APPENDIX

A IMPLEMENTATION DETAILS

Score Model: We use the implementation of the score model from Song et al. (2021) using the variance-exploding configuration, in particular the CIFAR10 configuration provided on their codebase. We augment the score-model with an Encoder E_ϕ which is implemented as the Wide-ResNet architecture (Zagoruyko & Komodakis, 2016) that maps the input with time embeddings to a vector in $\mathbb{R}^d$, where d is the dimensionality of the latent space and is set to 128 unless otherwise specified. The time embeddings for the encoder model are implemented in the exact same way as for the score inputs, as outlined in Song et al. (2021). We use the learning rate of 2×10^{-4} to optimize the score network.

Downstream Model: For Multi-Layer-Perceptron (MLP) based Classification model, we consider a network with a single hidden layer, ReLU activation function, and 512 neurons. For the Recurrent Neural Network (RNN) model, we use a GRU with 256 hidden units and for the transformer system, we use a Multi-Head attention system with 4 heads and two layers, with weight sharing between the layers. For our attention profile based analysis settings, we consider the same Multi-Head attention system but only use a single layer instead of two, as it allows to make the score (averaged over heads) more interpretable.

We train all the downstream models with dropout of 0.25 and perform hyperparameter optimization for the learning rate over the set $\{0.001, 0.00075, 0.0005, 0.00025, 0.0001, 0.00005\}$. In particular, we found the hyperparameter optimization important when considering the granularity analysis.

B CIFAR10, CIFAR100 AND MINI-IMAGENET

We train the score model for 70,000 iterations and then the downstream models for 100 epochs. For the performance of the models, we use a $2-$layered Transformer model while for attention score profiles, we use a single layered Transformer model.

C SYNTHETIC

We train the score model for 250,000 iterations and then the downstream models for 1500 epochs. Figure 11 shows some samples obtained from this dataset, showcasing the different features present as well as the diversity of these different features.

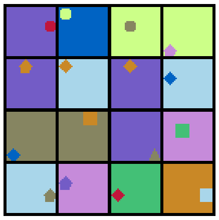


Figure 11: Samples from Synthetic Dataset

We additionally perform the Jensen-Shannon Divergence analysis between different features for different granularities, as well as visualize the attention score profiles for the different granularities as well. Furthermore, we do the same analysis with both the types of encoders; VDRL and DRL.

The corresponding plots for the attention score profiles are present in Figures 12 - 17 for different latent space dimensionalities, different granularities and the different types of encoding schemes (VDRL and DRL). Further analysis into the performance on different features with different granularities and dimensionalities can be found in Figure 18.

D COLORED MNIST

We train the score model for 250,000 iterations and then the downstream models for 1500 epochs. Figure 26 shows some samples obtained from this dataset, showcasing the different features present as well as the diversity of these different features.

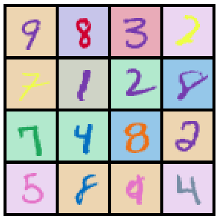
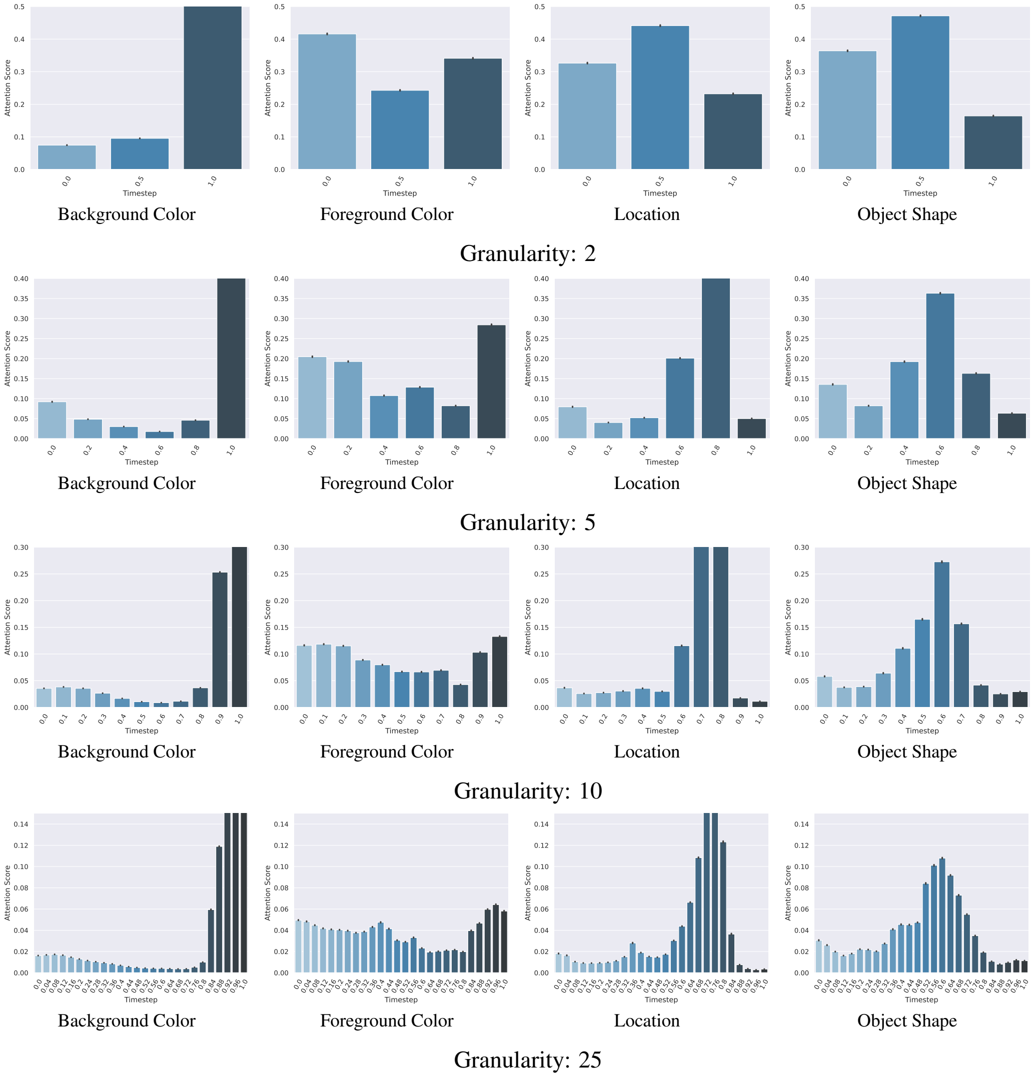


Figure 26: Samples from Colored-MNIST Dataset

We additionally perform the Jensen-Shannon Divergence analysis between different features for different granularities, as well as visualize the attention score profiles for the different granularities as well. Furthermore, we do the same analysis with both the types of encoders; VDRL and DRL.



Background Color Foreground Color Location Object Shape

Granularity: 2

Background Color Foreground Color Location Object Shape

Granularity: 5

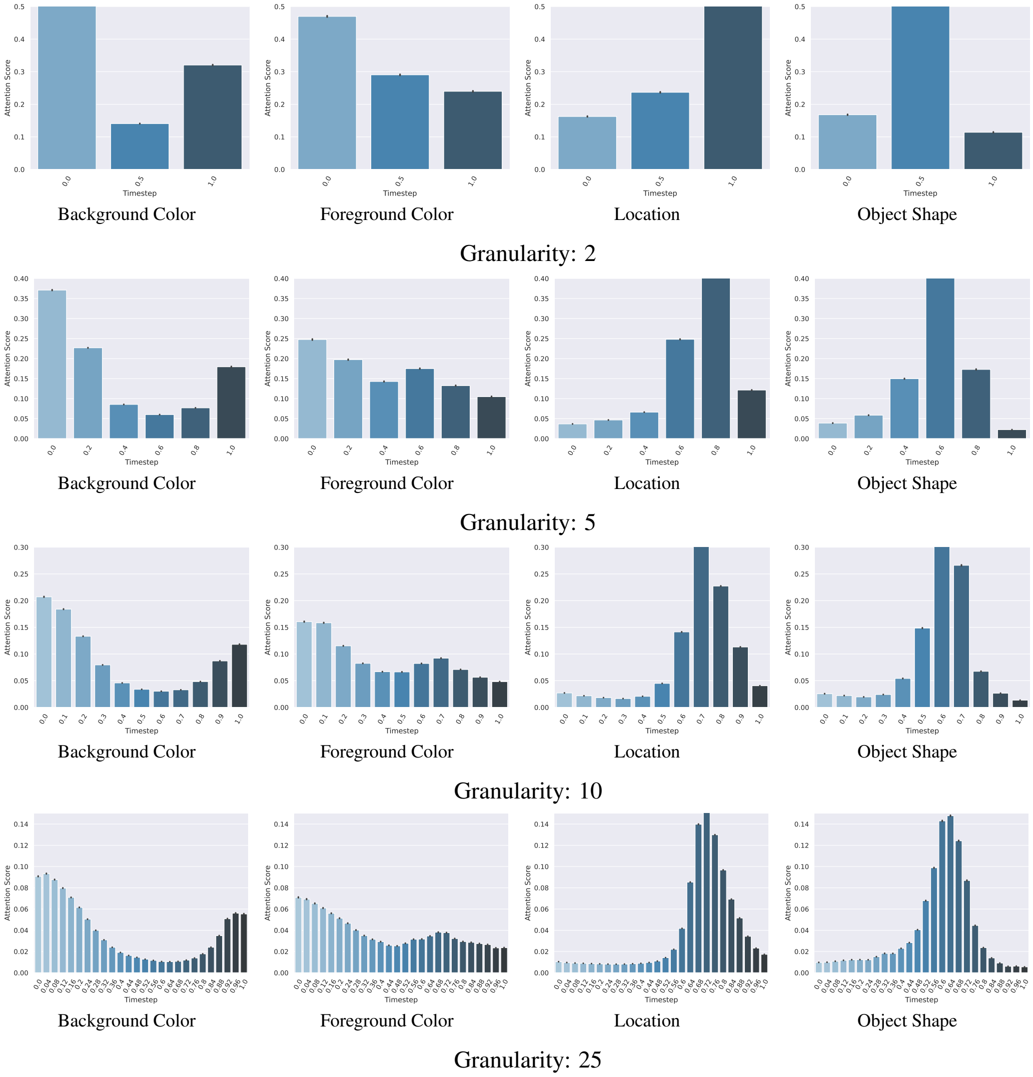
Background Color Foreground Color Location Object Shape

Granularity: 10

Background Color Foreground Color Location Object Shape

Granularity: 25

Figure 12: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 2 and the VDRL encoder.



Background Color Foreground Color Location Object Shape

Granularity: 2

Background Color Foreground Color Location Object Shape

Granularity: 5

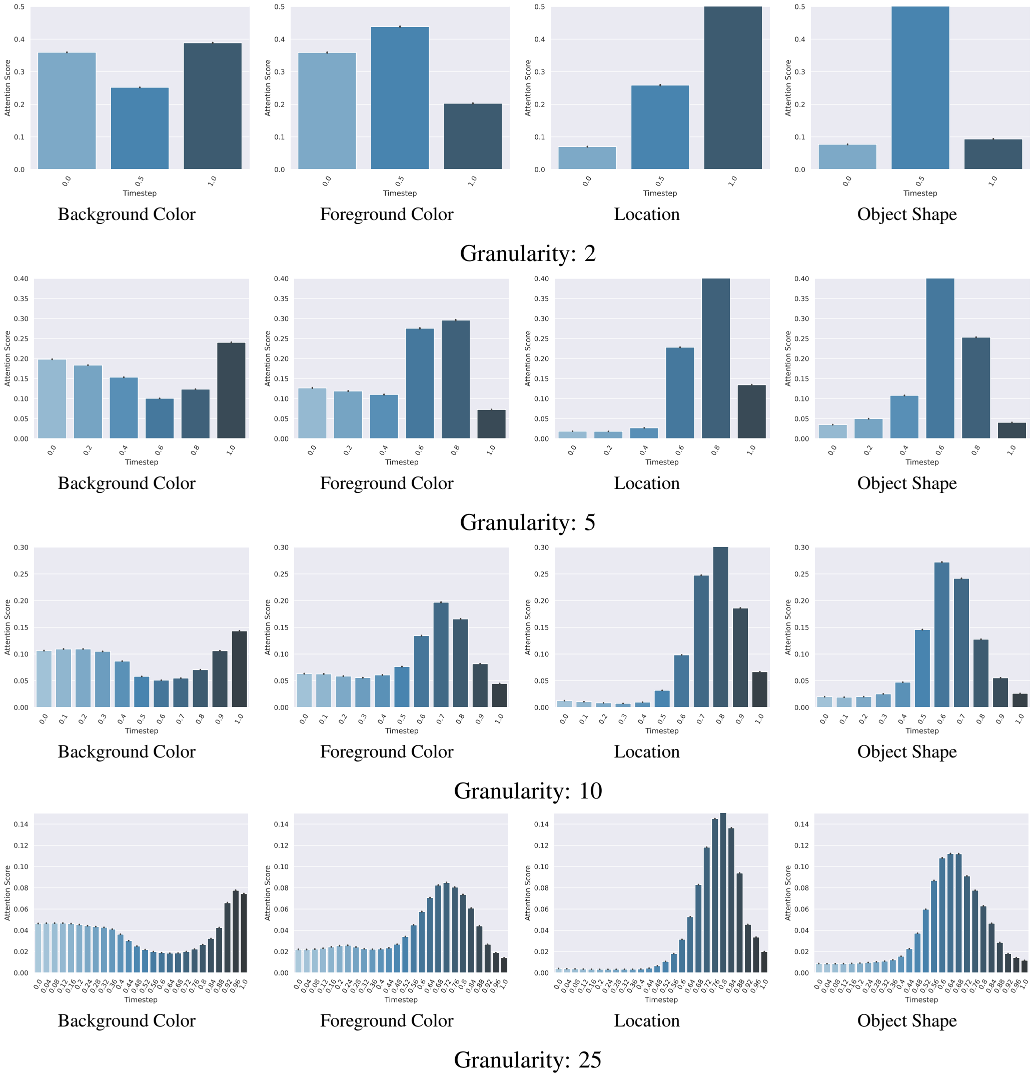
Background Color Foreground Color Location Object Shape

Granularity: 10

Background Color Foreground Color Location Object Shape

Granularity: 25

Figure 13: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 16 and the VDRL encoder.



Background Color

Foreground Color

Location

Object Shape

Granularity: 2

Background Color

Foreground Color

Location

Object Shape

Granularity: 5

Background Color

Foreground Color

Location

Object Shape

Granularity: 10

Background Color

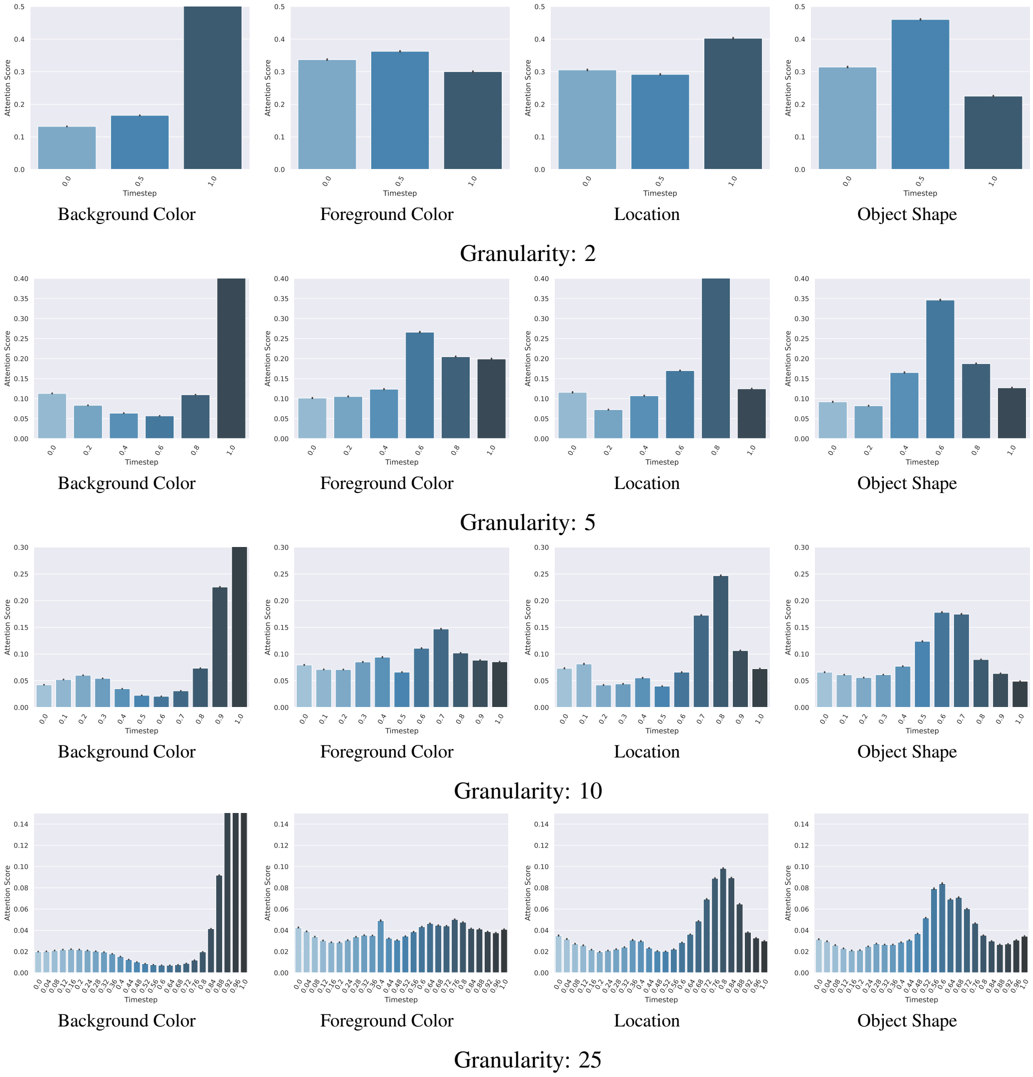
Foreground Color

Location

Object Shape

Granularity: 25

Figure 14: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 32 and the VDRL encoder.



Background Color Foreground Color Location Object Shape

Granularity: 2

Background Color Foreground Color Location Object Shape

Granularity: 5

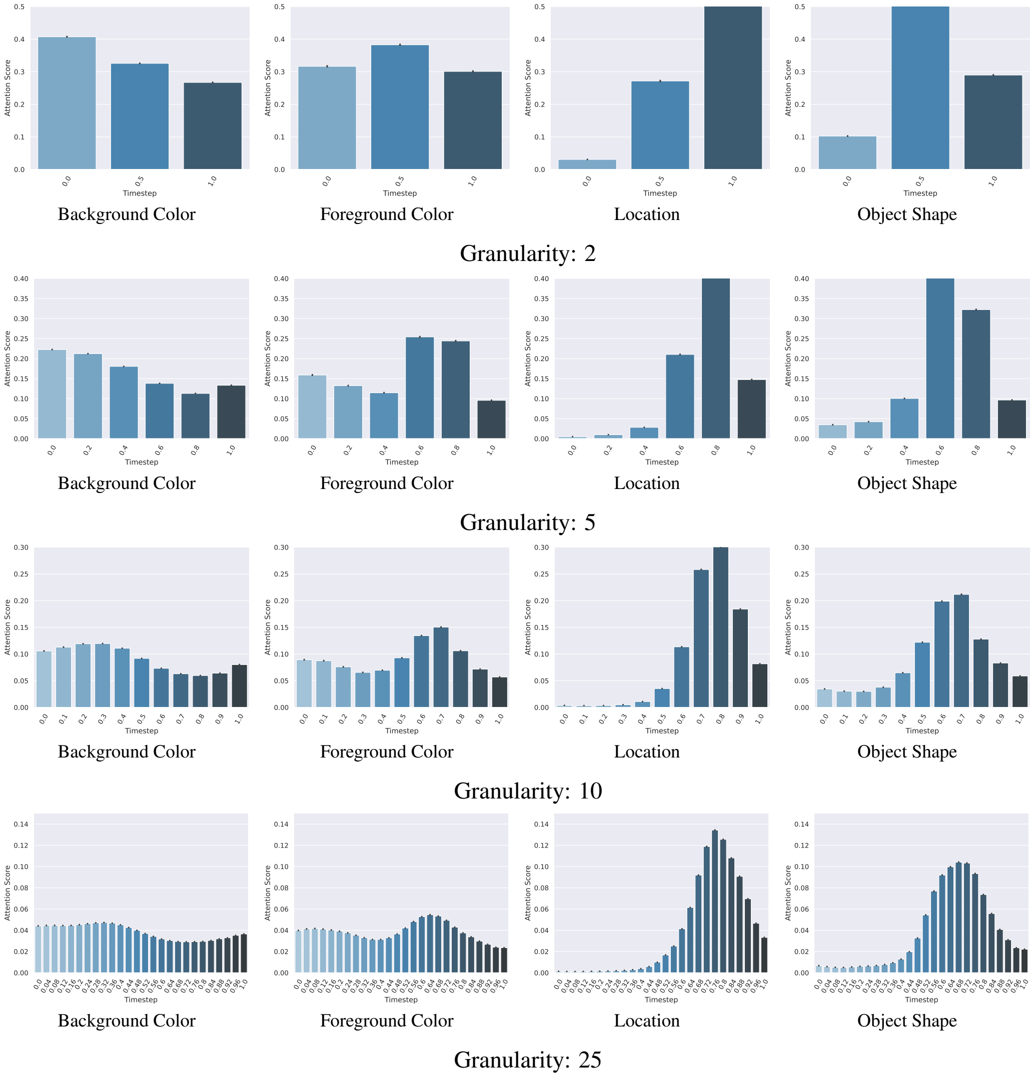
Background Color Foreground Color Location Object Shape

Granularity: 10

Background Color Foreground Color Location Object Shape

Granularity: 25

Figure 15: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 2 and the DRL encoder.



Background Color

Foreground Color

Location

Object Shape

Granularity: 2

Background Color

Foreground Color

Location

Object Shape

Granularity: 5

Background Color

Foreground Color

Location

Object Shape

Granularity: 10

Background Color

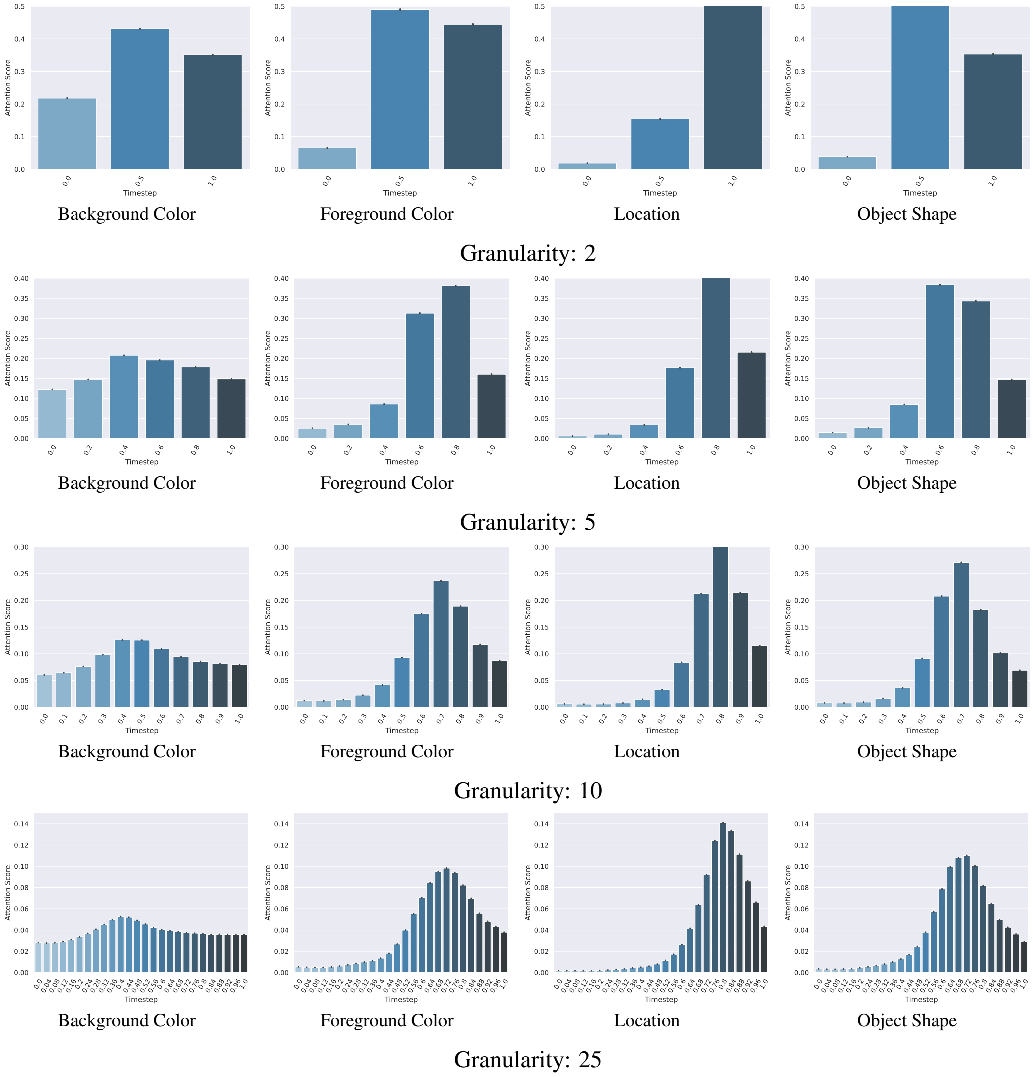
Foreground Color

Location

Object Shape

Granularity: 25

Figure 16: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 16 and the DRL encoder.



Background Color Foreground Color Location Object Shape

Granularity: 2

Background Color Foreground Color Location Object Shape

Granularity: 5

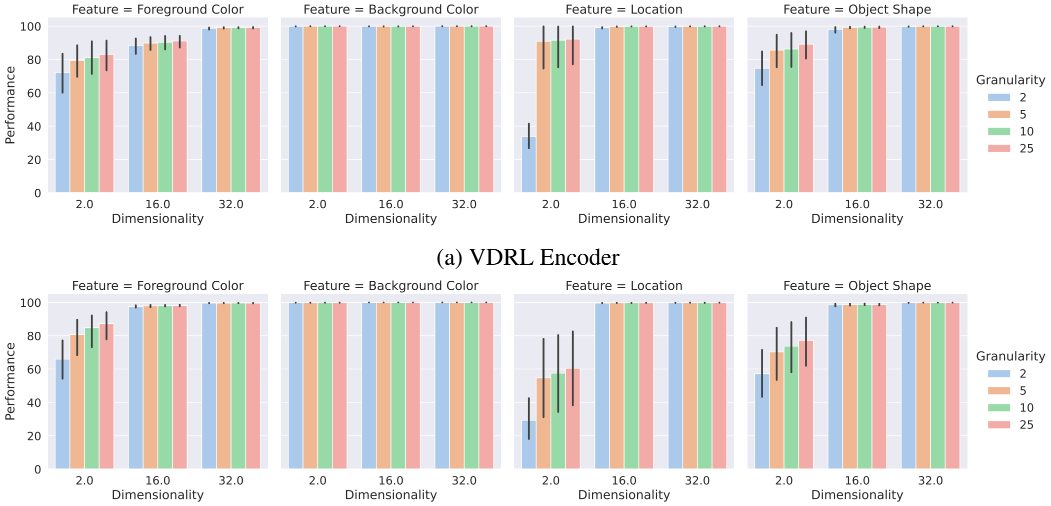
Background Color Foreground Color Location Object Shape

Granularity: 10

Background Color Foreground Color Location Object Shape

Granularity: 25

Figure 17: Attention score profiles for the synthetic dataset on the different features, using different granularities, with the dimensionality of the latent space as 32 and the DRL encoder.



(a) VDRL Encoder

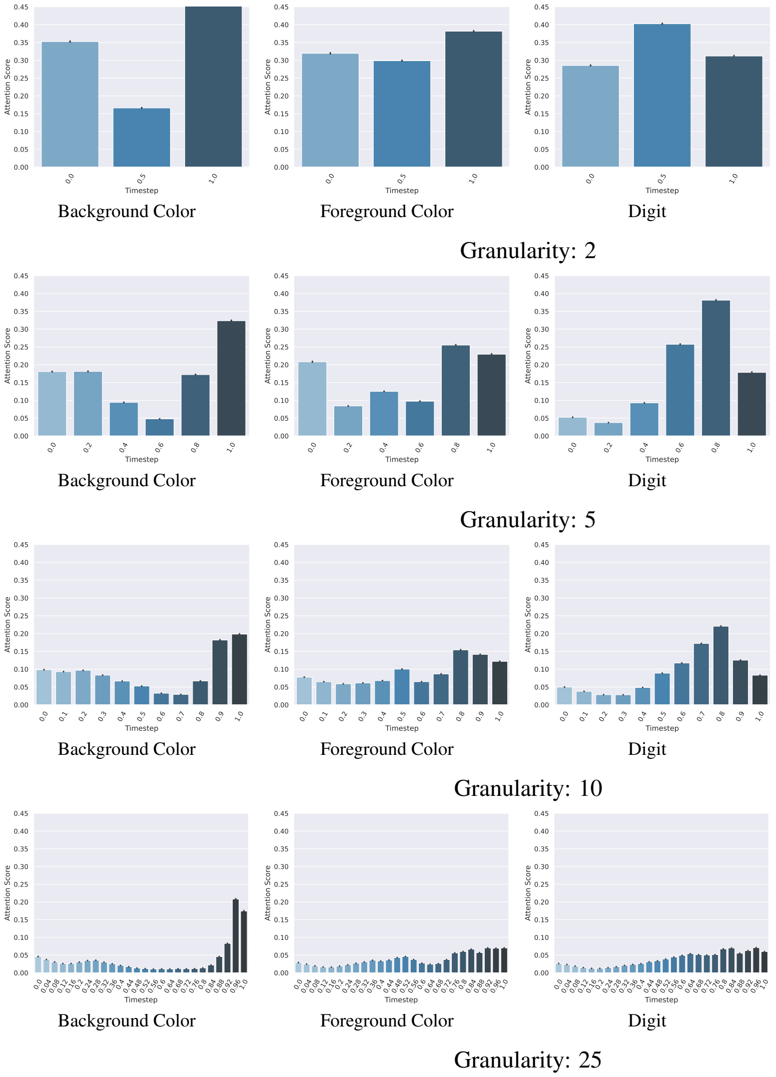
(b) DRL Encoder

Figure 18: Downstream performance plots for the Synthetic Dataset for different features, when the score model is trained with different latent dimensionality and the downstream models are trained with different granularities for discretization.

The corresponding plots for the attention score profiles are present in Figures 19 - 24 for different latent space dimensionalities, different granularities and the different types of encoding schemes (VDRL and DRL). Further analysis into the performance on different features with different granularities and dimensionalities can be found in Figure 25.

E CelebA

We train the score model for 250,000 iterations and then the downstream models for 100 epochs. We additionally perform the Jensen-Shannon Divergence analysis between different features for two different types of encoders; VDRL and DRL. The corresponding plots for these analysis, as well as for the attention score profiles and performances on different features, are present in Figures 27 - 30. The figures also enumerate the different attributes present in the dataset.



Background Color Foreground Color Digit

Granularity: 2

Background Color Foreground Color Digit

Granularity: 5

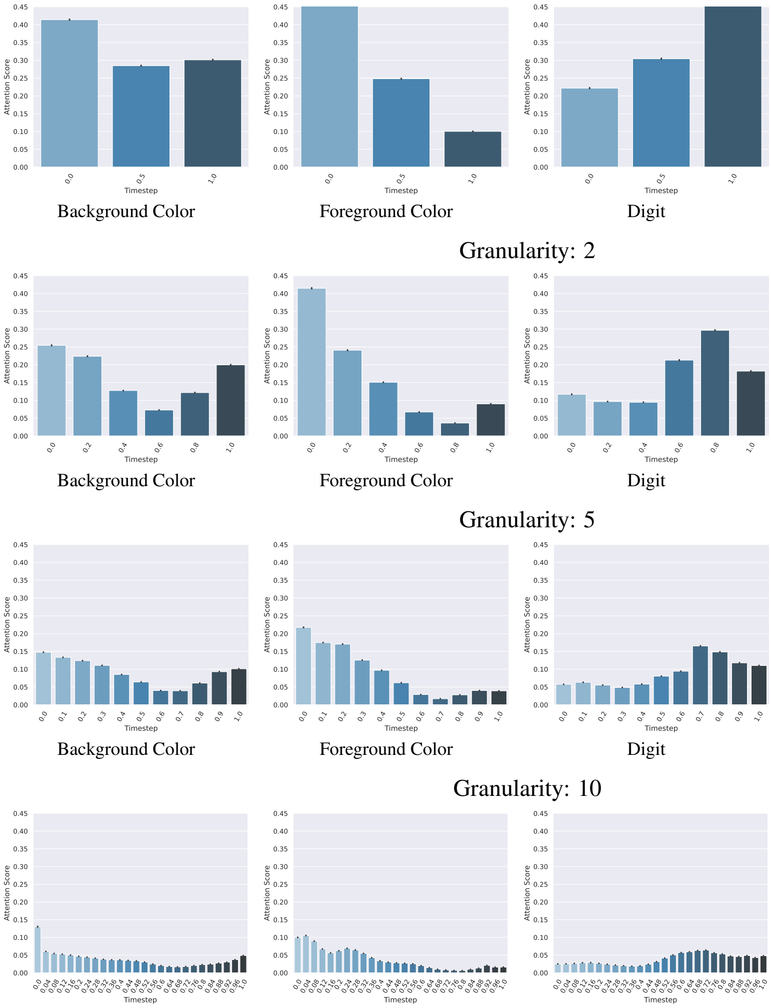
Background Color Foreground Color Digit

Granularity: 10

Background Color Foreground Color Digit

Granularity: 25

Figure 19: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 2 and the VDRL encoder.



Background Color

Foreground Color

Digit

Granularity: 2

Background Color

Foreground Color

Digit

Granularity: 5

Background Color

Foreground Color

Digit

Granularity: 10

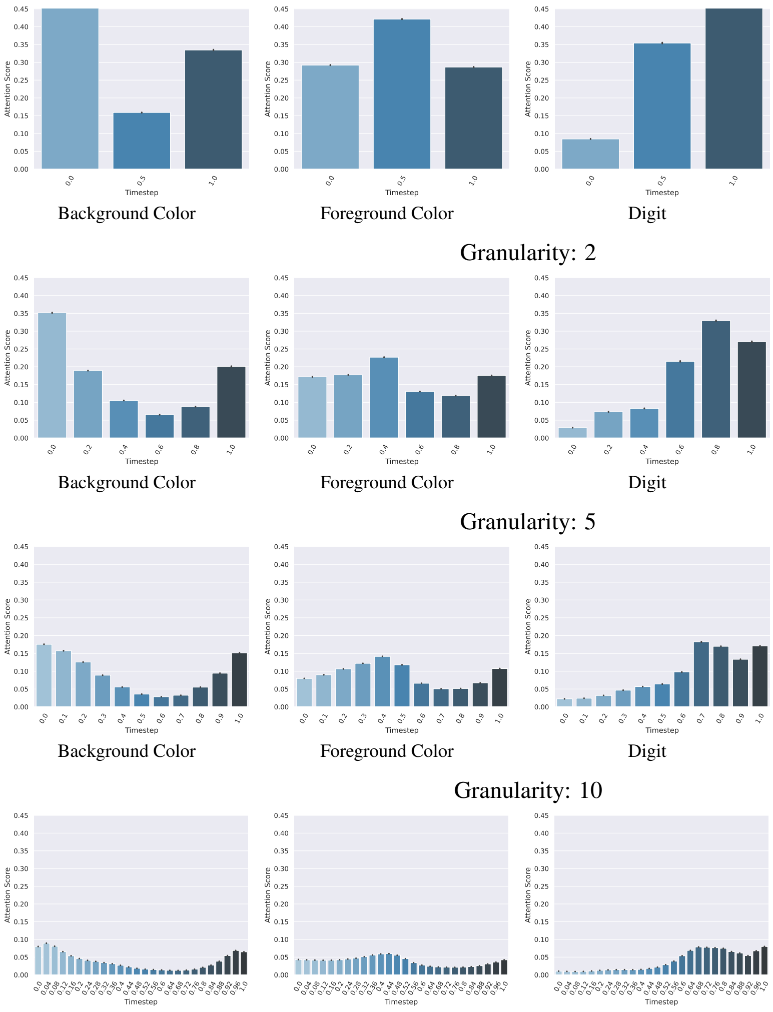
Background Color

Foreground Color

Digit

Granularity: 25

Figure 20: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 16 and the VDRL encoder.



Background Color Foreground Color Digit

Granularity: 2

Background Color Foreground Color Digit

Granularity: 5

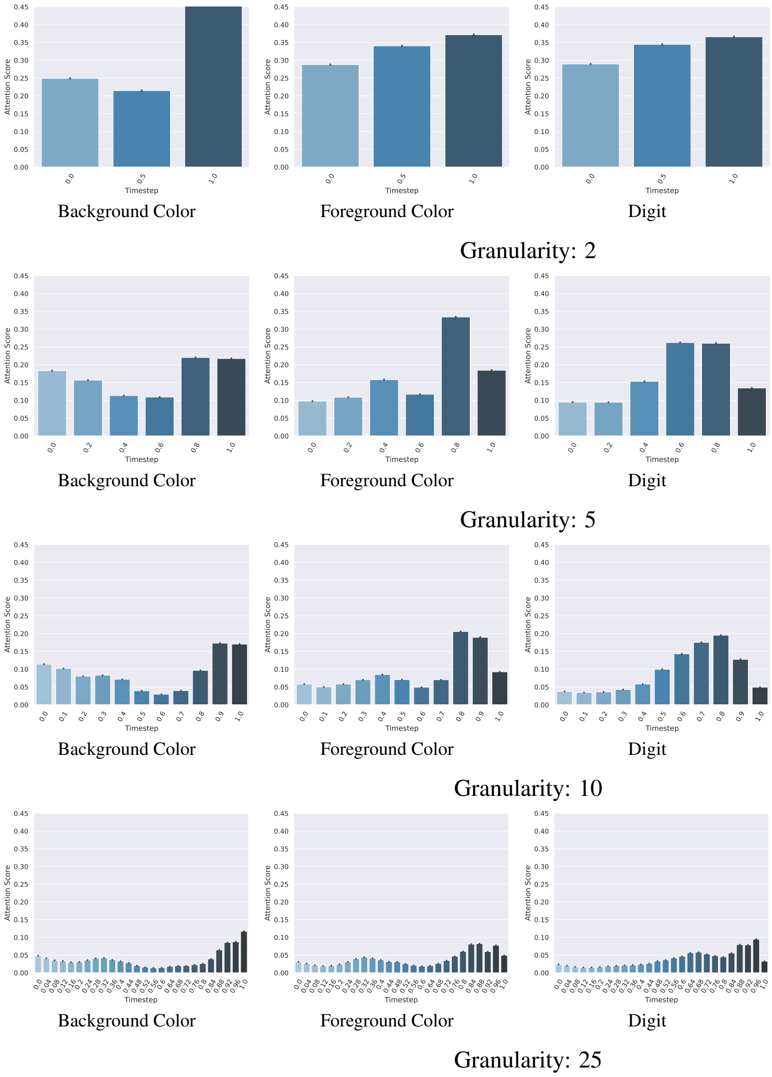
Background Color Foreground Color Digit

Granularity: 10

Background Color Foreground Color Digit

Granularity: 25

Figure 21: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 32 and the VDRL encoder.



Background Color Foreground Color Digit

Granularity: 2

Background Color Foreground Color Digit

Granularity: 5

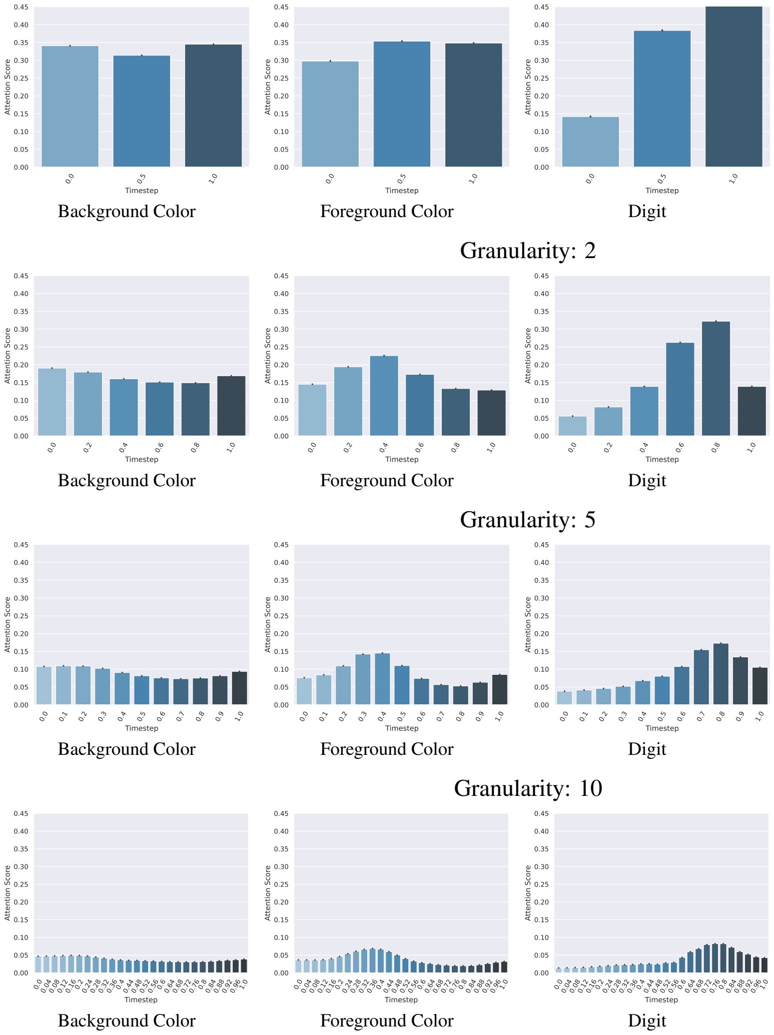
Background Color Foreground Color Digit

Granularity: 10

Background Color Foreground Color Digit

Granularity: 25

Figure 22: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 2 and the DRL encoder.



Background Color Foreground Color Digit

Granularity: 2

Background Color Foreground Color Digit

Granularity: 5

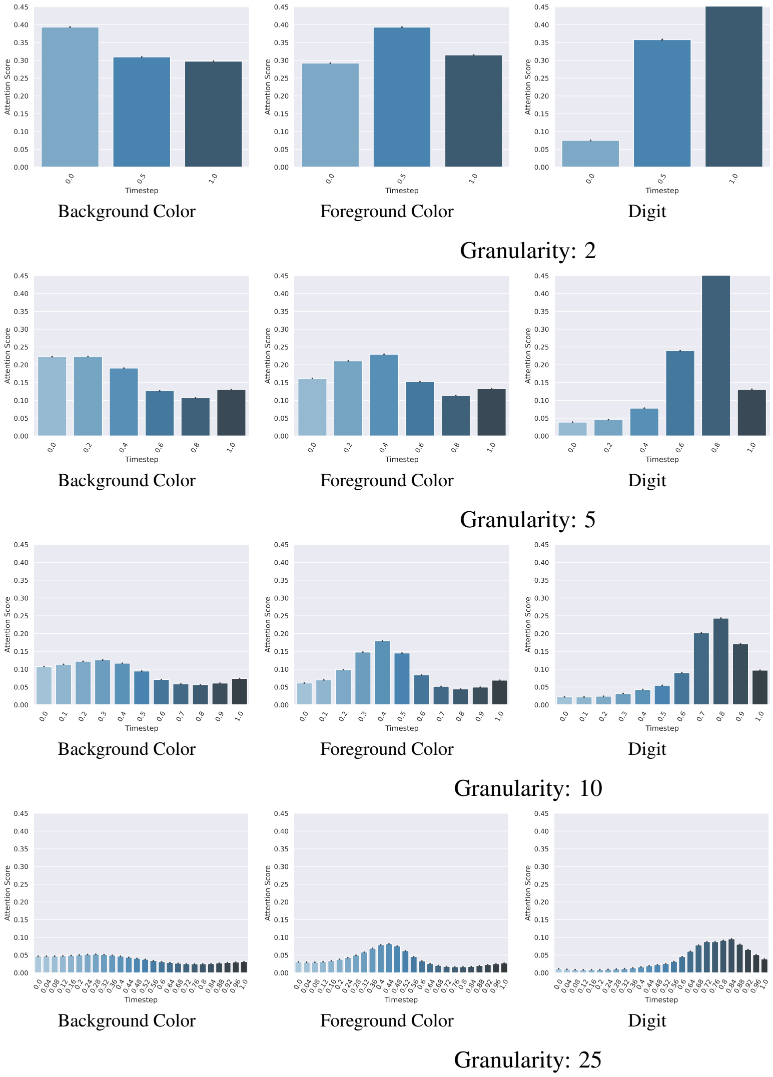
Background Color Foreground Color Digit

Granularity: 10

Background Color Foreground Color Digit

Granularity: 25

Figure 23: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 16 and the DRL encoder.



Background Color

Foreground Color

Digit

Granularity: 2

Background Color

Foreground Color

Digit

Granularity: 5

Background Color

Foreground Color

Digit

Granularity: 10

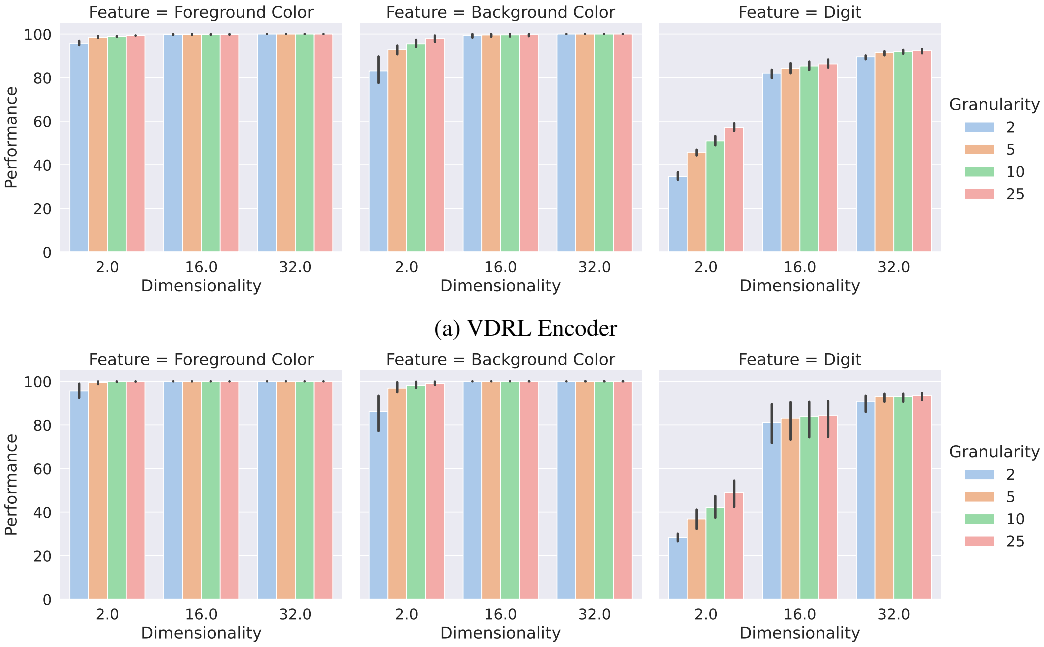
Background Color

Foreground Color

Digit

Granularity: 25

Figure 24: Attention score profiles for the Colored-MNIST dataset on the different features, using different granularities, with the dimensionality of the latent space as 32 and the DRL encoder.



(a) VDRL Encoder

(b) DRL Encoder

Figure 25: Downstream performance plots for the Colored-MNIST Dataset for different features, when the score model is trained with different latent dimensionality and the downstream models are trained with different granularities for discretization.

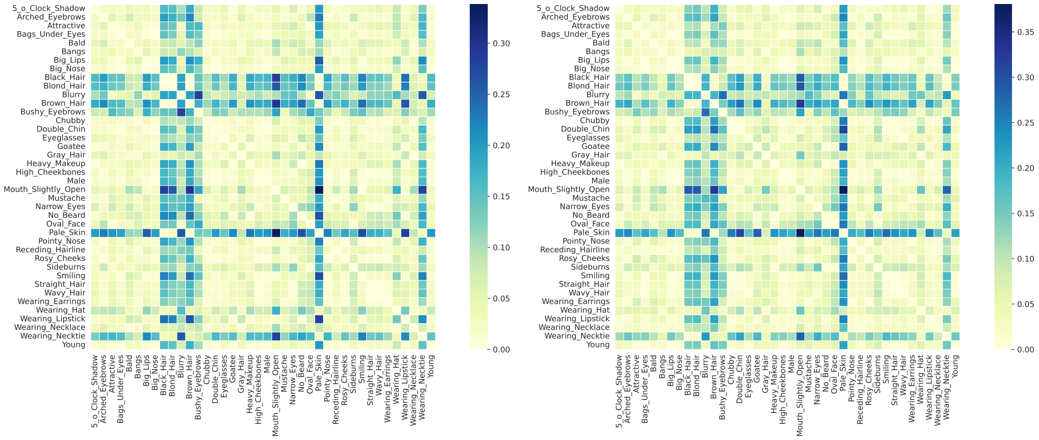


Figure 27: Jensen Shannon Divergence plot for the attention profiles for any pair of features in the *CelebA* dataset when using the *Left:* VDRL Encoder, and *Right:* DRL Encoder.

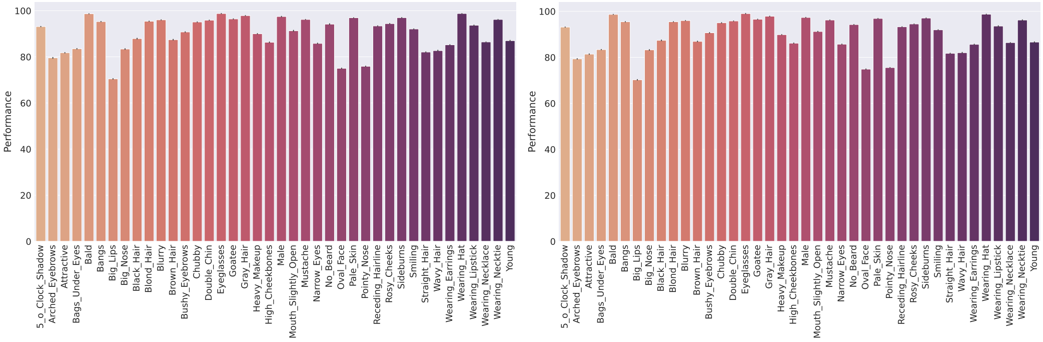


Figure 28: Downstream performance plots for the different features in the *CelebA* dataset when using the *Left:* VDRL Encoder, and *Right:* DRL Encoder.

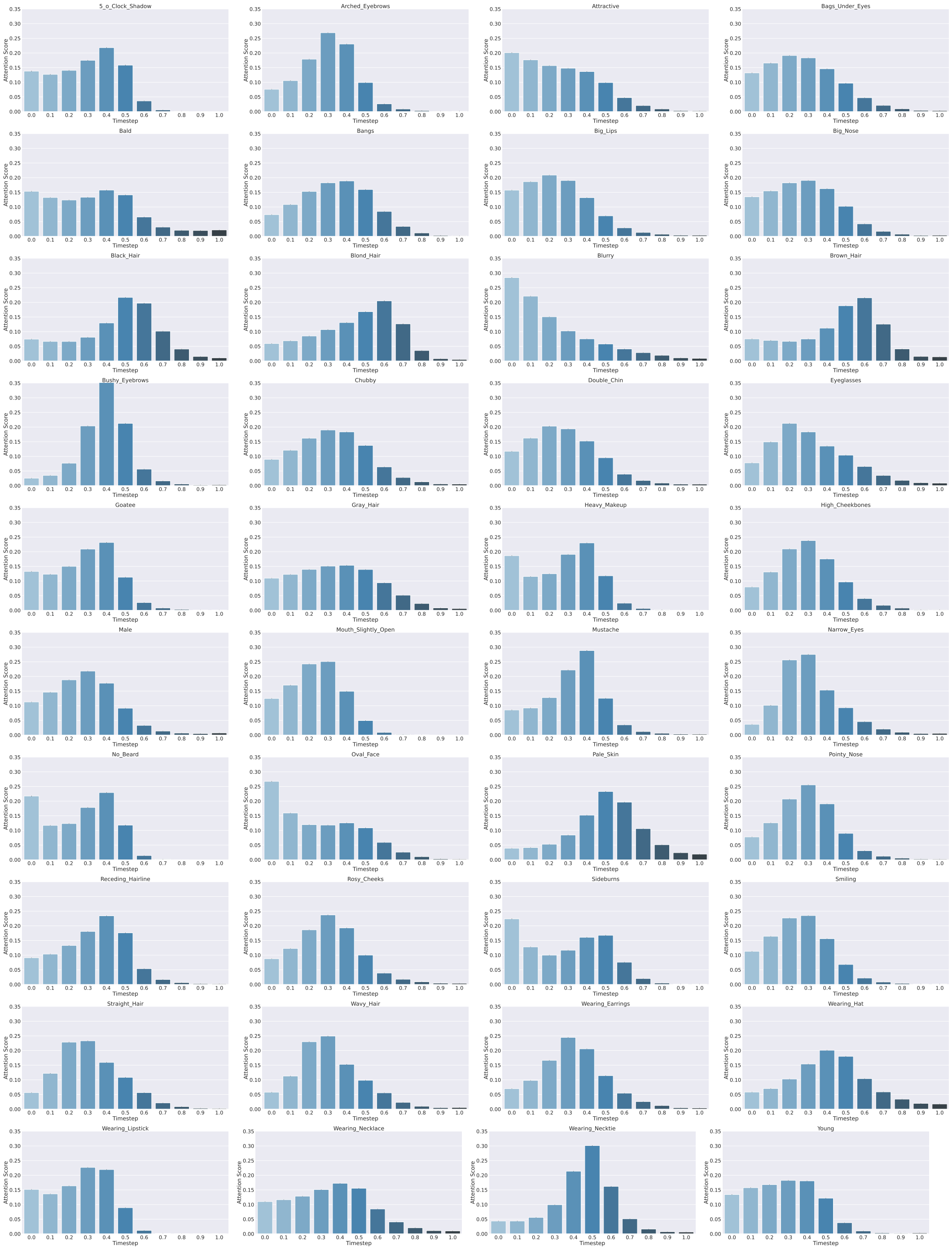

5_o_Clock_Shadow
Arched_Eyebrows
Attractive
Bags_Under_Eyes
Bald
Bangs
Big_Lips
Big_Nose
Black_Hair
Blond_Hair
Blurry
Brown_Hair
Bushy_Eyebrows
Chubby
Double_Chin
Eyeglasses
Goatee
Gray_Hair
Heavy_Makeup
High_Cheekbones
Male
Mouth_Slightly_Open
Mustache
Narrow_Eyes
No_Beard
Oval_Face
Pale_Skin
Pointy_Nose
Receding_Hairline
Rosy_Cheeks
Sideburns
Smiling
Straight_Hair
Wavy_Hair
Wearing_Earrings
Wearing_Hat
Wearing_Lipstick
Wearing_Necklace
Wearing_Necktie
Young

Figure 29: Attention Score profiles for different parts of the trajectory-based representation on CelebA when using the VDRL stochastic encoder.

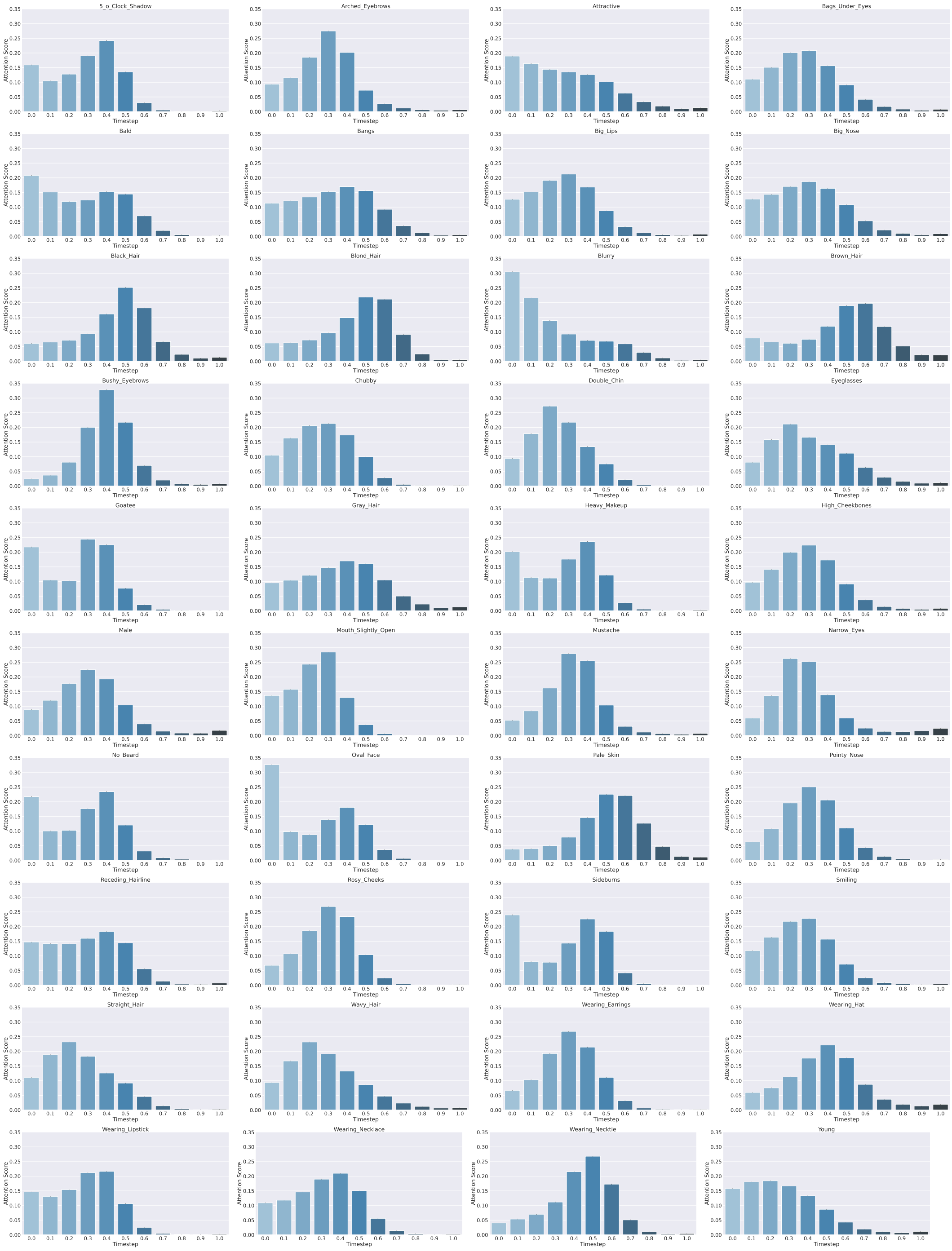


Figure 30: Attention Score profiles for different parts of the trajectory-based representation on CelebA when using the DRL deterministic encoder.