8/11

* Testing multi conditional guidance free image generation for pre-trained celeba model
* Testing clip + parse
  + Found that parse’s norm and grad norm is a lot bigger in terms of magnitude than clip
    - Thus, increase the weighing factor for clip, Nc (Set Nc to be 1000), so  
      Nc : Np = 1000:1
  + DEGRADED: Found that parse’s norm and grad norm continues to overpower clip’s
    - Have to add its contribution (grad norm) only at later stages, when time step > 200
      * So clip can guide the image generation at earlier stages
  + Found that if clip description is similar to the parse image (in terms of description), there will be better results. Parse gives directional/positional information while clip gives detailed facial descriptions
* Testing clip + faceid (arcface)
  + Nc:Na = 1:1 or 1:10 works well
  + Same, if clip description is similar to the parse image, there will be better results
* Testing clip + landmark
  + Nc:Nl = 1:1 or 1:10 works well
  + Same, if clip description is similar to the parse image, there will be better results
* Testing clip + sketch
  + Nc:Ns = 10:1 works well
  + Same, if clip description is similar to the parse image, there will be better results
* Final weights, A white background with black symbols

  Description automatically generated
  + Nc = 1, Np = 1/1000, Na = 1, Nl = 1, Ns = 1/10
* TODO: improve the weighted norm formula, with covariance information

15/12

Things done:

1. Implemented the multi conditional sampling (not implemented in code base, although formula is proposed)
2. Experimented on weighing parameter for each condition to get decently good results
   1. Methods like distance-based normalisation do not work well thus far
   2. Settled with empirical method, found suitable ratios

Problems:

1. Not yet improve the proposed multi conditional sampling formula
2. Reasons for poor results not yet confirmed
   1. How to pick suitable weights (not empirically)
   2. Is it really because of the dependence between conditions (?) have to account for covariance because we cannot simply add them up?

Potential Improvements:

20/12/24

* TODO:
  + Write Theory
  + Eval metrics

27/12/2024

Eval on multi conditions:

* L2 done

30/12/2024

Eval on multi conditions (all done)

* FID done

"clip\_landmark\_ref2334jpg" / "int\_terms"

* clip -> mean L2 distance = 12.077734375
* clip -> mean FID distance = 145.9729835993993
* landmark -> mean L2 distance = 0.029189802706241608
* landmark -> mean FID distance = 0.25196502906482793

"clip\_landmark\_ref2334jpg" / "no\_int\_terms"

* clip -> mean L2 distance = 12.107265625
* clip -> mean FID distance = 146.71101829213083
* landmark -> mean L2 distance = 0.029059896245598793
* landmark -> mean FID distance = 0.25966718652546983

2/1/2025

Done:

* Report research
* Report planning

TODO:

* Possible other kernels implementation
* Other conditions combi generation of images for evaluation
  + Eg clip + face id + landmark/segmentation map

"clip\_parse\_ref294jpg" / "no\_int\_terms"

* clip -> mean L2 distance = 12.293125
* clip -> mean FID distance = 151.27000188378273
* parse -> mean L2 distance = 3.8397459983825684
* parse -> mean FID distance = 0.006183114888091701

"clip\_parse\_ref294jpg" / “int\_terms"

* clip -> mean L2 distance = 12.29828125
* clip -> mean FID distance = 151.3704585654553
* parse -> mean L2 distance = 3.840965747833252
* parse -> mean FID distance = 0.005401201414072836