A. Theoretical Analysis of the Trade-off Between Pretrained and One-Point Velocities for Oracle MeanFlow Learning

Proposition A.1. *For any $\lambda \in [0,1]$, consider the combination of the one-point estimator and the pre-trained teacher velocity*

$$\mathbf{w}_\lambda := (1-\lambda)\hat{\mathbf{v}} + \lambda \mathbf{v}_{\boldsymbol{\phi}}.$$

Plugging $\mathbf{w}_\lambda$ into the target $\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w})$ (with $\mathbf{w} = \mathbf{w}_\lambda$) yields the corresponding loss $\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\theta}; \mathbf{w}_\lambda)$. Consider the following three residuals: the one-point velocity residual, the teacher–oracle velocity residual, and the oracle bias.

$$\begin{aligned}
\delta\hat{\mathbf{v}}_t &:= \hat{\mathbf{v}}(\mathbf{z}_t, t) - \mathbf{v}(\mathbf{z}_t, t), \\
\delta\mathbf{v}^{\boldsymbol{\phi}}_t &:= \mathbf{v}_{\boldsymbol{\phi}}(\mathbf{z}_t, t) - \mathbf{v}(\mathbf{z}_t, t), \\
\delta\mathbf{h}(\mathbf{z}_t, t, s) &:= \mathbf{h}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s) - \mathbf{h}(\mathbf{z}_t, t, s).
\end{aligned}$$

Then the MF loss admits the following decomposition:

$$\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\theta}; \mathbf{w}_\lambda) = \mathbb{E}_{t,\mathbf{z}_t}\Big\|\mathbf{h}_{\boldsymbol{\theta}} - \big(\mathbf{h} + \mathbf{B} + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}\, \delta\mathbf{v}^{\boldsymbol{\phi}}_t\big)\Big\|^2 + (1-\lambda)^2\, \mathbb{E}\big\|\mathbf{A}_{\boldsymbol{\theta}^-}\, \delta\hat{\mathbf{v}}_t\big\|^2,$$

where $\mathbf{B}(\mathbf{z}_t, t, s) := (t-s)\Big(\partial_t \delta\mathbf{h} + \big(\nabla_{\mathbf{x}} \delta\mathbf{h}\big)\mathbf{v}\Big)$ and $\mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s) := \mathbf{I} - (t-s)\, \nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-}$.

Proof. Throughout the proof we fix s and suppress the dependence on $(\mathbf{z}_t, t, s)$ whenever there is no ambiguity. All expectations are taken over $(t, \mathbf{z}_t)$ and when $\hat{\mathbf{v}}$ is random (e.g. a Monte–Carlo estimator).

We first derive the affine form of the teacher target around the oracle velocity $\mathbf{v}$. Recall from the MF construction that for any proxy velocity $\mathbf{w}$, the target associated with $\mathbf{h}_{\boldsymbol{\theta}^-}$ is

$$\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}) = \mathbf{w} - (t-s)\Big((\nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-})\, \mathbf{w} + \partial_t \mathbf{h}_{\boldsymbol{\theta}^-}\Big),$$

so for two velocities $\mathbf{w}$ and $\mathbf{v}$ we have

$$\begin{aligned}
\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}) - \mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{v}) &= \Big[\mathbf{w} - (t-s)(\nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-})\, \mathbf{w}\Big] - \Big[\mathbf{v} - (t-s)(\nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-})\, \mathbf{v}\Big] \\
&= \big(\mathbf{I} - (t-s)\, \nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-}\big)\, (\mathbf{w} - \mathbf{v}).
\end{aligned}$$

By the definition of $\mathbf{A}_{\boldsymbol{\theta}^-}$ in the proposition, $\mathbf{A}_{\boldsymbol{\theta}^-} := \mathbf{I} - (t-s)\, \nabla_{\mathbf{x}} \mathbf{h}_{\boldsymbol{\theta}^-}$, so this can be written as

$$\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}) = \mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{v}) + \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \big(\mathbf{w} - \mathbf{v}(\mathbf{z}_t, t)\big).$$

On the other hand, by the PDE relation between the oracle flow map $\mathbf{h}$ and the oracle velocity $\mathbf{v}$, and by the definition $\delta\mathbf{h} = \mathbf{h}_{\boldsymbol{\theta}^-} - \mathbf{h}$ and $\mathbf{B}(\mathbf{z}_t, t, s) := (t-s)\big(\partial_t \delta\mathbf{h} + (\nabla_{\mathbf{x}} \delta\mathbf{h})\, \mathbf{v}\big)$, we have (see the derivation in the main text)

$$\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{v}) = \mathbf{h}(\mathbf{z}_t, t, s) + \mathbf{B}(\mathbf{z}_t, t, s).$$

Combining the last two displays gives, for every proxy velocity $\mathbf{w}$,

$$\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}) = \mathbf{h}(\mathbf{z}_t, t, s) + \mathbf{B}(\mathbf{z}_t, t, s) + \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \big(\mathbf{w} - \mathbf{v}(\mathbf{z}_t, t)\big), \tag{5}$$

which is exactly the affine reparametrization of the target around the oracle velocity $\mathbf{v}$.

We now prove the claimed decomposition of the MF loss. By the definitions of the residual velocities,

$$\hat{\mathbf{v}} = \mathbf{v} + \delta\hat{\mathbf{v}}_t, \qquad \mathbf{v}_{\boldsymbol{\phi}} = \mathbf{v} + \delta\mathbf{v}^{\boldsymbol{\phi}}_t,$$

the mixed velocity satisfies, for any $\lambda \in [0,1]$,

$$\mathbf{w}_\lambda = (1-\lambda)\hat{\mathbf{v}} + \lambda\mathbf{v}_{\boldsymbol{\phi}} = (1-\lambda)\big(\mathbf{v} + \delta\hat{\mathbf{v}}_t\big) + \lambda\big(\mathbf{v} + \delta\mathbf{v}^{\boldsymbol{\phi}}_t\big) = \mathbf{v} + (1-\lambda)\delta\hat{\mathbf{v}}_t + \lambda\delta\mathbf{v}^{\boldsymbol{\phi}}_t. \tag{6}$$

Substituting $\mathbf{w} = \mathbf{w}_\lambda$ and Equation (6) into Equation (5) yields

$$\begin{aligned}\mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}_\lambda) &= \mathbf{h} + \mathbf{B} + \mathbf{A}_{\boldsymbol{\theta}^-}\big(\mathbf{w}_\lambda - \mathbf{v}\big) \\ &= \mathbf{h} + \mathbf{B} + \mathbf{A}_{\boldsymbol{\theta}^-}\big((1-\lambda)\delta\hat{\mathbf{v}}_t + \lambda\delta\mathbf{v}^{\boldsymbol{\phi}}_t\big) \\ &= \mathbf{h}(\mathbf{z}_t, t, s) + \mathbf{B}(\mathbf{z}_t, t, s) + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \delta\mathbf{v}^{\boldsymbol{\phi}}_t + (1-\lambda)\, \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \delta\hat{\mathbf{v}}_t. \end{aligned} \tag{7}$$

For the remainder of the proof we fix $(t, \mathbf{z}_t)$ and abbreviate

$$\mathbf{h}_{\boldsymbol{\theta}} := \mathbf{h}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, s), \quad \mathbf{h} := \mathbf{h}(\mathbf{z}_t, t, s), \quad \mathbf{A}_{\boldsymbol{\theta}^-} := \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s), \quad \mathbf{B} := \mathbf{B}(\mathbf{z}_t, t, s).$$

By Equation (2), the MF loss at $\mathbf{w}_\lambda$ is

$$\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\theta}; \mathbf{w}_\lambda) = \mathbb{E}\Big\|\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, s) - \mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}_\lambda)\Big\|_2^2.$$

Using Equation (7), we can write the pointwise residual as

$$\mathbf{h}_{\boldsymbol{\theta}} - \mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}_\lambda) = \Big[\mathbf{h}_{\boldsymbol{\theta}} - \big(\mathbf{h} + \mathbf{B} + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\mathbf{v}^{\boldsymbol{\phi}}_t\big)\Big] - (1-\lambda)\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t.$$

Let

$$\mathbf{Y} := \mathbf{h}_{\boldsymbol{\theta}} - \big(\mathbf{h} + \mathbf{B} + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\mathbf{v}^{\boldsymbol{\phi}}_t\big),$$

so that

$$\mathbf{h}_{\boldsymbol{\theta}} - \mathbf{h}^{\mathrm{tgt}}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s; \mathbf{w}_\lambda) = \mathbf{Y} - (1-\lambda)\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t.$$

Therefore

$$\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\theta}; \mathbf{w}_\lambda) = \mathbb{E}\big\|\mathbf{Y} - (1-\lambda)\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t\big\|_2^2 = \mathbb{E}\|\mathbf{Y}\|_2^2 - 2(1-\lambda)\, \mathbb{E}\big\langle \mathbf{Y}, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t\big\rangle + (1-\lambda)^2\, \mathbb{E}\big\|\mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t\big\|_2^2. \tag{8}$$

Since the one-point estimator is conditionally unbiased:

$$\mathbb{E}\big[\hat{\mathbf{v}}(\mathbf{z}_t, t) \,\big|\, \mathbf{z}_t\big] = \mathbf{v}(\mathbf{z}_t, t), \qquad \text{so} \qquad \mathbb{E}\big[\delta\hat{\mathbf{v}}_t \,\big|\, \mathbf{z}_t\big] = 0.$$

For fixed $(t, \mathbf{z}_t)$, the quantities $\mathbf{h}_{\boldsymbol{\theta}}$, $\mathbf{h}$, $\mathbf{B}$, $\mathbf{v}$, $\mathbf{v}_{\boldsymbol{\phi}}$, hence $\mathbf{Y}$ and $\mathbf{A}_{\boldsymbol{\theta}^-}$, are deterministic, and the only randomness comes from the internal noise of $\hat{\mathbf{v}}$. Thus

$$\begin{aligned}\mathbb{E}\big\langle \mathbf{Y}, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t\big\rangle &= \mathbb{E}_{t,\mathbf{z}_t}\Big[\mathbb{E}\big[\big\langle \mathbf{Y}, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\hat{\mathbf{v}}_t\big\rangle \big| \mathbf{z}_t\big]\Big] \\ &= \mathbb{E}_{t,\mathbf{z}_t}\Big[\big\langle \mathbf{Y}, \mathbf{A}_{\boldsymbol{\theta}^-}\, \mathbb{E}[\delta\hat{\mathbf{v}}_t | \mathbf{z}_t]\big\rangle\Big] \\ &= \mathbb{E}_{t,\mathbf{z}_t}\big\langle \mathbf{Y}, \mathbf{A}_{\boldsymbol{\theta}^-} \cdot \mathbf{0}\big\rangle = \mathbf{0}.\end{aligned}$$

Hence the cross term in Equation (8) is zero, and we obtain

$$\mathcal{L}_{\mathrm{MF}}(\boldsymbol{\theta}; \mathbf{w}_\lambda) = \mathbb{E}_{t,\mathbf{z}_t}\Big\|\mathbf{h}_{\boldsymbol{\theta}}(\mathbf{z}_t, t, s) - \big(\mathbf{h}(\mathbf{z}_t, t, s) + \mathbf{B}(\mathbf{z}_t, t, s) + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \delta\mathbf{v}^{\boldsymbol{\phi}}_t\big)\Big\|_2^2 + (1-\lambda)^2\, \mathbb{E}\big\|\mathbf{A}_{\boldsymbol{\theta}^-}(\mathbf{z}_t, t, s)\, \delta\hat{\mathbf{v}}_t\big\|_2^2,$$

which is exactly the claimed decomposition Equation (4). This completes the proof. □

From Proposition 3.1, several implications follow. First, using the pre-trained velocity $\mathbf{v}_{\boldsymbol{\phi}}$ improves stability: when $\lambda = 1$, the noisy one-point term vanishes, which reduces gradient variance and typically stabilizes and accelerates optimization.

Second, this comes at the cost of bias. The model no longer regresses to $\mathbf{h}$, but to $\mathbf{h} + \mathbf{B} + \lambda\, \mathbf{A}_{\boldsymbol{\theta}^-}\delta\mathbf{v}^{\boldsymbol{\phi}}_s$. When $\mathbf{v}_{\boldsymbol{\phi}}$ is accurate (small $\delta\mathbf{v}^{\boldsymbol{\phi}}_s$) and the time step is small so that $\mathbf{A}_{\boldsymbol{\theta}^-} \approx \mathbf{I}$, this additional bias is small. In the small-step regime, the shift is approximately $\lambda\, \delta\mathbf{v}^{\boldsymbol{\phi}}_s$, so the practical target is close to the oracle one whenever $\|\delta\mathbf{v}^{\boldsymbol{\phi}}_s\|$ is small. However, under domain shift, $\delta\mathbf{v}^{\boldsymbol{\phi}}_s$ can be non-negligible and the learned $\mathbf{h}_{\boldsymbol{\theta}}$ will partially compensate for this mismatch.

Third, the mixing weight λ offers a natural tuning mechanism. A schedule with λ gradually increasing to 1 can start from a regime with low bias and low variance: initially, the one-point label maintains an unbiased target while the pre-trained signal reduces variance; later, larger λ leverages the smoothness of $\mathbf{v}_{\boldsymbol{\phi}}$ for stable convergence. In practice, we use a simple two-stage schedule: $\lambda = 0$ in the distillation stage and $\lambda = 1$ in the (optional) bootstrapping stage.

Finally, the coupling to the self-teacher remains simple. Both $\mathbf{B}$ and $\mathbf{A}_{\boldsymbol{\theta}^-}$ are treated as stop-gradient with respect to $\boldsymbol{\theta}^-$, so updates in $\boldsymbol{\theta}$ still solve a least-squares fit to a shifted target. As the self-teacher improves over training, $\delta\mathbf{h}$ shrinks and $\mathbf{B}$ decreases, further reducing the induced bias.

B. More Hyperparameter Settings

MF-RAE's Flow Matching Pre-Training Setup. We directly use the flow matching model checkpoint provided by RAE [48], where the model on ImageNet 256 is trained for 800 epochs, and that on ImageNet 512 is trained for 400 epochs.

MF-RAE's CMT Setup. The CMT stage's learning rate is also 1e-4, but the EMA β is set to 0.999 for faster convergence, thanks to CMT's stability and fixed target.

MF-RAE's CMT Setup on ImageNet 256. CMT generates eight trajectories from the teacher DiT^{DH}-XL without guidance during each iteration. We use the Euler ODE solver with 16 steps (FID=2.3). For each trajectory, it generates $16 \times 15/2$ =120 pairs, but we randomly select 64 pairs for optimization to fit on the H100 GPU. The total batch size is thus $64 \times 8 = 512$ pairs. We conduct this training on 8 H100 GPUs for 27K iterations.

MF-RAE's CMT Setup on ImageNet 512. CMT generates two trajectories from the teacher DiT^{DH}-XL without guidance during each iteration. We use the Euler ODE solver with 16 steps (FID=1.66 without any guidance). For each trajectory, we randomly select 8 pairs for optimization to fit on the H100 GPU. The total batch size is thus $8 \times 2 = 16$ pairs. We accumulate gradients for eight rounds to enlarge the batch size. We conduct this training on 2 H100 GPUs for 27K iterations.

C. Generated Samples

Figure 2. ImageNet 256 MF-RAE 1-step samples on class 437: beacon, lighthouse, beacon light, pharos.

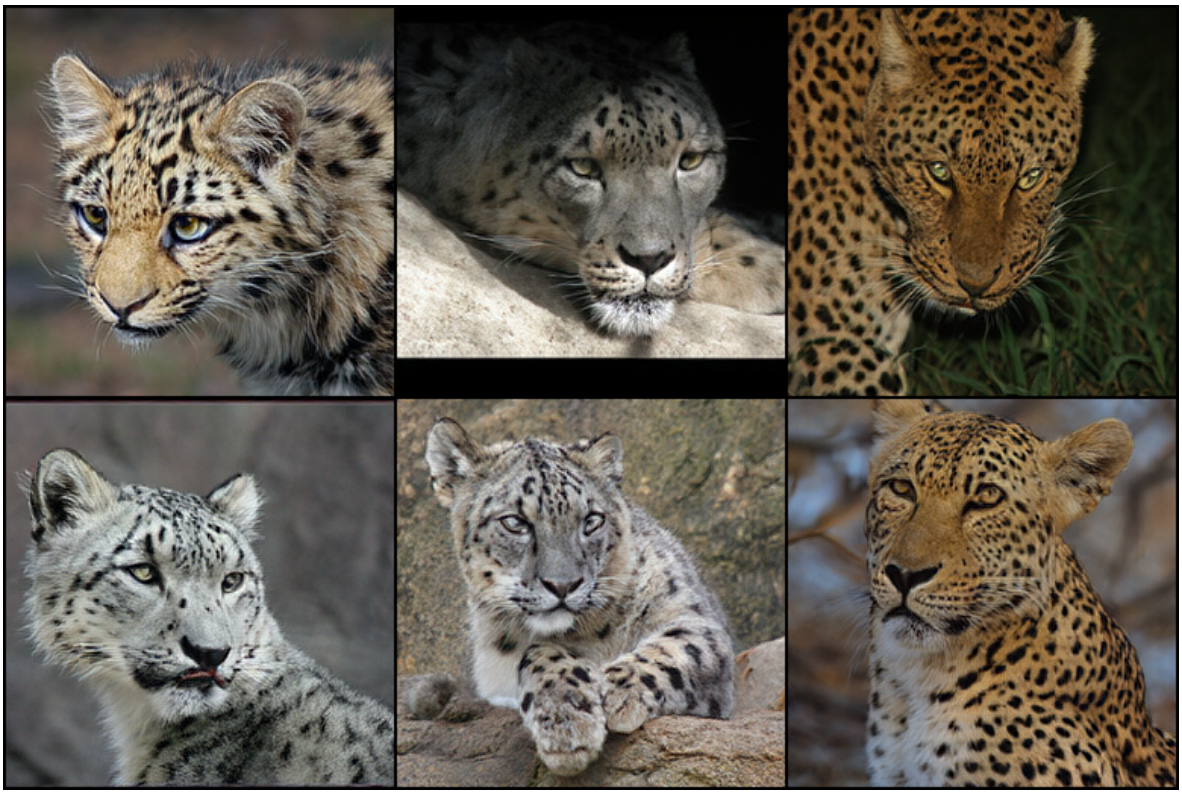

Figure 3. ImageNet 256 MF-RAE 1-step samples on classes 288 and 290: leopard and snow leopard.

Figure 4. ImageNet 256 MF-RAE 1-step samples on classes 13, 14, 94, and 134: snowbird, indigo bird, hummingbird, and crane bird.

Figure 5. ImageNet 256 MF-RAE 1-step samples for various dogs.

Figure 6. ImageNet 256 MF-RAE 1-step samples for class 933: cheeseburger.

Figure 7. ImageNet 256 MF-RAE 1-step samples for class 959: carbonara.

Figure 8. ImageNet 256 MF-RAE 1-step samples for class 947: mushroom.