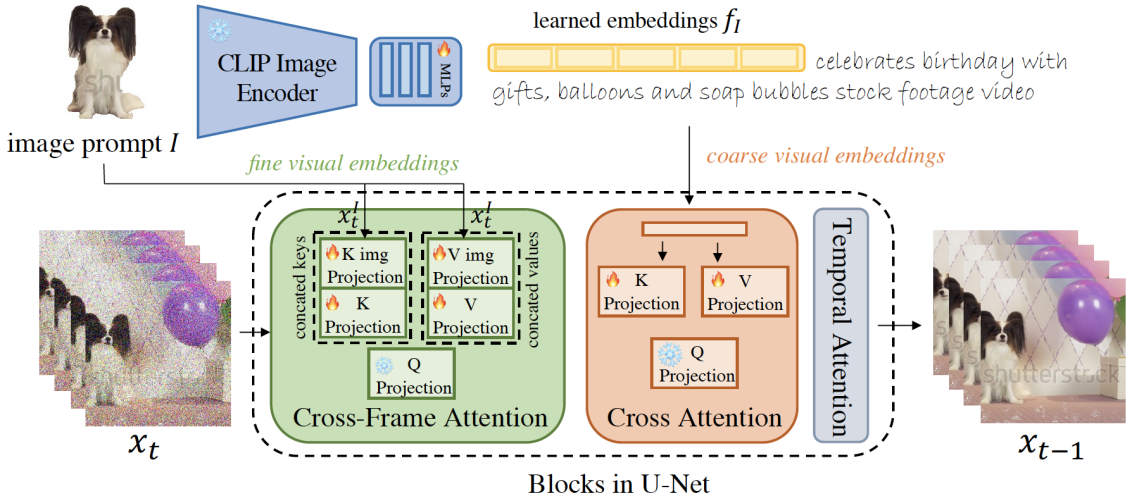
VideoBooth: Diffusion-based Video Generation with Image Prompts

https://arxiv.org/abs/2312.00777

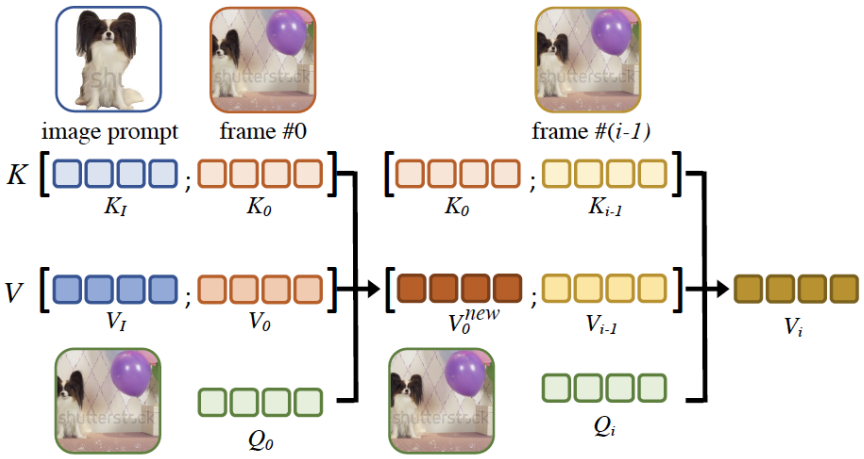
- アブスト
  - 画像をプロンプトとした動画生成モデル、VideoBoothを提案
    - 画像のEmbeddingを大まかから詳細な情報へと段階的に処理する
      - 画像のEncoderを通したEmbeddingでは画像の全体的な特徴を捉えるのに対し、提案したattention injection moduleによるEmbeddingでは、画像のプロンプトに対する詳細かつマルチスケールの情報を提供する
    - attention injection moduleによるマルチスケールのEmbeddingは、様々なcross-frame attention層のkeyやvalueとして入力される。この追加の空間情報は最初のフレームの特徴を捉えており、それが残りのフレームに伝播され、時間的な一貫性を維持する
- イントロ
  - text-to-imageタスクにおいて画像をプロンプトに用いる方法はいくつかある
    - 参照画像を用いてfew-shotで一部のパラメータをファインチューニングする
      - 複数の同様な参照画像を用意する必要がありコストが高い
    - プロンプト画像のembeddingをtext-to-imageのモデルに埋め込む
      - 推論時にファインチューニングをする必要がない
  - text-to-videoは時間的な一貫性も必要になるので、text-to-imageよりも難しい。そこでVideoBoothを提案

    - 2つのEmbedding
        - 画像のエンコーダを用いた大まかな画像の情報に関するEmbedding
            - 画像のEmbeddingによってテキストのEmbeddingの一部を代替する
        - attention injectionによる詳細な画像の情報に関するEmbedding
            - text-to-video生成に関して、画像プロンプトをマルチスケールの潜在変数にすることで制御する
    - プロンプト画像をCLIPによって画像の特徴量にする。その後、テキストのEmbedding空間に落とし込まれ、一部のテキストのEmbeddingの代わりになる
    - また、マルチスケールのプロンプト画像の潜在表現が様々なcross-attention層のkeysやvaluesに用いられる
        - 特に最初のフレームの生成に用いられる
            - 最初のフレームに関して更新されたvaluesを用いて残りのフレームの生成に用いる
- 提案手法
    - モデル構造



    - coarse visual embeddings
        - プロンプト画像によるembeddingとテキストによるembeddingを混ぜる

        - 例えば、プロンプト画像として、Papillon dogの画像があり、テキストプロンプトとして、"Papillon dog celebrates birthday with gifts"とあった場合、テキストEmbeddingのPapillon dogの部分をPapillon dogの画像のEmbeddingで代替する
    - fine visual embeddings
        - プロンプト画像をまず、Stable DiffusionのVAEで潜在変数にする
            - 中間ステップでは、入力の潜在変数はノイズを含んでいるため、このままの状態で条件付けるとドメインの不一致が起きる。そのため、拡散過程(forward process)によってプロンプト画像の潜在変数にノイズを付与する
            - cross-attentionは、生成フレームの時間的な一貫性を実現する
                - 一つ一つのフレームに対して、最初のフレームと直前のフレームの特徴量をconcatしたものをkeysとvaluesに用いる
                - 最初のフレームのvaluesについては、画像プロンプトを用いてアップデートしたものを用いる



    - coarse-to-fine training strategy
        - coarse-to-fine mannerで訓練する
            - 大まかなプロンプト画像の情報について学習させるために、cross-attentionのパラメータを先に学習させる

        - その後に、cross-frame attention層にプロンプト画像のEmbeddingを挿入するためにattention injection moduleを学習させる
        - これらを一緒に学習させると、fine attention injection moduleがプロンプト画像に関する詳細な情報をリークしてしまい、coarse encoderが何も学習しなくなってしまう
- VideoBooth Dataset
    - WebVid datasetを用いて作成
        - 最初のフレームに対して、Grounded-SAMを用いて物体を検出してそれらをプロンプト画像として用いる
- 評価指標
    - CLIP-Text metric
        - CLIPを用いてテキストEmbeddingと生成したフレームのEmbeddingのコサイン類似度を測る
    - CLIP-Image metric
        - CLIPを用いてプロンプト画像のEmbeddingと、生成したフレームのEmbeddingのコサイン類似度を測る
    - DINO similarity metric
        - ViT-S/16モデルを用いてプロンプト画像と生成フレームを特徴量に変換して、DINOを用いて類似度を測る